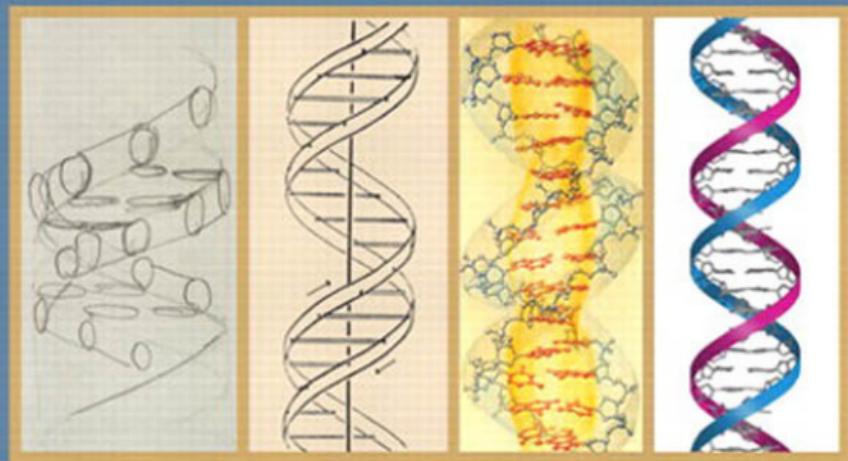


# MOLECULAR BIOLOGY OF THE GENE

SEVENTH EDITION



WATSON • BAKER • BELL  
GANN • LEVINE • LOSICK

# MOLECULAR BIOLOGY OF THE GENE

SEVENTH EDITION

*This page intentionally left blank*

# MOLECULAR BIOLOGY OF THE GENE

SEVENTH EDITION

**JAMES D. WATSON**

*Cold Spring Harbor Laboratory*

**ALEXANDER GANN**

*Cold Spring Harbor Laboratory*

**TANIA A. BAKER**

*Massachusetts Institute of Technology*

**MICHAEL LEVINE**

*University of California, Berkeley*

**STEPHEN P. BELL**

*Massachusetts Institute of Technology*

**RICHARD LOSICK**

*Harvard University*

*With*

**STEPHEN C. HARRISON**

*Harvard Medical School*

*(Chapter 6: The Structure of Proteins)*

**PEARSON**

Boston Columbus Indianapolis New York San Francisco Upper Saddle River  
Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montréal Toronto  
Delhi Mexico City São Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo



**COLD SPRING HARBOR LABORATORY PRESS**  
*Cold Spring Harbor, New York*

**PEARSON**

**Editor-in-Chief:** Beth Wilbur  
**Senior Acquisitions Editor:** Josh Frost  
**Executive Director of Development:** Deborah Gale  
**Assistant Editor:** Katherine Harrison-Adcock  
**Managing Editor:** Michael Early  
**Production Project Manager:** Lori Newman  
**Illustrators:** Dragonfly Media Group  
**Manufacturing Buyer:** Michael Penne  
**Director of Marketing:** Christy Lesko  
**Executive Marketing Manager:** Lauren Harp  
**Executive Media Producer:** Laura Tommasi  
**Editorial Media Producer:** Lee Ann Doctor  
**Supervising Media Project Manager:** David Chavez  
**Director of Content Development, MasteringBiology:** Natania Mlawer  
**Content Specialist, MasteringBiology:** J. Zane Barlow, PhD

**COLD SPRING HARBOR LABORATORY PRESS**

**Publisher and Sponsoring Editor:** John Inglis  
**Editorial Director:** Alexander Gann  
**Director of Editorial Development:** Jan Argentine  
**Managing Editor and Developmental Editor:** Kaaren Janssen  
**Project Manager:** Inez Sialiano  
**Production Manager:** Denise Weiss  
**Production Editor:** Kathleen Bubbeo  
**Permissions Coordinator:** Carol Brown  
**Crystal Structure Images:** Leemor Joshua-Tor and Stephen C. Harrison  
**Cover Designer:** Mike Albano

*Front and Back Cover Images:* Far left, drawing by Francis Crick, Wellcome Library, London. Second from left, from Watson J.D. and Crick F.H.C. 1953. *Nature* **171**: 737–738. Second from right, Irving Geis illustration. Rights owned by Howard Hughes Medical Institute. Not to be reproduced without permission. Far right, structure by Leemor Joshua-Tor (image prepared with PyMOL).

Credits and acknowledgments for materials borrowed from other sources and reproduced, with permission, in this textbook appear on the appropriate page within the text.

---

Copyright © 2014, 2008, 2004 Pearson Education, Inc. All rights reserved. Manufactured in the United States of America. This publication is protected by Copyright, and permission should be obtained from the publisher prior to any prohibited reproduction, storage in a retrieval system, or transmission in any form or by any means, electronic, mechanical, photocopying, recording, or likewise. To obtain permission(s) to use material from this work, please submit a written request to Pearson Education, Inc., Permissions Department, 1900 E. Lake Ave., Glenview, IL 60025. For information regarding permissions, call (847) 486-2635.

Readers may view, browse, and/or download material for temporary copying purposes only, provided these uses are for noncommercial personal purposes. Except as provided by law, this material may not be further reproduced, distributed, transmitted, modified, adapted, performed, displayed, published, or sold in whole or in part, without prior written permission from the publisher.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and the publisher was aware of a trademark claim, the designations have been printed in initial caps or all caps.

MasteringBiology and BioFlix are trademarks, in the U.S. and/or other countries, of Pearson Education, Inc. or its affiliates.

**Library of Congress Cataloging-in-Publication Data**

Watson, James D.

Molecular biology of the gene / James D. Watson, Cold Spring Harbor Laboratory, Tania A. Baker, Massachusetts Institute of Technology, Alexander Gann, Cold Spring Harbor Laboratory, Michael Levine, University of California, Berkeley, Richard Losick, Harvard University.

pages cm

Includes bibliographical references and index.

ISBN-13: 978-0-321-76243-6 (hardcover (student ed))

ISBN-10: 0-321-76243-6 (hardcover (student ed))

ISBN-13: 978-0-321-90537-6 (paper (a la carte))

ISBN-10: 0-321-90537-7 (paper (a la carte))

[etc.]

1. Molecular biology--Textbooks. 2. Molecular genetics--Textbooks. I. Title.

QH506.M6627 2013

572'.33--dc23

2012046093

1 2 3 4 5 6 7 8 9 10—DOW—17 16 15 14 13

**PEARSON**

[www.pearsonhighered.com](http://www.pearsonhighered.com)



COLD SPRING HARBOR  
LABORATORY PRESS  
[www.cshlpress.org](http://www.cshlpress.org)

ISBN 10: 0-321-76243-6 (Student Edition)

ISBN 13: 978-0-321-76243-6 (Student Edition)

ISBN 10: 0-321-90264-5 (Instructor's Review Copy)

ISBN 13: 978-0-321-90264-1 (Instructor's Review Copy)

ISBN 10: 0-321-90537-7 (Books à la Carte Edition)

ISBN 13: 978-0-321-90537-6 (Books à la Carte Edition)

# Preface

THE NEW EDITION OF *MOLECULAR BIOLOGY OF THE GENE* appears in this, its 7th edition, on the 60th anniversary of the discovery of the structure of DNA in 1953, an occasion celebrated by our cover design. The double-helical structure, held together by specific pairing between the bases on the two strands, has become one of the iconic images of science. The image of the microscope was perhaps the icon of science in the late 19th century, displaced by the mid 20th century by the graphical representation of the atom with its orbiting electrons. But by the end of the century that image had in turn given way to the double helix.

The field of molecular biology as we understand it today was born out of the discovery of the DNA structure and the agenda for research that that structure immediately provided. The paper by Watson and Crick proposing the double helix ended with a now famous sentence: “It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.” The structure suggested how DNA could replicate, opening the way to investigate, in molecular terms, how genes are passed down through generations. It was also immediately apparent that the order of bases along a DNA molecule could represent a “genetic code,” and so an attack on that second great mystery of genetics—how genes encode characteristics—could also be launched.

By the time the first edition of *Molecular Biology of the Gene* was published, just 12 years later in 1965, it had been confirmed that DNA replicated in the manner suggested by the model, the genetic code had all but been cracked, and the mechanism by which genes are expressed, and how that expression is regulated, had been established at least in outline. The field of molecular biology was ripe for its first textbook, defining for the first time the curriculum for undergraduate courses in this topic.

Our understanding of the mechanisms underlying these processes has hugely increased over the last 48 years since that first edition, often driven by technological advances, including DNA sequencing (another anniversary this year is the 10th anniversary of completion of the human genome project). The current edition of *Molecular Biology of the Gene* celebrates both the central intellectual framework of the field put in place in that first edition and the extraordinary mechanistic, biological, and evolutionary understanding that has since been achieved.

## New to This Edition

There are a number of major changes to the new edition. As well as wide-ranging updates, these include changes in organization, addition of completely new chapters, and the addition of new topics within existing chapters.

- **New Part 2 on the Structure and Study of Macromolecules.** In this new section, each of the three major macromolecules gets its own chapter. The DNA chapter is retained from the previous edition, but what was previously just a short section at the end of that chapter is now expanded into a whole new chapter on the structure of RNA. The chapter on the structure of proteins is completely new and was written for this edition by Stephen Harrison (Harvard University).

- **Techniques chapter moved from the end of the book into Part 2.** This revised and relocated chapter introduces the important techniques that will be referred to throughout the book. In addition to many of the basic techniques of molecular biology, this chapter now includes an updated section on many genomics techniques routinely employed by molecular biologists. Techniques more specialized for particular chapters appear as boxes within those chapters.
- **Completely new chapter on The Origin and Early Evolution of Life.** This chapter shows how the techniques of molecular biology and biochemistry allow us to consider—even reconstruct—how life might have arisen and addresses the prospect of creating life in a test tube (synthetic biology). The chapter also reveals how, even at the very early stages of life, molecular processes were subject to evolution.
- **New material on many aspects of gene regulation.** Part 5 of the book is concerned with gene regulation. In this edition we have introduced significant new topics, such as quorum sensing in bacterial populations, the bacterial CRISPR defense system and piRNAs in animals, the function of Polycomb, and increased discussion of other so-called “epigenetic” mechanisms of gene regulation in higher eukaryotes. The regulation of “paused polymerase” at many genes during animal development and the critical involvement of nucleosome positioning and remodeling at promoters during gene activation are also new topics to this edition.
- **End-of-chapter questions.** Appearing for the first time in this edition, these include both short answer and data analysis questions. The answers to the even-numbered questions are included as Appendix 2 at the back of the book.
- **New experiments and experimental approaches reflecting recent advances in research.** Integrated within the text are new experimental approaches and applications that broaden the horizons of research. These include, for example, a description of how the genetic code can be experimentally expanded to generate novel proteins, creation of a synthetic genome to identify the minimal features required for life, discussion of new genome-wide analysis of nucleosome positioning, experiments on bimodal switches in bacteria, and how new antibacterial drugs are being designed that target the quorum-sensing pathways required for pathogenesis.

## Supplements

### *MasteringBiology* [www.masteringbiology.com](http://www.masteringbiology.com)

MasteringBiology is an online homework, tutorial, and assessment system that delivers self-paced tutorials that provide individualized coaching, focus on your course objectives, and are responsive to each student’s progress. The Mastering system helps instructors maximize class time with customizable, easy-to-assign, and automatically graded assessments that motivate students to learn outside of class and arrive prepared for lecture. MasteringBiology includes the book’s end-of-chapter problems, eighteen 3D structure tutorials, reading quizzes, animations, videos, and a wide variety of activities. The eText is also available through MasteringBiology, providing access to the complete textbook and featuring powerful interactive and customization functions.

### *Instructor Resource DVD* 978-0-321-88342-1/0-321-88342-X

Available free to all adopters, this dual-platform DVD-ROM contains all art and tables from the book in JPEG and PowerPoint in high-resolution (150 dpi) files. The PowerPoint slides include problems formatted for use with Classroom Response Systems. This DVD-ROM also contains an answer key for all of the end-of-chapter Critical Thinking questions included in MasteringBiology.

### *Transparency Acetates* 978-0-321-88341-4/0-321-88341-1

Features approximately 90 four-color illustrations from the text. These transparencies are free to all adopters.

## Cold Spring Harbor Laboratory Photographs

As in the previous edition, each part opener includes photographs, some newly added to this edition. These pictures, selected from the archives of Cold Spring Harbor Laboratory, were all taken at the Lab, the great majority during the Symposia hosted there almost every summer since 1933. Captions identify who is in each picture and when it was taken. Many more examples of these historic photos can be found at the CSHL archives website (<http://archives.cshl.edu/>).

## Acknowledgments

Parts of the current edition grew out of an introductory course on molecular biology taught by one of us (RL) at Harvard University, and this author is grateful to Steve Harrison and Jim Wang who contributed to this course in past years. In the case of Steve Harrison, we are additionally indebted to him for writing and illustrating a brand new chapter on protein structure especially for this new edition. No one could be better qualified for such a task, and we are the grateful beneficiaries of—and the book is immeasurably improved by—his contribution.

We are also grateful to Craig Hunter, who earlier wrote the section on the worm for Appendix 1, and to Rob Martienssen, who wrote the section on plants for that same appendix.

We have shown sections of the manuscript to various colleagues and their comments have been extremely helpful. Specifically we thank Katsura Asano, Stephen Blacklow, Jamie Cate, Amy Caudy, Irene Chen, Victoria D’Souza, Richard Ebright, Mike Eisen, Chris Fromme, Brenton Graveley, Chris Hammell, Steve Hahn, Oliver Hobert, Ann Hochschild, Jim Hu, David Jerulzalmi, Leemor Joshua-Tor, Sandy Johnson, Andrew Knoll, Adrian Krainer, Julian Lewis, Sue Lovett, Karolin Luger, Kristen Lynch, Rob Martienssen, Bill McGinnis, Matt Michael, Lily Mirels, Nipam Patel, Mark Ptashne, Danny Reinberg, Dimitar Sasselov, David Shechner, Sarah T. Stewart-Mukhopadhyay, Bruce Stillman, and Jack Szostak.

We also thank those who provided us with figures, or the wherewithal to create them: Sean Carroll, Seth Darst, Paul Fransz, Brenton Graveley, Ann Hochschild, Julian Lewis, Bill McGinnis, Phoebe Rice, Dan Rokhsar, Nori Satoh, Matt Scott, Ali Shilatifard, Peter Sorger, Tom Steitz, Andrzej Stasiak, Dan Voytas, and Steve West.

New to this edition are end-of-chapter questions, provided by Mary Ellen Wiltout, and we thank her for these efforts that have enhanced the new edition. In addition, Mary Ellen helped with revisions to the DNA repair chapter.

We are indebted to Leemor Joshua-Tor, who so beautifully rendered the majority of the structure figures throughout the book. Her skill and patience are much appreciated.

We are also grateful to those who provided their software<sup>1</sup>: Per Kraulis, Robert Esnouf, Ethan Merritt, Barry Honig, and Warren Delano. Coordinates were obtained from the Protein Data Bank ([www.rcsb.org/pdb/](http://www.rcsb.org/pdb/)), and citations to those who solved each structure are included in the figure legends.

Our art program was again executed by a team from the Dragonfly Media Group, led by Craig Durant. Denise Weiss and Mike Albano produced a beautiful cover design. We thank Clare Bunce and the CSHL Archive for providing the photos for the part openers and for much help tracking them down.

We thank Josh Frost at Pearson who oversaw our efforts and was always on hand to help us out or provide advice. In development at CSHL Press, Jan Argentine provided great support, guidance, and perspective throughout the process. Our heartfelt thanks to Kaaren Janssen who was once again our constant savior—editing and organizing, encouraging and understanding—and unstintingly good-humored even on the darkest days. Inez Sialiano kept track of the output, and Carol Brown dealt with the permissions as efficiently as ever. In production, we relied heavily on the extraordinary efforts and patience

of Kathleen Bubbeo, for which we are most grateful. And we must also thank Denise Weiss, who oversaw production and ensured that the book looked so good by finessing the page layout and creating the design. John Inglis as ever created the environment in which this could all take place.

And once again, we thank our families for putting up with this book for a third time!

JAMES D. WATSON  
TANIA A. BAKER  
STEPHEN P. BELL  
ALEXANDER GANN  
MICHAEL LEVINE  
RICHARD LOSICK

<sup>1</sup>Per Kraulis granted permission to use MolScript (Kraulis P.J. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.* **24**: 946–950). Robert Esnouf gave permission to use BobScript (Esnouf R.M. 1997. *J. Mol. Graph.* **15**: 132–134). In addition, Ethan Merritt gave us use of Raster3D (Merritt E.A. and Bacon D.J. 1997. Raster3D: Photorealistic molecular graphics. *Methods Enzymol.* **277**: 505–524), and Barry Honig granted permission to use GRASP (Nicolls A., Sharp K.A., and Honig B. 1991. Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**: 281–296). Warren DeLano agreed to the use of PyMOL (DeLano W.L. 2002. *The PyMOL Molecular Graphics System*. DeLano Scientific, Palo Alto, California).

# About the Authors

**JAMES D. WATSON** is Chancellor Emeritus at Cold Spring Harbor Laboratory, where he was previously its Director from 1968 to 1993, President from 1994 to 2003, and Chancellor from 2003 to 2007. He spent his undergraduate years at the University of Chicago and received his Ph.D. in 1950 from Indiana University. Between 1950 and 1953, he did postdoctoral research in Copenhagen and Cambridge, England. While at Cambridge, he began the collaboration that resulted in the elucidation of the double-helical structure of DNA in 1953. (For this discovery, Watson, Francis Crick, and Maurice Wilkins were awarded the Nobel Prize in 1962.) Later in 1953, he went to the California Institute of Technology. He moved to Harvard in 1955, where he taught and did research on RNA synthesis and protein synthesis until 1976. He was the first Director of the National Center for Genome Research of the National Institutes of Health from 1989 to 1992. Dr. Watson was sole author of the first, second, and third editions of *Molecular Biology of the Gene*, and a co-author of the fourth, fifth and sixth editions. These were published in 1965, 1970, 1976, 1987, 2003, and 2007, respectively. He is also a co-author of two other textbooks, *Molecular Biology of the Cell* and *Recombinant DNA*, as well as author of the celebrated 1968 memoir, *The Double Helix*, which in 2012 was listed by the Library of Congress as one of the 88 Books That Shaped America.

**TANIA A. BAKER** is the Head of the Department and Whitehead Professor of Biology at the Massachusetts Institute of Technology and an Investigator of the Howard Hughes Medical Institute. She received a B.S. in biochemistry from the University of Wisconsin, Madison, and a Ph.D. in biochemistry from Stanford University in 1988. Her graduate research was carried out in the laboratory of Professor Arthur Kornberg and focused on mechanisms of initiation of DNA replication. She did postdoctoral research in the laboratory of Dr. Kiyoshi Mizuuchi at the National Institutes of Health, studying the mechanism and regulation of DNA transposition. Her current research explores mechanisms and regulation of genetic recombination, enzyme-catalyzed protein unfolding, and ATP-dependent protein degradation. Professor Baker received the 2001 Eli Lilly Research Award from the American Society of Microbiology and the 2000 MIT School of Science Teaching Prize for Undergraduate Education and is a Fellow of the American Academy of Arts and Sciences since 2004 and was elected to the National Academy of Sciences in 2007. She is co-author (with Arthur Kornberg) of the book *DNA Replication*, Second Edition.

**STEPHEN P. BELL** is a Professor of Biology at the Massachusetts Institute of Technology and an Investigator of the Howard Hughes Medical Institute. He received B.A. degrees from the Department of Biochemistry, Molecular Biology, and Cell Biology and the Integrated Sciences Program at Northwestern University and a Ph.D. in biochemistry at the University of California, Berkeley, in 1991. His graduate research was carried out in the laboratory of Dr. Robert Tjian and focused on eukaryotic transcription. He did postdoctoral research in the laboratory of Dr. Bruce Stillman at Cold Spring Harbor Laboratory, working on the initiation of eukaryotic DNA replication. His current research focuses on the mechanisms controlling the duplication of eukaryotic chromosomes. Professor Bell received the 2001 ASBMB–Schering Plough Scientific Achievement Award, the

1998 Everett Moore Baker Memorial Award for Excellence in Undergraduate Teaching at MIT, the 2006 MIT School of Science Teaching Award, and the 2009 National Academy of Sciences Molecular Biology Award.

**ALEXANDER GANN** is the Lita Annenberg Hazen Dean and Professor in the Watson School of Biological Sciences at Cold Spring Harbor Laboratory. He is also a Senior Editor at Cold Spring Harbor Laboratory Press. He received his B.Sc. in microbiology from University College London and a Ph.D. in molecular biology from The University of Edinburgh in 1989. His graduate research was carried out in the laboratory of Noreen Murray and focused on DNA recognition by restriction enzymes. He did postdoctoral research in the laboratory of Mark Ptashne at Harvard, working on transcriptional regulation, and that of Jeremy Brockes at the Ludwig Institute of Cancer Research at University College London, where he worked on newt limb regeneration. He was a Lecturer at Lancaster University, United Kingdom, from 1996 to 1999, before moving to Cold Spring Harbor Laboratory. He is co-author (with Mark Ptashne) of the book *Genes & Signals* (2002) and co-editor (with Jan Witkowski) of *The Annotated and Illustrated Double Helix* (2012).

**MICHAEL LEVINE** is a Professor of Genetics, Genomics and Development at the University of California, Berkeley, and is also Co-Director of the Center for Integrative Genomics. He received his B.A. from the Department of Genetics at the University of California, Berkeley, and his Ph.D. with Alan Garen in the Department of Molecular Biophysics and Biochemistry from Yale University in 1981. As a Postdoctoral Fellow with Walter Gehring and Gerry Rubin from 1982 to 1984, he studied the molecular genetics of *Drosophila* development. Professor Levine's research group currently studies the gene networks responsible for the gastrulation of the *Drosophila* and *Ciona* (sea squirt) embryos. He holds the F. Williams Chair in Genetics and Development at University of California, Berkeley. He was awarded the Monsanto Prize in Molecular Biology from the National Academy of Sciences in 1996 and was elected to the American Academy of Arts and Sciences in 1996 and the National Academy of Sciences in 1998.

**RICHARD LOSICK** is the Maria Moors Cabot Professor of Biology, a Harvard College Professor, and a Howard Hughes Medical Institute Professor in the Faculty of Arts and Sciences at Harvard University. He received his A.B. in chemistry at Princeton University and his Ph.D. in biochemistry at the Massachusetts Institute of Technology. Upon completion of his graduate work, Professor Losick was named a Junior Fellow of the Harvard Society of Fellows when he began his studies on RNA polymerase and the regulation of gene transcription in bacteria. Professor Losick is a past Chairman of the Departments of Cellular and Developmental Biology and Molecular and Cellular Biology at Harvard University. He received the Camille and Henry Dreyfus Teacher-Scholar Award and is a member of the National Academy of Sciences, a Fellow of the American Academy of Arts and Sciences, a Fellow of the American Association for the Advancement of Science, a Fellow of the American Academy of Microbiology, a member of the American Philosophical Society, and a former Visiting Scholar of the Phi Beta Kappa Society. Professor Losick is the 2007 winner of the Selman A. Waksman Award of the National Academy of Sciences, a 2009 winner of the Canada Gairdner Award, a 2012 winner of the Louisa Gross Horwitz Prize for Biology or Biochemistry of Columbia University, and a 2012 winner of the Harvard University Fannie Cox Award for Excellence in Science Teaching.

# Class Testers and Reviewers

We wish to thank all of the instructors for their thoughtful suggestions and comments on versions of many chapters in this book.

## Chapter Reviewers

- Ann Aguanno, *Marymount Manhattan College*  
David P. Aiello, *Austin College*  
Charles F. Austerberry, *Creighton University*  
David G. Bear, *University of New Mexico Health Sciences Center*  
Margaret E. Beard, *College of the Holy Cross*  
Gail S. Begley, *Northeastern University*  
Sanford Bernstein, *San Diego State University*  
Michael Blaber, *Florida State University*  
Nicole Bournias, *California State University, San Bernardino*  
John Boyle, *Mississippi State University*  
Suzanne Bradshaw, *University of Cincinnati*  
John G. Burr, *University of Texas at Dallas*  
Michael A. Campbell, *Pennsylvania State University, Erie, The Behrend College*  
Aaron Cassill, *University of Texas at San Antonio*  
Shirley Coomber, *King's College, University of London*  
Anne Cordon, *University of Toronto*  
Sumana Datta, *Texas A&M University*  
Jeff DeJong, *University of Texas at Dallas*  
Jurgen Denecke, *University of Leeds*  
Susan M. DiBartolomeis, *Millersville University*  
Santosh R. D'Mello, *University of Texas at Dallas*  
Robert J. Duronio, *University of North Carolina, Chapel Hill*  
Steven W. Edwards, *University of Liverpool*  
David Frick, *University of Wisconsin*  
Allen Gathman, *Southeast Missouri State University*  
Anthony D.M. Glass, *University of British Columbia*  
Elliott S. Goldstein, *Arizona State University*  
Ann Grens, *Indiana University, South Bend*  
Gregory B. Hecht, *Rowan University*  
Robert B. Helling, *University of Michigan*  
David C. Higgs, *University of Wisconsin, Parkside*  
Mark Kainz, *Colgate University*  
Gregory M. Kelly, *University of Western Ontario*  
Ann Kleinschmidt, *Allegheny College*  
Dan Krane, *Wright State University*  
Mark Levinthal, *Purdue University*  
Gary J. Lindquester, *Rhodes College*  
James Lodolce, *Loyola University Chicago*  
Curtis Loer, *University of San Diego*  
Virginia McDonough, *Hope College*  
Michael J. McPherson, *University of Leeds*  
Victoria Meller, *Tufts University*  
William L. Miller, *North Carolina State University*  
Dragana Miskovic, *University of Waterloo*  
David Mullin, *Tulane University*  
Jeffrey D. Newman, *Lycoming College*  
James B. Olesen, *Ball State University*  
Anthony J. Otsuka, *Illinois State University*  
Karen Palter, *Temple University*  
James G. Patton, *Vanderbilt University*  
Ian R. Phillips, *Queen Mary, University of London*  
Steve Picksley, *University of Bradford*  
Debra Pires, *University of California, Los Angeles*  
Todd P. Primm, *University of Texas at El Paso*  
Phillip E. Ryals, *The University of West Florida*  
Eva Sapi, *University of New Haven*  
Jon B. Scales, *Midwestern State University*  
Michael Schultze, *University of York*  
Venkat Sharma, *University of West Florida*

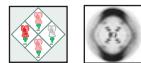
Erica L. Shelley, *University of Toronto at Mississauga*  
Elizabeth A. Shephard, *University College, London*  
Margaret E. Stevens, *Ripon College*  
Akif Uzman, *University of Houston, Downtown*  
Quinn Vega, *Montclair State University*  
Jeffrey M. Voight, *Albany College of Pharmacy*  
Lori L. Wallrath, *University of Iowa*  
Robert Wiggers, *Stephen F. Austin State University*  
Bruce C. Wightman, *Muhlenberg College*  
Bob Zimmermann, *University of Massachusetts*

### Class Testers

Charles F. Austerberry, *Creighton University*  
Christine E. Bezotté, *Elmira College*  
Astrid Helfant, *Hamilton College*  
Gerald Joyce, *The Scripps Research Institute*  
Jocelyn Krebs, *University of Alaska, Anchorage*  
Cran Lucas, *Louisiana State University in Shreveport*  
Anthony J. Otsuka, *Illinois State University*  
Charles Polson, *Florida Institute of Technology*  
Ming-Che Shih, *University of Iowa*

# Brief Contents

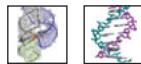
## PART 1



### HISTORY, 1

- 1 The Mendelian View of the World, 5
- 2 Nucleic Acids Convey Genetic Information, 21

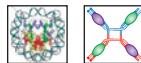
## PART 2



### STRUCTURE AND STUDY OF MACROMOLECULES, 45

- 3 The Importance of Weak and Strong Chemical Bonds, 51
- 4 The Structure of DNA, 77
- 5 The Structure and Versatility of RNA, 107
- 6 The Structure of Proteins, 121
- 7 Techniques of Molecular Biology, 147

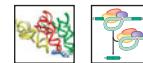
## PART 3



### MAINTENANCE OF THE GENOME, 193

- 8 Genome Structure, Chromatin, and the Nucleosome, 199
- 9 The Replication of DNA, 257
- 10 The Mutability and Repair of DNA, 313
- 11 Homologous Recombination at the Molecular Level, 341
- 12 Site-Specific Recombination and Transposition of DNA, 377

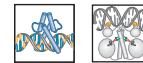
## PART 4



### EXPRESSION OF THE GENOME, 423

- 13 Mechanisms of Transcription, 429
- 14 RNA Splicing, 467
- 15 Translation, 509
- 16 The Genetic Code, 573
- 17 The Origin and Early Evolution of Life, 593

## PART 5



### REGULATION, 609

- 18 Transcriptional Regulation in Prokaryotes, 615
- 19 Transcriptional Regulation in Eukaryotes, 657
- 20 Regulatory RNAs, 701
- 21 Gene Regulation in Development and Evolution, 733
- 22 Systems Biology, 775

## PART 6



### APPENDICES, 793

- 1 Model Organisms, 797
- 2 Answers, 831

Index, 845

*This page intentionally left blank*

# Detailed Contents

## PART 1: HISTORY, 1



### 1 The Mendelian View of the World, 5

MENDEL'S DISCOVERIES, 6	
The Principle of Independent Segregation, 6	
<b>ADVANCED CONCEPTS BOX 1-1</b> <i>Mendelian Laws</i> , 6	
Some Alleles Are neither Dominant nor Recessive, 7	
Principle of Independent Assortment, 8	
CHROMOSOMAL THEORY OF HEREDITY, 8	
GENE LINKAGE AND CROSSING OVER, 9	
<b>KEY EXPERIMENTS BOX 1-2</b> <i>Genes Are Linked to Chromosomes</i> , 10	
CHROMOSOME MAPPING, 11	

THE ORIGIN OF GENETIC VARIABILITY THROUGH MUTATIONS, 13	
---	--

EARLY SPECULATIONS ABOUT WHAT GENES ARE AND HOW THEY ACT, 15	
PRELIMINARY ATTEMPTS TO FIND A GENE–PROTEIN RELATIONSHIP, 16	
SUMMARY, 17	
BIBLIOGRAPHY, 17	
QUESTIONS, 18	



### 2 Nucleic Acids Convey Genetic Information, 21

AVERY'S BOMBSHELL: DNA CAN CARRY GENETIC SPECIFICITY, 22	
Viral Genes Are Also Nucleic Acids, 23	
THE DOUBLE HELIX, 24	
<b>KEY EXPERIMENTS BOX 2-1</b> <i>Chargaff's Rules</i> , 26	
Finding the Polymerases That Make DNA, 26	
Experimental Evidence Favors Strand Separation during DNA Replication, 27	
THE GENETIC INFORMATION WITHIN DNA IS CONVEYED BY THE SEQUENCE OF ITS FOUR NUCLEOTIDE BUILDING BLOCKS, 30	
<b>KEY EXPERIMENTS BOX 2-2</b> <i>Evidence That Genes Control Amino Acid Sequences in Proteins</i> , 31	
DNA Cannot Be the Template That Directly Orders Amino Acids during Protein Synthesis, 32	
RNA Is Chemically Very Similar to DNA, 32	
THE CENTRAL DOGMA, 33	

The Adaptor Hypothesis of Crick, 34	
Discovery of Transfer RNA, 34	
The Paradox of the Nonspecific-Appearing Ribosomes, 35	
Discovery of Messenger RNA (mRNA), 35	
Enzymatic Synthesis of RNA upon DNA Templates, 35	
Establishing the Genetic Code, 37	

ESTABLISHING THE DIRECTION OF PROTEIN SYNTHESIS, 38	
Start and Stop Signals Are Also Encoded within DNA, 40	
THE ERA OF GENOMICS, 40	
SUMMARY, 41	
BIBLIOGRAPHY, 42	
QUESTIONS, 42	

## PART 2: STRUCTURE AND STUDY OF MACROMOLECULES, 45

---



### 3 The Importance of Weak and Strong Chemical Bonds, 51

CHARACTERISTICS OF CHEMICAL BONDS,	51
Chemical Bonds Are Explainable in Quantum-Mechanical Terms,	52
Chemical-Bond Formation Involves a Change in the Form of Energy,	53
Equilibrium between Bond Making and Breaking,	53
THE CONCEPT OF FREE ENERGY,	54
$K_{eq}$ Is Exponentially Related to $\Delta G$ ,	54
Covalent Bonds Are Very Strong,	54
WEAK BONDS IN BIOLOGICAL SYSTEMS,	55
Weak Bonds Have Energies between 1 and 7 kcal/mol,	55
Weak Bonds Are Constantly Made and Broken at Physiological Temperatures,	55
The Distinction between Polar and Nonpolar Molecules,	55
van der Waals Forces,	56
Hydrogen Bonds,	57
Some Ionic Bonds Are Hydrogen Bonds,	58
Weak Interactions Demand Complementary Molecular Surfaces,	58
Water Molecules Form Hydrogen Bonds,	59
Weak Bonds between Molecules in Aqueous Solutions,	59
Organic Molecules That Tend to Form Hydrogen Bonds Are Water Soluble,	60
Hydrophobic “Bonds” Stabilize Macromolecules,	60
<b>ADVANCED CONCEPTS BOX 3-1</b> <i>The Uniqueness of Molecular Shapes and the Concept of Selective Stickiness,</i>	61
The Advantage of $\Delta G$ between 2 and 5 kcal/mol,	62

Weak Bonds Attach Enzymes to Substrates, 62

Weak Bonds Mediate Most Protein–DNA and Protein–Protein Interactions, 62

#### HIGH-ENERGY BONDS, 63

MOLECULES THAT DONATE ENERGY ARE THERMODYNAMICALLY UNSTABLE, 63

ENZYME LOWER ACTIVATION ENERGIES IN BIOCHEMICAL REACTIONS, 65

#### FREE ENERGY IN BIOMOLECULES, 66

High-Energy Bonds Hydrolyze with Large Negative  $\Delta G$ , 66

#### HIGH-ENERGY BONDS IN BIOSYNTHETIC REACTIONS, 67

Peptide Bonds Hydrolyze Spontaneously, 68  
Coupling of Negative with Positive  $\Delta G$ , 69

#### ACTIVATION OF PRECURSORS IN GROUP TRANSFER REACTIONS, 69

ATP Versatility in Group Transfer, 70  
Activation of Amino Acids by Attachment of AMP, 70

Nucleic Acid Precursors Are Activated by the Presence of  $\text{P} \sim \text{P}$ , 71

The Value of  $\text{P} \sim \text{P}$  Release in Nucleic Acid Synthesis, 72

$\text{P} \sim \text{P}$  Splits Characterize Most Biosynthetic Reactions, 73

#### SUMMARY, 74

#### BIBLIOGRAPHY, 75

#### QUESTIONS, 75



### 4 The Structure of DNA, 77

DNA STRUCTURE,	78
DNA Is Composed of Polynucleotide Chains,	78
Each Base Has Its Preferred Tautomeric Form,	80
The Two Strands of the Double Helix Are Wound around Each Other in an Antiparallel Orientation,	81
The Two Chains of the Double Helix Have Complementary Sequences,	81

The Double Helix Is Stabilized by Base Pairing and Base Stacking, 82

Hydrogen Bonding Is Important for the Specificity of Base Pairing, 83

Bases Can Flip Out from the Double Helix, 83

DNA Is Usually a Right-Handed Double Helix, 83

**KEY EXPERIMENTS BOX 4-1** *DNA Has 10.5 bp per Turn of the Helix in Solution: The Mica Experiment,* 84

The Double Helix Has Minor and Major Grooves, 84
The Major Groove Is Rich in Chemical Information, 85
The Double Helix Exists in Multiple Conformations, 86
DNA Can Sometimes Form a Left-Handed Helix, 87
<b>KEY EXPERIMENTS BOX 4-2</b> How Spots on an X-Ray Film Reveal the Structure of DNA, 88
DNA Strands Can Separate (Denature) and Reassociate, 89
Some DNA Molecules Are Circles, 92
<b>DNA TOPOLOGY</b> , 93
Linking Number Is an Invariant Topological Property of Covalently Closed, Circular DNA, 93
Linking Number Is Composed of Twist and Writhe, 93
$Lk^o$ Is the Linking Number of Fully Relaxed cccDNA under Physiological Conditions, 94
DNA in Cells Is Negatively Supercoiled, 95
Nucleosomes Introduce Negative Supercoiling in Eukaryotes, 96



## 5 The Structure and Versatility of RNA, 107

RNA CONTAINS RIBOSE AND URACIL AND IS USUALLY SINGLE-STRANDED, 107
RNA CHAINS FOLD BACK ON THEMSELVES TO FORM LOCAL REGIONS OF DOUBLE HELIX SIMILAR TO A-FORM DNA, 108
RNA CAN FOLD UP INTO COMPLEX TERTIARY STRUCTURES, 110
NUCLEOTIDE SUBSTITUTIONS IN COMBINATION WITH CHEMICAL PROBING PREDICT RNA STRUCTURE, 111
<b>MEDICAL CONNECTIONS BOX 5-1</b> An RNA Switch Controls Protein Synthesis by Murine Leukemia Virus, 112



## 6 The Structure of Proteins, 121

THE BASICS, 121
Amino Acids, 121
The Peptide Bond, 122
Polypeptide Chains, 123
Three Amino Acids with Special Conformational Properties, 124
<b>ADVANCED CONCEPT BOX 6-1</b> Ramachandran Plot: Permitted Combinations of Main-Chain Torsion Angles $\phi$ and $\psi$ , 124

Topoisomerases Can Relax Supercoiled DNA, 97
Prokaryotes Have a Special Topoisomerase That Introduces Supercoils into DNA, 97
Topoisomerases Also Unknot and Disentangle DNA Molecules, 98
Topoisomerases Use a Covalent Protein–DNA Linkage to Cleave and Rejoin DNA Strands, 99
Topoisomerases Form an Enzyme Bridge and Pass DNA Segments through Each Other, 100
DNA Topoisomers Can Be Separated by Electrophoresis, 102
Ethidium Ions Cause DNA to Unwind, 102
<b>KEY EXPERIMENTS BOX 4-3</b> Proving that DNA Has a Helical Periodicity of $\sim 10.5$ bp per Turn from the Topological Properties of DNA Rings, 103
<b>SUMMARY</b> , 103
<b>BIBLIOGRAPHY</b> , 104
<b>QUESTIONS</b> , 104

DIRECTED EVOLUTION SELECTS RNAs THAT BIND SMALL MOLECULES, 114

SOME RNAs ARE ENZYMES, 114

**TECHNIQUES BOX 5-2** Creating an RNA Mimetic of the Green Fluorescent Protein by Directed Evolution, 115

The Hammerhead Ribozyme Cleaves RNA by the Formation of a 2', 3' Cyclic Phosphate, 116

A Ribozyme at the Heart of the Ribosome Acts on a Carbon Center, 118

**SUMMARY**, 118

**BIBLIOGRAPHY**, 118

**QUESTIONS**, 118

IMPORTANCE OF WATER, 125

PROTEIN STRUCTURE CAN BE DESCRIBED AT FOUR LEVELS, 126

PROTEIN DOMAINS, 130

Polypeptide Chains Typically Fold into One or More Domains, 130

**ADVANCED CONCEPTS BOX 6-2** Glossary of Terms, 130  
Basic Lessons from the Study of Protein Structures, 131

Classes of Protein Domains, 132	PROTEINS AS AGENTS OF SPECIFIC MOLECULAR RECOGNITION, 137
Linkers and Hinges, 133	Proteins That Recognize DNA Sequence, 137
Post-Translational Modifications, 133	Protein–Protein Interfaces, 140
<b>ADVANCED CONCEPTS BOX 6-3</b> <i>The Antibody Molecule as an Illustration of Protein Domains</i> , 133	Proteins That Recognize RNA, 141
FROM AMINO-ACID SEQUENCE TO THREE-DIMENSIONAL STRUCTURE, 134	ENZYMES: PROTEINS AS CATALYSTS, 141
Protein Folding, 134	REGULATION OF PROTEIN ACTIVITY, 142
<b>KEY EXPERIMENTS BOX 6-4</b> <i>Three-Dimensional Structure of a Protein Is Specified by Its Amino Acid Sequence (Anfinsen Experiment)</i> , 135	SUMMARY, 143
Predicting Protein Structure from Amino Acid Sequence, 135	BIBLIOGRAPHY, 144
CONFORMATIONAL CHANGES IN PROTEINS, 136	QUESTIONS, 144



## 7 Techniques of Molecular Biology, 147

NUCLEIC ACIDS: BASIC METHODS, 148	GENOMICS, 168
Gel Electrophoresis Separates DNA and RNA Molecules according to Size, 148	Bioinformatics Tools Facilitate the Genome-Wide Identification of Protein-Coding Genes, 169
Restriction Endonucleases Cleave DNA Molecules at Particular Sites, 149	Whole-Genome Tiling Arrays Are Used to Visualize the Transcriptome, 169
DNA Hybridization Can Be Used to Identify Specific DNA Molecules, 151	Regulatory DNA Sequences Can Be Identified by Using Specialized Alignment Tools, 171
Hybridization Probes Can Identify Electrophoretically Separated DNAs and RNAs, 151	Genome Editing Is Used to Precisely Alter Complex Genomes, 172
Isolation of Specific Segments of DNA, 153	PROTEINS, 173
DNA Cloning, 154	Specific Proteins Can Be Purified from Cell Extracts, 173
Vector DNA Can Be Introduced into Host Organisms by Transformation, 155	Purification of a Protein Requires a Specific Assay, 173
Libraries of DNA Molecules Can Be Created by Cloning, 156	Preparation of Cell Extracts Containing Active Proteins, 174
Hybridization Can Be Used to Identify a Specific Clone in a DNA Library, 156	Proteins Can Be Separated from One Another Using Column Chromatography, 174
Chemical Synthesis of Defined DNA Sequences, 157	Separation of Proteins on Polyacrylamide Gels, 176
The Polymerase Chain Reaction Amplifies DNAs by Repeated Rounds of DNA Replication In Vitro, 158	Antibodies Are Used to Visualize Electrophoretically Separated Proteins, 176
Nested Sets of DNA Fragments Reveal Nucleotide Sequences, 158	Protein Molecules Can Be Directly Sequenced, 177
<b>TECHNIQUES BOX 7-1</b> <i>Forensics and the Polymerase Chain Reaction</i> , 160	PROTEOMICS, 179
Shotgun Sequencing a Bacterial Genome, 162	Combining Liquid Chromatography with Mass Spectrometry Identifies Individual Proteins within a Complex Extract, 179
The Shotgun Strategy Permits a Partial Assembly of Large Genome Sequences, 162	Proteome Comparisons Identify Important Differences between Cells, 181
<b>KEY EXPERIMENTS BOX 7-2</b> <i>Sequenators Are Used for High-Throughput Sequencing</i> , 163	Mass Spectrometry Can Also Monitor Protein Modification States, 181
The Paired-End Strategy Permits the Assembly of Large-Genome Scaffolds, 165	Protein–Protein Interactions Can Yield Information regarding Protein Function, 182
The \$1000 Human Genome Is within Reach, 167	

NUCLEIC ACID–PROTEIN INTERACTIONS, 182
The Electrophoretic Mobility of DNA Is Altered by Protein Binding, 183
DNA-Bound Protein Protects the DNA from Nucleases and Chemical Modification, 184
Chromatin Immunoprecipitation Can Detect Protein Association with DNA in the Cell, 185

Chromosome Conformation Capture Assays Are Used to Analyze Long-Range Interactions, 187
In Vitro Selection Can Be Used to Identify a Protein’s DNA- or RNA-Binding Site, 189
BIBLIOGRAPHY, 190
QUESTIONS, 190

## PART 3: MAINTENANCE OF THE GENOME, 193

---



### 8 Genome Structure, Chromatin, and the Nucleosome, 199

#### GENOME SEQUENCE AND CHROMOSOME DIVERSITY, 200

- Chromosomes Can Be Circular or Linear, 200
- Every Cell Maintains a Characteristic Number of Chromosomes, 201
- Genome Size Is Related to the Complexity of the Organism, 202
- The *E. coli* Genome Is Composed Almost Entirely of Genes, 203
- More Complex Organisms Have Decreased Gene Density, 204
- Genes Make Up Only a Small Proportion of the Eukaryotic Chromosomal DNA, 205
- The Majority of Human Intergenic Sequences Are Composed of Repetitive DNA, 207

#### CHROMOSOME DUPLICATION AND SEGREGATION, 208

- Eukaryotic Chromosomes Require Centromeres, Telomeres, and Origins of Replication to Be Maintained during Cell Division, 208
- Eukaryotic Chromosome Duplication and Segregation Occur in Separate Phases of the Cell Cycle, 210
- Chromosome Structure Changes as Eukaryotic Cells Divide, 212
- Sister-Chromatid Cohesion and Chromosome Condensation Are Mediated by SMC Proteins, 214
- Mitosis Maintains the Parental Chromosome Number, 214
- During Gap Phases, Cells Prepare for the Next Cell Cycle Stage and Check That the Previous Stage Is Completed Correctly, 217
- Meiosis Reduces the Parental Chromosome Number, 217
- Different Levels of Chromosome Structure Can Be Observed by Microscopy, 219

#### THE NUCLEOSOME, 220

#### Nucleosomes Are the Building Blocks of Chromosomes, 220

- Histones Are Small, Positively Charged Proteins, 221
- The Atomic Structure of the Nucleosome, 224
- Histones Bind Characteristic Regions of DNA within the Nucleosome, 224
- KEY EXPERIMENTS BOX 8-1** *Micrococcal Nuclease and the DNA Associated with the Nucleosome*, 226
- Many DNA Sequence–Independent Contacts Mediate the Interaction between the Core Histones and DNA, 227
- The Histone Amino-Terminal Tails Stabilize DNA Wrapping around the Octamer, 227
- Wrapping of the DNA around the Histone Protein Core Stores Negative Superhelicity, 228

#### HIGHER-ORDER CHROMATIN STRUCTURE, 229

- Heterochromatin and Euchromatin, 229
- KEY EXPERIMENTS BOX 8-2** *Nucleosomes and Superhelical Density*, 230
- Histone H1 Binds to the Linker DNA between Nucleosomes, 232
- Nucleosome Arrays Can Form More Complex Structures: The 30-nm Fiber, 232
- The Histone Amino-Terminal Tails Are Required for the Formation of the 30-nm Fiber, 234
- Further Compaction of DNA Involves Large Loops of Nucleosomal DNA, 234
- Histone Variants Alter Nucleosome Function, 234

#### REGULATION OF CHROMATIN STRUCTURE, 236

- The Interaction of DNA with the Histone Octamer Is Dynamic, 236
- Nucleosome-Remodeling Complexes Facilitate Nucleosome Movement, 237
- Some Nucleosomes Are Found in Specific Positions: Nucleosome Positioning, 240

The Amino-Terminal Tails of the Histones Are Frequently Modified, 241
Protein Domains in Nucleosome-Remodeling and -Modifying Complexes Recognize Modified Histones, 244
<b>KEY EXPERIMENTS BOX 8-3</b> Determining Nucleosome Position in the Cell, 245
Specific Enzymes Are Responsible for Histone Modification, 248
Nucleosome Modification and Remodeling Work Together to Increase DNA Accessibility, 249

NUCLEOSOME ASSEMBLY, 249
Nucleosomes Are Assembled Immediately after DNA Replication, 249
Assembly of Nucleosomes Requires Histone "Chaperones", 253
SUMMARY, 254
BIBLIOGRAPHY, 255
QUESTIONS, 255



## 9 The Replication of DNA, 257

THE CHEMISTRY OF DNA SYNTHESIS, 258
DNA Synthesis Requires Deoxynucleoside Triphosphates and a Primer:Template Junction, 258
DNA Is Synthesized by Extending the 3' End of the Primer, 259
Hydrolysis of Pyrophosphate Is the Driving Force for DNA Synthesis, 260
THE MECHANISM OF DNA POLYMERASE, 260
DNA Polymerases Use a Single Active Site to Catalyze DNA Synthesis, 260
<b>TECHNIQUES BOX 9-1</b> Incorporation Assays Can Be Used to Measure Nucleic Acid and Protein Synthesis, 261
DNA Polymerases Resemble a Hand That Grips the Primer:Template Junction, 263
DNA Polymerases Are Processive Enzymes, 265
Exonucleases Proofread Newly Synthesized DNA, 267
<b>MEDICAL CONNECTIONS BOX 9-2</b> Anticancer and Antiviral Agents Target DNA Replication, 268
THE REPLICATION FORK, 269
Both Strands of DNA Are Synthesized Together at the Replication Fork, 269
The Initiation of a New Strand of DNA Requires an RNA Primer, 270
RNA Primers Must Be Removed to Complete DNA Replication, 271
DNA Helicases Unwind the Double Helix in Advance of the Replication Fork, 272
DNA Helicase Pulls Single-Stranded DNA through a Central Protein Pore, 273
Single-Stranded DNA-Binding Proteins Stabilize ssDNA before Replication, 273
Topoisomerases Remove Supercoils Produced by DNA Unwinding at the Replication Fork, 275
Replication Fork Enzymes Extend the Range of DNA Polymerase Substrates, 275

## THE SPECIALIZATION OF DNA POLYMERASES, 277

DNA Polymerases Are Specialized for Different Roles in the Cell, 277
Sliding Clamps Dramatically Increase DNA Polymerase Processivity, 278
Sliding Clamps Are Opened and Placed on DNA by Clamp Loaders, 281
<b>ADVANCED CONCEPTS BOX 9-3</b> ATP Control of Protein Function: Loading a Sliding Clamp, 282

## DNA SYNTHESIS AT THE REPLICATION FORK, 283

Interactions between Replication Fork Proteins Form the *E. coli* Replisome, 286

## INITIATION OF DNA REPLICATION, 288

Specific Genomic DNA Sequences Direct the Initiation of DNA Replication, 288
The Replicon Model of Replication Initiation, 288
Replicator Sequences Include Initiator-Binding Sites and Easily Unwound DNA, 289

## **KEY EXPERIMENTS BOX 9-4** The Identification of Origins of Replication and Replicators, 290

## BINDING AND UNWINDING: ORIGIN SELECTION AND ACTIVATION BY THE INITIATOR PROTEIN, 293

Protein–Protein and Protein–DNA Interactions Direct the Initiation Process, 293
<b>ADVANCED CONCEPTS BOX 9-5</b> <i>E. coli</i> DNA Replication Is Regulated by DnaA-ATP Levels and SeqA, 294
Eukaryotic Chromosomes Are Replicated Exactly Once per Cell Cycle, 297
Helicase Loading Is the First Step in the Initiation of Replication in Eukaryotes, 298
Helicase Loading and Activation Are Regulated to Allow Only a Single Round of Replication during Each Cell Cycle, 300

Similarities between Eukaryotic and Prokaryotic DNA Replication Initiation, 301	
<b>FINISHING REPLICATION, 302</b>	
Type II Topoisomerases Are Required to Separate Daughter DNA Molecules, 303	
Lagging-Strand Synthesis Is Unable to Copy the Extreme Ends of Linear Chromosomes, 303	
Telomerase Is a Novel DNA Polymerase That Does Not Require an Exogenous Template, 305	
Telomerase Solves the End Replication Problem by Extending the 3' End of the Chromosome, 305	

## 10 The Mutability and Repair of DNA, 313

<b>REPLICATION ERRORS AND THEIR REPAIR, 314</b>	
The Nature of Mutations, 314	
Some Replication Errors Escape Proofreading, 315	
<b>MEDICAL CONNECTIONS BOX 10-1 Expansion of Triple Repeats Causes Disease, 316</b>	
Mismatch Repair Removes Errors That Escape Proofreading, 316	
<b>DNA DAMAGE, 320</b>	
DNA Undergoes Damage Spontaneously from Hydrolysis and Deamination, 320	
<b>MEDICAL CONNECTIONS BOX 10-2 The Ames Test, 321</b>	
DNA Is Damaged by Alkylation, Oxidation, and Radiation, 322	
<b>ADVANCED CONCEPTS BOX 10-3 Quantitation of DNA Damage and Its Effects on Cellular Survival and Mutagenesis, 323</b>	
Mutations Are Also Caused by Base Analogs and Intercalating Agents, 323	
<b>REPAIR AND TOLERANCE OF DNA DAMAGE, 324</b>	
Direct Reversal of DNA Damage, 325	

## 11 Homologous Recombination at the Molecular Level, 341

<b>DNA BREAKS ARE COMMON AND INITIATE RECOMBINATION, 342</b>	
<b>MODELS FOR HOMOLOGOUS RECOMBINATION, 342</b>	
Strand Invasion Is a Key Early Step in Homologous Recombination, 344	
Resolving Holliday Junctions Is a Key Step to Finishing Genetic Exchange, 346	
The Double-Strand Break–Repair Model Describes Many Recombination Events, 346	

<b>MEDICAL CONNECTIONS BOX 9-6 Aging, Cancer, and the Telomere Hypothesis, 307</b>	
Telomere-Binding Proteins Regulate Telomerase Activity and Telomere Length, 307	
Telomere-Binding Proteins Protect Chromosome Ends, 308	
<b>SUMMARY, 310</b>	
<b>BIBLIOGRAPHY, 311</b>	
<b>QUESTIONS, 312</b>	

Base Excision Repair Enzymes Remove Damaged Bases by a Base-Flipping Mechanism, 326	
Nucleotide Excision Repair Enzymes Cleave Damaged DNA on Either Side of the Lesion, 328	
<b>MEDICAL CONNECTIONS BOX 10-4 Linking Nucleotide Excision Repair and Translesion Synthesis to a Genetic Disorder in Humans, 330</b>	
Recombination Repairs DNA Breaks by Retrieving Sequence Information from Undamaged DNA, 330	
DSBs in DNA Are Also Repaired by Direct Joining of Broken Ends, 331	
<b>MEDICAL CONNECTIONS BOX 10-5 Nonhomologous End Joining, 332</b>	
Translesion DNA Synthesis Enables Replication to Proceed across DNA Damage, 333	
<b>ADVANCED CONCEPTS BOX 10-6 The Y Family of DNA Polymerases, 336</b>	
<b>SUMMARY, 338</b>	
<b>BIBLIOGRAPHY, 338</b>	
<b>QUESTIONS, 339</b>	

<b>HOMOLOGOUS RECOMBINATION PROTEIN MACHINES, 349</b>	
---	--

<b>ADVANCED CONCEPTS BOX 11-1 How to Resolve a Recombination Intermediate with Two Holliday Junctions, 350</b>	
The RecBCD Helicase/Nuclease Processes Broken DNA Molecules for Recombination, 351	
Chi Sites Control RecBCD, 354	
RecA Protein Assembles on Single-Stranded DNA and Promotes Strand Invasion, 355	

Newly Base-Paired Partners Are Established within the RecA Filament, 356	<b>MEDICAL CONNECTIONS BOX 11-2</b> <i>The Product of the Tumor Suppressor Gene BRCA2 Interacts with Rad51 Protein and Controls Genome Stability, 367</i>
RecA Homologs Are Present in All Organisms, 359	<b>MEDICAL CONNECTIONS BOX 11-3</b> <i>Proteins Associated with Premature Aging and Cancer Promote an Alternative Pathway for Holliday Junction Processing, 368</i>
The RuvAB Complex Specifically Recognizes Holliday Junctions and Promotes Branch Migration, 359	
RuvC Cleaves Specific DNA Strands at the Holliday Junction to Finish Recombination, 361	
<b>HOMOLOGOUS RECOMBINATION IN EUKARYOTES, 362</b>	
Homologous Recombination Has Additional Functions in Eukaryotes, 362	MATING-TYPE SWITCHING, 369
Homologous Recombination Is Required for Chromosome Segregation during Meiosis, 362	Mating-Type Switching Is Initiated by a Site-Specific Double-Strand Break, 370
Programmed Generation of Double-Stranded DNA Breaks Occurs during Meiosis, 363	Mating-Type Switching Is a Gene Conversion Event and Not Associated with Crossing Over, 370
MRX Protein Processes the Cleaved DNA Ends for Assembly of the RecA-Like Strand-Exchange Proteins, 364	<b>GENETIC CONSEQUENCES OF THE MECHANISM OF HOMOLOGOUS RECOMBINATION, 371</b>
Dmc1 Is a RecA-Like Protein That Specifically Functions in Meiotic Recombination, 366	One Cause of Gene Conversion Is DNA Repair during Recombination, 373
Many Proteins Function Together to Promote Meiotic Recombination, 366	<b>SUMMARY, 374</b>
	<b>BIBLIOGRAPHY, 375</b>
	<b>QUESTIONS, 376</b>



## 12 Site-Specific Recombination and Transposition of DNA, 377

<b>CONSERVATIVE SITE-SPECIFIC RECOMBINATION, 378</b>	The Hin Recombinase Inverts a Segment of DNA Allowing Expression of Alternative Genes, 389
Site-Specific Recombination Occurs at Specific DNA Sequences in the Target DNA, 378	Hin Recombination Requires a DNA Enhancer, 390
Site-Specific Recombinases Cleave and Rejoin DNA Using a Covalent Protein–DNA Intermediate, 380	Recombinases Convert Multimeric Circular DNA Molecules into Monomers, 391
Serine Recombinases Introduce Double-Strand Breaks in DNA and Then Swap Strands to Promote Recombination, 382	There Are Other Mechanisms to Direct Recombination to Specific Segments of DNA, 391
Structure of the Serine Recombinase–DNA Complex Indicates that Subunits Rotate to Achieve Strand Exchange, 383	<b>ADVANCED CONCEPTS BOX 12-2</b> <i>The Xer Recombinase Catalyzes the Monomerization of Bacterial Chromosomes and of Many Bacterial Plasmids, 392</i>
Tyrosine Recombinases Break and Rejoin One Pair of DNA Strands at a Time, 383	<b>TRANSPOSITION, 393</b>
Structures of Tyrosine Recombinases Bound to DNA Reveal the Mechanism of DNA Exchange, 384	Some Genetic Elements Move to New Chromosomal Locations by Transposition, 393
<b>MEDICAL CONNECTIONS BOX 12-1</b> <i>Application of Site-Specific Recombination to Genetic Engineering, 386</i>	There Are Three Principal Classes of Transposable Elements, 395
<b>BIOLOGICAL ROLES OF SITE-SPECIFIC RECOMBINATION, 386</b>	DNA Transposons Carry a Transposase Gene, Flanked by Recombination Sites, 395
λ Integrase Promotes the Integration and Excision of a Viral Genome into the Host-Cell Chromosome, 386	Transposons Exist as Both Autonomous and Nonautonomous Elements, 396
Bacteriophage λ Excision Requires a New DNA-Bending Protein, 389	Virus-Like Retrotransposons and Retroviruses Carry Terminal Repeat Sequences and Two Genes Important for Recombination, 396
	Poly-A Retrotransposons Look Like Genes, 396
	DNA Transposition by a Cut-and-Paste Mechanism, 397

The Intermediate in Cut-and-Paste Transposition is Finished by Gap Repair, 398	Phage Mu Is an Extremely Robust Transposon, 411
There Are Multiple Mechanisms for Cleaving the Nontransferred Strand during DNA Transposition, 399	Mu Uses Target Immunity to Avoid Transposing into Its Own DNA, 411
DNA Transposition by a Replicative Mechanism, 401	Tc1/ <i>mariner</i> Elements Are Highly Successful DNA Elements in Eukaryotes, 411
Virus-Like Retrotransposons and Retroviruses Move Using an RNA Intermediate, 403	<b>ADVANCED CONCEPTS BOX 12-4</b> Mechanism of Transposition Target Immunity, 413
DNA Transposases and Retroviral Integrases Are Members of a Protein Superfamily, 403	Yeast Ty Elements Transpose into Safe Havens in the Genome, 414
Poly-A Retrotransposons Move by a “Reverse Splicing” Mechanism, 405	LINEs Promote Their Own Transposition and Even Transpose Cellular RNAs, 414
EXAMPLES OF TRANSPOSSABLE ELEMENTS AND THEIR REGULATION, 406	V(D)J RECOMBINATION, 416
<b>KEY EXPERIMENTS BOX 12-3</b> Maize Elements and Discovery of Transposons, 408	The Early Events in V(D)J Recombination Occur by a Mechanism Similar to Transposon Excision, 418
IS4 Family Transposons Are Compact Elements with Multiple Mechanisms for Copy Number Control, 409	SUMMARY, 420
	BIBLIOGRAPHY, 420
	QUESTIONS, 421

## PART 4: EXPRESSION OF THE GENOME, 423

---



### 13 Mechanisms of Transcription, 429

RNA POLYMERASES AND THE TRANSCRIPTION CYCLE, 430	The Elongating Polymerase Is a Processive Machine That Synthesizes and Proofreads RNA, 442
RNA Polymerases Come in Different Forms but Share Many Features, 430	<b>ADVANCED CONCEPTS BOX 13-2</b> The Single-Subunit RNA Polymerases, 443
Transcription by RNA Polymerase Proceeds in a Series of Steps, 432	RNA Polymerase Can Become Arrested and Need Removing, 445
Transcription Initiation Involves Three Defined Steps, 434	Transcription Is Terminated by Signals within the RNA Sequence, 445
THE TRANSCRIPTION CYCLE IN BACTERIA, 434	TRANSCRIPTION IN EUKARYOTES, 448
Bacterial Promoters Vary in Strength and Sequence but Have Certain Defining Features, 434	RNA Polymerase II Core Promoters Are Made Up of Combinations of Different Classes of Sequence Element, 448
<b>TECHNIQUES BOX 13-1</b> Consensus Sequences, 436	RNA Polymerase II Forms a Preinitiation Complex with General Transcription Factors at the Promoter, 449
The $\sigma$ Factor Mediates Binding of Polymerase to the Promoter, 437	Promoter Escape Requires Phosphorylation of the Polymerase “Tail,” 449
Transition to the Open Complex Involves Structural Changes in RNA Polymerase and in the Promoter DNA, 438	TBP Binds to and Distorts DNA Using a $\beta$ Sheet Inserted into the Minor Groove, 451
Transcription Is Initiated by RNA Polymerase without the Need for a Primer, 440	The Other General Transcription Factors Also Have Specific Roles in Initiation, 452
During Initial Transcription, RNA Polymerase Remains Stationary and Pulls Downstream DNA into Itself, 441	In Vivo, Transcription Initiation Requires Additional Proteins, Including the Mediator Complex, 453
Promoter Escape Involves Breaking Polymerase–Promoter Interactions and Polymerase Core– $\sigma$ Interactions, 442	

The Elongating Polymerase Is a Processive Machine That Synthesizes and Proofreads RNA, 442

**ADVANCED CONCEPTS BOX 13-2** The Single-Subunit RNA Polymerases, 443

RNA Polymerase Can Become Arrested and Need Removing, 445

Transcription Is Terminated by Signals within the RNA Sequence, 445

#### TRANSCRIPTION IN EUKARYOTES, 448

RNA Polymerase II Core Promoters Are Made Up of Combinations of Different Classes of Sequence Element, 448

RNA Polymerase II Forms a Preinitiation Complex with General Transcription Factors at the Promoter, 449

Promoter Escape Requires Phosphorylation of the Polymerase “Tail,” 449

TBP Binds to and Distorts DNA Using a  $\beta$  Sheet Inserted into the Minor Groove, 451

The Other General Transcription Factors Also Have Specific Roles in Initiation, 452

In Vivo, Transcription Initiation Requires Additional Proteins, Including the Mediator Complex, 453

Mediator Consists of Many Subunits, Some Conserved from Yeast to Human, 454	TRANSCRIPTION BY RNA POLYMERASES I AND III, 462
A New Set of Factors Stimulates Pol II Elongation and RNA Proofreading, 455	RNA Pol I and Pol III Recognize Distinct Promoters but Still Require TBP, 462
Elongating RNA Polymerase Must Deal with Histones in Its Path, 456	Pol I Transcribes Just the rRNA Genes, 462
Elongating Polymerase Is Associated with a New Set of Protein Factors Required for Various Types of RNA Processing, 457	Pol III Promoters Are Found Downstream from the Transcription Start Site, 463
Transcription Termination Is Linked to RNA Destruction by a Highly Processive RNase, 460	SUMMARY, 463
	BIBLIOGRAPHY, 464
	QUESTIONS, 465



## 14 RNA Splicing, 467

THE CHEMISTRY OF RNA SPLICING, 469	Several Mechanisms Exist to Ensure Mutually Exclusive Splicing, 486
Sequences within the RNA Determine Where Splicing Occurs, 469	The Curious Case of the <i>Drosophila Dscam</i> Gene: Mutually Exclusive Splicing on a Grand Scale, 487
The Intron Is Removed in a Form Called a Lariat as the Flanking Exons Are Joined, 470	Mutually Exclusive Splicing of <i>Dscam</i> Exon 6 Cannot Be Accounted for by Any Standard Mechanism and Instead Uses a Novel Strategy, 488
<b>KEY EXPERIMENTS BOX 14-1</b> Adenovirus and the Discovery of Splicing, 471	<b>KEY EXPERIMENTS BOX 14-3</b> Identification of Docking Site and Selector Sequences, 490
THE SPLICEOSOME MACHINERY, 473	Alternative Splicing Is Regulated by Activators and Repressors, 491
RNA Splicing Is Performed by a Large Complex Called the Spliceosome, 473	Regulation of Alternative Splicing Determines the Sex of Flies, 493
SPLICING PATHWAYS, 474	An Alternative Splicing Switch Lies at the Heart of Pluripotency, 495
Assembly, Rearrangements, and Catalysis within the Spliceosome: The Splicing Pathway, 474	EXON SHUFFLING, 497
Spliceosome Assembly Is Dynamic and Variable and Its Disassembly Ensures That the Splicing Reaction Goes Only Forward in the Cell, 476	Exons Are Shuffled by Recombination to Produce Genes Encoding New Proteins, 497
Self-Splicing Introns Reveal That RNA Can Catalyze RNA Splicing, 477	<b>MEDICAL CONNECTIONS BOX 14-4</b> Defects in Pre-mRNA Splicing Cause Human Disease, 497
Group I Introns Release a Linear Intron Rather Than a Lariat, 478	RNA EDITING, 500
<b>KEY EXPERIMENTS BOX 14-2</b> Converting Group I Introns into Ribozymes, 479	RNA Editing Is Another Way of Altering the Sequence of an mRNA, 500
How Does the Spliceosome Find the Splice Sites Reliably?, 480	Guide RNAs Direct the Insertion and Deletion of Uridines, 501
VARIANTS OF SPLICING, 482	<b>MEDICAL CONNECTIONS BOX 14-5</b> Deaminases and HIV, 503
Exons from Different RNA Molecules Can Be Fused by <i>Trans</i> -Splicing, 482	mRNA TRANSPORT, 503
A Small Group of Introns Is Spliced by an Alternative Spliceosome Composed of a Different Set of snRNPs, 483	Once Processed, mRNA Is Packaged and Exported from the Nucleus into the Cytoplasm for Translation, 503
ALTERNATIVE SPLICING, 483	SUMMARY, 505
Single Genes Can Produce Multiple Products by Alternative Splicing, 483	BIBLIOGRAPHY, 506
	QUESTIONS, 507



## 15 Translation, 509

MESSENGER RNA, 510	INITIATION OF TRANSLATION, 528
Polypeptide Chains Are Specified by Open Reading Frames, 510	Prokaryotic mRNAs Are Initially Recruited to the Small Subunit by Base Pairing to rRNA, 528
Prokaryotic mRNAs Have a Ribosome-Binding Site That Recruits the Translational Machinery, 512	A Specialized tRNA Charged with a Modified Methionine Binds Directly to the Prokaryotic Small Subunit, 528
Eukaryotic mRNAs Are Modified at their 5' and 3' Ends to Facilitate Translation, 512	Three Initiation Factors Direct the Assembly of an Initiation Complex That Contains mRNA and the Initiator tRNA, 529
TRANSFER RNA, 513	Eukaryotic Ribosomes Are Recruited to the mRNA by the 5' Cap, 530
tRNAs Are Adaptors between Codons and Amino Acids, 513	Translation Initiation Factors Hold Eukaryotic mRNAs in Circles, 532
<b>ADVANCED CONCEPTS BOX 15-1 CCA-Adding Enzymes: Synthesizing RNA without a Template, 513</b>	<b>ADVANCED CONCEPTS BOX 15-3 uORFs and IRESS: Exceptions That Prove the Rule, 533</b>
tRNAs Share a Common Secondary Structure That Resembles a Cloverleaf, 514	The Start Codon Is Found by Scanning Downstream from the 5' End of the mRNA, 535
tRNAs Have an L-Shaped Three-Dimensional Structure, 514	
ATTACHMENT OF AMINO ACIDS TO tRNA, 515	TRANSLATION ELONGATION, 535
tRNAs Are Charged by the Attachment of an Amino Acid to the 3'-Terminal Adenosine Nucleotide via a High-Energy Acyl Linkage, 515	Aminoacyl-tRNAs Are Delivered to the A-Site by Elongation Factor EF-Tu, 537
Aminoacyl-tRNA Synthetases Charge tRNAs in Two Steps, 515	The Ribosome Uses Multiple Mechanisms to Select against Incorrect Aminoacyl-tRNAs, 537
Each Aminoacyl-tRNA Synthetase Attaches a Single Amino Acid to One or More tRNAs, 515	The Ribosome Is a Ribozyme, 538
tRNA Synthetases Recognize Unique Structural Features of Cognate tRNAs, 517	Peptide-Bond Formation Initiates Translocation in the Large Subunit, 541
Aminoacyl-tRNA Formation Is Very Accurate, 518	EF-G Drives Translocation by Stabilizing Intermediates in Translocation, 542
Some Aminoacyl-tRNA Synthetases Use an Editing Pocket to Charge tRNAs with High Accuracy, 518	EF-Tu–GDP and EF-G–GDP Must Exchange GDP for GTP before Participating in a New Round of Elongation, 543
The Ribosome Is Unable to Discriminate between Correctly and Incorrectly Charged tRNAs, 519	A Cycle of Peptide-Bond Formation Consumes Two Molecules of GTP and One Molecule of ATP, 543
THE RIBOSOME, 519	TERMINATION OF TRANSLATION, 544
<b>ADVANCED CONCEPTS BOX 15-2 Selenocysteine, 520</b>	Release Factors Terminate Translation in Response to Stop Codons, 544
The Ribosome Is Composed of a Large and a Small Subunit, 521	Short Regions of Class I Release Factors Recognize Stop Codons and Trigger Release of the Peptidyl Chain, 544
The Large and Small Subunits Undergo Association and Dissociation during Each Cycle of Translation, 522	<b>ADVANCED CONCEPTS BOX 15-4 GTP-Binding Proteins, Conformational Switching, and the Fidelity and Ordering of the Events of Translation, 546</b>
New Amino Acids Are Attached to the Carboxyl Terminus of the Growing Polypeptide Chain, 523	GDP/GTP Exchange and GTP Hydrolysis Control the Function of the Class II Release Factor, 547
Peptide Bonds Are Formed by Transfer of the Growing Polypeptide Chain from One tRNA to Another, 524	The Ribosome Recycling Factor Mimics a tRNA, 548
Ribosomal RNAs Are Both Structural and Catalytic Determinants of the Ribosome, 524	REGULATION OF TRANSLATION, 549
The Ribosome Has Three Binding Sites for tRNA, 525	Protein or RNA Binding near the Ribosome-Binding Site Negatively Regulates Bacterial Translation Initiation, 549
Channels through the Ribosome Allow the mRNA and Growing Polypeptide to Enter and/or Exit the Ribosome, 527	Regulation of Prokaryotic Translation: Ribosomal Proteins Are Translational Repressors of Their Own Synthesis, 551

**MEDICAL CONNECTIONS BOX 15-5** Antibiotics Arrest Cell Division by Blocking Specific Steps in Translation, 552

Global Regulators of Eukaryotic Translation Target Key Factors Required for mRNA Recognition and Initiator tRNA Ribosome Binding, 556

Spatial Control of Translation by mRNA-Specific 4E-BPs, 556

An Iron-Regulated, RNA-Binding Protein Controls Translation of Ferritin, 557

Translation of the Yeast Transcriptional Activator Gcn4 Is Controlled by Short Upstream ORFs and Ternary Complex Abundance, 558

**TECHNIQUES BOX 15-6** Ribosome and Polysome Profiling, 561

TRANSLATION-DEPENDENT REGULATION OF mRNA AND PROTEIN STABILITY, 563

The SsrA RNA Rescues Ribosomes That Translate Broken mRNAs, 563

**MEDICAL CONNECTIONS BOX 15-7** A Frontline Drug in Tuberculosis Therapy Targets SsrA Tagging, 565

Eukaryotic Cells Degrade mRNAs That Are Incomplete or Have Premature Stop Codons, 565

SUMMARY, 567

BIBLIOGRAPHY, 570

QUESTIONS, 570



## 16 The Genetic Code, 573

THE CODE IS DEGENERATE, 573

Perceiving Order in the Makeup of the Code, 575

Wobble in the Anticodon, 575

Three Codons Direct Chain Termination, 577

How the Code Was Cracked, 577

Stimulation of Amino Acid Incorporation by Synthetic mRNAs, 578

Poly-U Codes for Polyphenylalanine, 579

Mixed Copolymers Allowed Additional Codon Assignments, 579

Transfer RNA Binding to Defined Trinucleotide Codons, 579

Codon Assignments from Repeating Copolymers, 581

THREE RULES GOVERN THE GENETIC CODE, 582

Three Kinds of Point Mutations Alter the Genetic Code, 582

Genetic Proof That the Code Is Read in Units of Three, 583

SUPPRESSOR MUTATIONS CAN RESIDE IN THE SAME OR A DIFFERENT GENE, 584

Intergenic Suppression Involves Mutant tRNAs, 584

Nonsense Suppressors Also Read Normal Termination Signals, 585

Proving the Validity of the Genetic Code, 586

THE CODE IS NEARLY UNIVERSAL, 587

**ADVANCED CONCEPTS BOX 16-1** Expanding the Genetic Code, 589

SUMMARY, 590

BIBLIOGRAPHY, 590

QUESTIONS, 591



## 17 The Origin and Early Evolution of Life, 593

WHEN DID LIFE ARISE ON EARTH?, 594

WHAT WAS THE BASIS FOR PREBIOTIC ORGANIC CHEMISTRY?, 595

DID LIFE EVOLVE FROM AN RNA WORLD?, 599

CAN SELF-REPLICATING RIBOZYMES BE CREATED BY DIRECTED EVOLUTION?, 599

DOES DARWINIAN EVOLUTION REQUIRE SELF-REPLICATING PROTOCELLS?, 603

DID LIFE ARISE ON EARTH?, 606

SUMMARY, 607

BIBLIOGRAPHY, 607

QUESTIONS, 607

## PART 5: REGULATION, 609

---



### 18 Transcriptional Regulation in Prokaryotes, 615

#### PRINCIPLES OF TRANSCRIPTIONAL REGULATION, 615

- Gene Expression Is Controlled by Regulatory Proteins, 615
- Most Activators and Repressors Act at the Level of Transcription Initiation, 616
- Many Promoters Are Regulated by Activators That Help RNA Polymerase Bind DNA and by Repressors That Block That Binding, 616
- Some Activators and Repressors Work by Allostery and Regulate Steps in Transcriptional Initiation after RNA Polymerase Binding, 618
- Action at a Distance and DNA Looping, 618
- Cooperative Binding and Allostery Have Many Roles in Gene Regulation, 619
- Antitermination and Beyond: Not All of Gene Regulation Targets Transcription Initiation, 620

#### REGULATION OF TRANSCRIPTION INITIATION: EXAMPLES FROM PROKARYOTES, 620

- An Activator and a Repressor Together Control the *lac* Genes, 620
- CAP and Lac Repressor Have Opposing Effects on RNA Polymerase Binding to the *lac* Promoter, 622
- CAP Has Separate Activating and DNA-Binding Surfaces, 622
- CAP and Lac Repressor Bind DNA Using a Common Structural Motif, 623
- KEY EXPERIMENTS BOX 18-1** Activator Bypass Experiments, 624
- The Activities of Lac Repressor and CAP Are Controlled Allosterically by Their Signals, 626
- Combinatorial Control: CAP Controls Other Genes As Well, 627
- KEY EXPERIMENTS BOX 18-2** Jacob, Monod, and the Ideas behind Gene Regulation, 628
- Alternative σ Factors Direct RNA Polymerase to Alternative Sets of Promoters, 630
- NtrC and MerR: Transcriptional Activators That Work by Allostery Rather than by Recruitment, 630
- NtrC Has ATPase Activity and Works from DNA Sites Far from the Gene, 631
- MerR Activates Transcription by Twisting Promoter DNA, 632

Some Repressors Hold RNA Polymerase at the Promoter Rather than Excluding It, 633

AraC and Control of the *araBAD* Operon by Antiactivation, 634

**MEDICAL CONNECTIONS 18-3** Blocking Virulence by Silencing Pathways of Intercellular Communication, 635

#### THE CASE OF BACTERIOPHAGE λ: LAYERS OF REGULATION, 636

- Alternative Patterns of Gene Expression Control Lytic and Lysogenic Growth, 636
- Regulatory Proteins and Their Binding Sites, 638
- λ Repressor Binds to Operator Sites Cooperatively, 639
- Repressor and Cro Bind in Different Patterns to Control Lytic and Lysogenic Growth, 640
- ADVANCED CONCEPTS BOX 18-4** Concentration, Affinity, and Cooperative Binding, 641
- Lysogenic Induction Requires Proteolytic Cleavage of λ Repressor, 642
- Negative Autoregulation of Repressor Requires Long-Distance Interactions and a Large DNA Loop, 643
- Another Activator, λ CII, Controls the Decision between Lytic and Lysogenic Growth upon Infection of a New Host, 644
- KEY EXPERIMENTS BOX 18-5** Evolution of the λ Switch, 645
- The Number of Phage Particles Infecting a Given Cell Affects Whether the Infection Proceeds Lytically or Lysogenically, 647
- Growth Conditions of *E. coli* Control the Stability of CII Protein and Thus the Lytic/Lysogenic Choice, 648
- Transcriptional Antitermination in λ Development, 648
- KEY EXPERIMENTS BOX 18-6** Genetic Approaches That Identified Genes Involved in the Lytic/Lysogenic Choice, 649
- Retroregulation: An Interplay of Controls on RNA Synthesis and Stability Determines *int* Gene Expression, 651
- SUMMARY, 652
- BIBLIOGRAPHY, 653
- QUESTIONS, 654



## 19 Transcriptional Regulation in Eukaryotes, 657

CONSERVED MECHANISMS OF TRANSCRIPTIONAL REGULATION FROM YEAST TO MAMMALS, 659

Activators Have Separate DNA-Binding and Activating Functions, 660

Eukaryotic Regulators Use a Range of DNA-Binding Domains, But DNA Recognition Involves the Same Principles as Found in Bacteria, 661

Activating Regions Are Not Well-Defined Structures, 663

**TECHNIQUES BOX 19-1** *The Two-Hybrid Assay*, 664

RECRUITMENT OF PROTEIN COMPLEXES TO GENES BY EUKARYOTIC ACTIVATORS, 665

Activators Recruit the Transcriptional Machinery to the Gene, 665

**TECHNIQUES BOX 19-2** *The ChIP-Chip and ChIP-Seq Assays Are the Best Method for Identifying Enhancers*, 666

Activators Also Recruit Nucleosome Modifiers That Help the Transcriptional Machinery Bind at the Promoter or Initiate Transcription, 667

Activators Recruit Additional Factors Needed for Efficient Initiation or Elongation at Some Promoters, 669

**MEDICAL CONNECTIONS BOX 19-3** *Histone Modifications, Transcription Elongation, and Leukemia*, 670

Action at a Distance: Loops and Insulators, 672

Appropriate Regulation of Some Groups of Genes Requires Locus Control Regions, 673

SIGNAL INTEGRATION AND COMBINATORIAL CONTROL, 675

Activators Work Synergistically to Integrate Signals, 675

Signal Integration: The *HO* Gene Is Controlled by Two Regulators—One Recruits Nucleosome Modifiers, and the Other Recruits Mediator, 675

Signal Integration: Cooperative Binding of Activators at the Human  $\beta$ -Interferon Gene, 676

Combinatorial Control Lies at the Heart of the Complexity and Diversity of Eukaryotes, 678

Combinatorial Control of the Mating-Type Genes from *S. cerevisiae*, 680

TRANSCRIPTIONAL REPRESSORS, 681

SIGNAL TRANSDUCTION AND THE CONTROL OF TRANSCRIPTIONAL REGULATORS, 682

Signals Are Often Communicated to Transcriptional Regulators through Signal Transduction Pathways, 682

**KEY EXPERIMENTS BOX 19-4** *Evolution of a Regulatory Circuit*, 683

Signals Control the Activities of Eukaryotic Transcriptional Regulators in a Variety of Ways, 686

GENE “SILENCING” BY MODIFICATION OF HISTONES AND DNA, 687

Silencing in Yeast Is Mediated by Deacetylation and Methylation of Histones, 688

In *Drosophila*, HP1 Recognizes Methylated Histones and Condenses Chromatin, 689

Repression by Polycomb Also Uses Histone Methylation, 690

**ADVANCED CONCEPTS BOX 19-5** *Is There a Histone Code?*, 691

DNA Methylation Is Associated with Silenced Genes in Mammalian Cells, 692

EPIGENETIC GENE REGULATION, 694

Some States of Gene Expression Are Inherited through Cell Division Even When the Initiating Signal Is No Longer Present, 694

**MEDICAL CONNECTIONS BOX 19-6** *Transcriptional Repression and Human Disease*, 696

SUMMARY, 697

BIBLIOGRAPHY, 698

QUESTIONS, 699



## 20 Regulatory RNAs, 701

REGULATION BY RNAs IN BACTERIA, 701

Riboswitches Reside within the Transcripts of Genes Whose Expression They Control through Changes in Secondary Structure, 703

RNAs as Defense Agents in Prokaryotes and Archaea, 705

CRISPRs Are a Record of Infections Survived and Resistance Gained, 706

**ADVANCED CONCEPTS BOX 20-1** *Amino Acid Biosynthetic Operons Are Controlled by Attenuation*, 707

Spacer Sequences Are Acquired from Infecting Viruses, 710

A CRISPR Is Transcribed as a Single Long RNA, Which Is Then Processed into Shorter RNA Species That Target Destruction of Invading DNA or RNA, 710

## REGULATORY RNAs ARE WIDESPREAD IN EUKARYOTES, 711

Short RNAs That Silence Genes Are Produced from a Variety of Sources and Direct the Silencing of Genes in Three Different Ways, 712

## SYNTHESIS AND FUNCTION OF miRNA MOLECULES, 714

miRNAs Have a Characteristic Structure That Assists in Identifying Them and Their Target Genes, 714

An Active miRNA Is Generated through a Two-Step Nucleolytic Processing, 716

Dicer Is the Second RNA-Cleaving Enzyme Involved in miRNA Production and the Only One Needed for siRNA Production, 717

## SILENCING GENE EXPRESSION BY SMALL RNAs, 718

Incorporation of a Guide Strand RNA into RISC Makes the Mature Complex That Is Ready to Silence Gene Expression, 718

Small RNAs Can Transcriptionally Silence Genes by Directing Chromatin Modification, 719

RNAi Is a Defense Mechanism That Protects against Viruses and Transposons, 721

**KEY EXPERIMENTS BOX 20-2** Discovery of miRNAs and RNAi, 722

RNAi Has Become a Powerful Tool for Manipulating Gene Expression, 725

**MEDICAL CONNECTIONS BOX 20-3** microRNAs and Human Disease, 727

## LONG NON-CODING RNAs AND X-INACTIVATION, 728

Long Non-Coding RNAs Have Many Roles in Gene Regulation, Including *Cis* and *Trans* Effects on Transcription, 728

X-Inactivation Creates Mosaic Individuals, 728

*Xist* Is a Long Non-Coding RNA That Inactivates a Single X Chromosome in Female Mammals, 729

## SUMMARY, 730

## BIBLIOGRAPHY, 731

## QUESTIONS, 732



## 21 Gene Regulation in Development and Evolution, 733

**MEDICAL CONNECTIONS BOX 21-1** Formation of iPS Cells, 734

## THREE STRATEGIES BY WHICH CELLS ARE INSTRUCTED TO EXPRESS SPECIFIC SETS OF GENES DURING DEVELOPMENT, 735

Some mRNAs Become Localized within Eggs and Embryos Because of an Intrinsic Polarity in the Cytoskeleton, 735

Cell-to-Cell Contact and Secreted Cell-Signaling Molecules Both Elicit Changes in Gene Expression in Neighboring Cells, 736

Gradients of Secreted Signaling Molecules Can Instruct Cells to Follow Different Pathways of Development Based on Their Location, 737

## EXAMPLES OF THE THREE STRATEGIES FOR ESTABLISHING DIFFERENTIAL GENE EXPRESSION, 738

The Localized Ash1 Repressor Controls Mating Type in Yeast by Silencing the *HO* Gene, 738

A Localized mRNA Initiates Muscle Differentiation in the Sea Squirt Embryo, 740

**ADVANCED CONCEPTS BOX 21-2** Review of Cytoskeleton: Asymmetry and Growth, 741

Cell-to-Cell Contact Elicits Differential Gene Expression in the Sporulating Bacterium, *Bacillus subtilis*, 743

A Skin–Nerve Regulatory Switch Is Controlled by Notch Signaling in the Insect Central Nervous System, 743

A Gradient of the Sonic Hedgehog Morphogen Controls the Formation of Different Neurons in the Vertebrate Neural Tube, 744

## THE MOLECULAR BIOLOGY OF DROSOPHILA EMBRYOGENESIS, 746

An Overview of *Drosophila* Embryogenesis, 746

A Regulatory Gradient Controls Dorsoventral Patterning of the *Drosophila* Embryo, 747

**ADVANCED CONCEPTS BOX 21-3** Overview of *Drosophila* Development, 748

Segmentation Is Initiated by Localized RNAs at the Anterior and Posterior Poles of the Unfertilized Egg, 751

**KEY EXPERIMENTS BOX 21-4** Activator Synergy, 752

Bicoid and Nanos Regulate *hunchback*, 753

Multiple Enhancers Ensure Precision of *hunchback* Regulation, 754

The Gradient of Hunchback Repressor Establishes Different Limits of Gap Gene Expression, 754

**MEDICAL CONNECTIONS BOX 21-5** Stem Cell Niche, 755

**ADVANCED CONCEPTS BOX 21-6** Gradient Thresholds, 757

Hunchback and Gap Proteins Produce Segmentation Stripes of Gene Expression, 758	<b>ADVANCED CONCEPTS BOX 21-8</b> Homeotic Genes of <i>Drosophila</i> Are Organized in Special Chromosome Clusters, 764
<b>KEY EXPERIMENTS BOX 21-7</b> <i>cis</i> -Regulatory Sequences in Animal Development and Evolution, 759	How Insects Lost Their Abdominal Limbs, 766
Gap Repressor Gradients Produce Many Stripes of Gene Expression, 760	Modification of Flight Limbs Might Arise from the Evolution of Regulatory DNA Sequences, 767
Short-Range Transcriptional Repressors Permit Different Enhancers to Work Independently of One Another within the Complex <i>eve</i> Regulatory Region, 761	<b>GENOME EVOLUTION AND HUMAN ORIGINS</b> , 769
<b>HOMEOTIC GENES: AN IMPORTANT CLASS OF DEVELOPMENTAL REGULATORS</b> , 762	Diverse Animals Contain Remarkably Similar Sets of Genes, 769
Changes in Homeotic Gene Expression Are Responsible for Arthropod Diversity, 763	Many Animals Contain Anomalous Genes, 769
Changes in <i>Ubx</i> Expression Explain Modifications in Limbs among the Crustaceans, 763	Synteny Is Evolutionarily Ancient, 770
	Deep Sequencing Is Being Used to Explore Human Origins, 772
	<b>SUMMARY</b> , 772
	<b>BIBLIOGRAPHY</b> , 773
	<b>QUESTIONS</b> , 774



## 22 Systems Biology, 775

<b>REGULATORY CIRCUITS</b> , 776	<b>FEED-FORWARD LOOPS</b> , 784
<b>AUTOREGULATION</b> , 776	Feed-Forward Loops Are Three-Node Networks with Beneficial Properties, 784
Negative Autoregulation Dampens Noise and Allows a Rapid Response Time, 777	Feed-Forward Loops Are Used in Development, 786
Gene Expression Is Noisy, 777	<b>OSCILLATING CIRCUITS</b> , 786
Positive Autoregulation Delays Gene Expression, 779	Some Circuits Generate Oscillating Patterns of Gene Expression, 786
<b>BISTABILITY</b> , 780	Synthetic Circuits Mimic Some of the Features of Natural Regulatory Networks, 789
Some Regulatory Circuits Persist in Alternative Stable States, 780	<b>SUMMARY</b> , 790
Bimodal Switches Vary in Their Persistence, 781	<b>BIBLIOGRAPHY</b> , 791
<b>KEY EXPERIMENTS BOX 22-1</b> Bistability and Hysteresis, 782	<b>QUESTIONS</b> , 791

## PART 6: APPENDICES, 793



### APPENDIX 1: Model Organisms, 797

<b>BACTERIOPHAGE</b> , 798	Bacteria Exchange DNA by Sexual Conjugation, Phage-Mediated Transduction, and DNA-Mediated Transformation, 803
Assays of Phage Growth, 800	Bacterial Plasmids Can Be Used as Cloning Vectors, 805
The Single-Step Growth Curve, 800	Transposons Can Be Used to Generate Insertional Mutations and Gene and Operon Fusions, 805
Phage Crosses and Complementation Tests, 801	Studies on the Molecular Biology of Bacteria Have Been Enhanced by Recombinant DNA Technology,
Transduction and Recombinant DNA, 801	
<b>BACTERIA</b> , 802	
Assays of Bacterial Growth, 803	

Whole-Genome Sequencing, and Transcriptional Profiling, 806	THE NEMATODE WORM, <i>CAENORHABDITIS ELEGANS</i> , 816
Biochemical Analysis Is Especially Powerful in Simple Cells with Well-Developed Tools of Traditional and Molecular Genetics, 806	<i>C. elegans</i> Has a Very Rapid Life Cycle, 816
Bacteria Are Accessible to Cytological Analysis, 807	<i>C. elegans</i> Is Composed of Relatively Few, Well-Studied Cell Lineages, 817
Phage and Bacteria Told Us Most of the Fundamentals Things about the Gene, 807	The Cell Death Pathway Was Discovered in <i>C. elegans</i> , 818
Synthetic Circuits and Regulatory Noise, 808	RNAi Was Discovered in <i>C. elegans</i> , 818
<b>BAKER'S YEAST, <i>SACCHAROMYCES CEREVISIAE</i>, 808</b>	<b>THE FRUIT FLY, <i>DROSOPHILA MELANOGASTER</i>, 819</b>
The Existence of Haploid and Diploid Cells Facilitates Genetic Analysis of <i>S. cerevisiae</i> , 809	<i>Drosophila</i> Has a Rapid Life Cycle, 819
Generating Precise Mutations in Yeast Is Easy, 810	The First Genome Maps Were Produced in <i>Drosophila</i> , 820
<i>S. cerevisiae</i> Has a Small, Well-Characterized Genome, 810	Genetic Mosaics Permit the Analysis of Lethal Genes in Adult Flies, 822
<i>S. cerevisiae</i> Cells Change Shape as They Grow, 810	The Yeast FLP Recombinase Permits the Efficient Production of Genetic Mosaics, 823
<b>ARABIDOPSIS, 811</b>	It Is Easy to Create Transgenic Fruit Flies that Carry Foreign DNA, 824
<i>Arabidopsis</i> Has a Fast Life Cycle with Haploid and Diploid Phases, 812	<b>THE HOUSE MOUSE, <i>MUS MUSCULUS</i>, 825</b>
<i>Arabidopsis</i> Is Easily Transformed for Reverse Genetics, 813	Mouse Embryonic Development Depends on Stem Cells, 826
<i>Arabidopsis</i> Has a Small Genome That Is Readily Manipulated, 813	It Is Easy to Introduce Foreign DNA into the Mouse Embryo, 827
Epigenetics, 814	Homologous Recombination Permits the Selective Ablation of Individual Genes, 827
Plants Respond to the Environment, 815	Mice Exhibit Epigenetic Inheritance, 829
Development and Pattern Formation, 815	<b>BIBLIOGRAPHY, 830</b>



## APPENDIX 2: Answers, 831

Chapter 1, 831	Chapter 12, 837
Chapter 2, 831	Chapter 13, 838
Chapter 3, 832	Chapter 14, 839
Chapter 4, 833	Chapter 15, 839
Chapter 5, 833	Chapter 16, 840
Chapter 6, 834	Chapter 17, 841
Chapter 7, 834	Chapter 18, 841
Chapter 8, 835	Chapter 19, 843
Chapter 9, 835	Chapter 20, 843
Chapter 10, 836	Chapter 21, 843
Chapter 11, 837	Chapter 22, 844

## Index, 845

*This page intentionally left blank*

# Box Contents

## ADVANCED CONCEPTS

- BOX 1-1** Mendelian Laws, 6  
**BOX 3-1** The Uniqueness of Molecular Shapes and the Concept of Selective Stickiness, 61  
**BOX 6-1** Ramachandran Plot: Permitted Combinations of Main-Chain Torsion Angles  $\phi$  and  $\psi$ , 124  
**BOX 6-2** Glossary of Terms, 130  
**BOX 6-3** The Antibody Molecule as an Illustration of Protein Domains, 133  
**BOX 9-3** ATP Control of Protein Function: Loading a Sliding Clamp, 282  
**BOX 9-5** *E. coli* DNA Replication Is Regulated by DnaA·ATP Levels and SeqA, 294  
**BOX 10-3** Quantitation of DNA Damage and Its Effects on Cellular Survival and Mutagenesis, 323  
**BOX 10-6** The Y Family of DNA Polymerases, 336  
**BOX 11-1** How to Resolve a Recombination Intermediate with Two Holliday Junctions, 350  
**BOX 12-2** The Xer Recombinase Catalyzes the Monomerization of Bacterial Chromosomes and of Many Bacterial Plasmids, 392  
**BOX 12-4** Mechanism of Transposition Target Immunity, 413

- BOX 13-2** The Single-Subunit RNA Polymerases, 443  
**BOX 15-1** CCA-Adding Enzymes: Synthesizing RNA without a Template, 513  
**BOX 15-2** Selenocysteine, 520  
**BOX 15-3** uORFs and IRESs: Exceptions That Prove the Rule, 533  
**BOX 15-4** GTP-Binding Proteins, Conformational Switching, and the Fidelity and Ordering of the Events of Translation, 546  
**BOX 16-1** Expanding the Genetic Code, 589  
**BOX 18-4** Concentration, Affinity, and Cooperative Binding, 641  
**BOX 19-5** Is There a Histone Code?, 691  
**BOX 20-1** Amino Acid Biosynthetic Operons Are Controlled by Attenuation, 707  
**BOX 21-2** Review of Cytoskeleton: Asymmetry and Growth, 741  
**BOX 21-3** Overview of *Drosophila* Development, 748  
**BOX 21-6** Gradient Thresholds, 757  
**BOX 21-8** Homeotic Genes of *Drosophila* Are Organized in Special Chromosome Clusters, 764

## KEY EXPERIMENTS

- BOX 1-2** Genes Are Linked to Chromosomes, 10  
**BOX 2-1** Chargaff's Rules, 26  
**BOX 2-2** Evidence That Genes Control Amino Acid Sequences in Proteins, 31  
**BOX 4-1** DNA Has 10.5 bp per Turn of the Helix in Solution: The Mica Experiment, 84  
**BOX 4-2** How Spots on an X-Ray Film Reveal the Structure of DNA, 88  
**BOX 4-3** Proving that DNA Has a Helical Periodicity of  $\sim$ 10.5 bp per Turn from the Topological Properties of DNA Rings, 103  
**BOX 6-4** Three-Dimensional Structure of a Protein Is Specified by Its Amino Acid Sequence (Anfinsen Experiment), 135  
**BOX 7-2** Sequenators Are Used for High-Throughput Sequencing, 163  
**BOX 8-1** Micrococcal Nuclease and the DNA Associated with the Nucleosome, 226

- BOX 8-2** Nucleosomes and Superhelical Density, 230  
**BOX 8-3** Determining Nucleosome Position in the Cell, 245  
**BOX 9-4** The Identification of Origins of Replication and Replicators, 290  
**BOX 12-3** Maize Elements and the Discovery of Transposons, 408  
**BOX 14-1** Adenovirus and the Discovery of Splicing, 471  
**BOX 14-2** Converting Group I Introns into Ribozymes, 479  
**BOX 14-3** Identification of Docking Site and Selector Sequences, 490  
**BOX 18-1** Activator Bypass Experiments, 624  
**BOX 18-2** Jacob, Monod, and the Ideas behind Gene Regulation, 628  
**BOX 18-5** Evolution of the  $\lambda$  Switch, 645  
**BOX 18-6** Genetic Approaches That Identified Genes Involved in the Lytic/Lysogenic Choice, 649

- BOX 19-4** Evolution of a Regulatory Circuit, 683  
**BOX 20-2** Discovery of miRNAs and RNAi, 722  
**BOX 21-4** Activator Synergy, 752

## MEDICAL CONNECTIONS

- BOX 5-1** An RNA Switch Controls Protein Synthesis by Murine Leukemia Virus, 112  
**BOX 9-2** Anticancer and Antiviral Agents Target DNA Replication, 268  
**BOX 9-6** Aging, Cancer, and the Telomere Hypothesis, 307  
**BOX 10-1** Expansion of Triple Repeats Causes Disease, 316  
**BOX 10-2** The Ames Test, 321  
**BOX 10-4** Linking Nucleotide Excision Repair and Translesion Synthesis to a Genetic Disorder in Humans, 330  
**BOX 10-5** Nonhomologous End Joining, 332  
**BOX 11-2** The Product of the Tumor Suppressor Gene *BRCA2* Interacts with Rad51 Protein and Controls Genome Stability, 367  
**BOX 11-3** Proteins Associated with Premature Aging and Cancer Promote an Alternative Pathway for Holliday Junction Processing, 368

## TECHNIQUES

- BOX 5-2** Creating an RNA Mimetic of the Green Fluorescent Protein by Directed Evolution, 115  
**BOX 7-1** Forensics and the Polymerase Chain Reaction, 160  
**BOX 9-1** Incorporation Assays Can Be Used to Measure Nucleic Acid and Protein Synthesis, 261

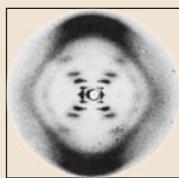
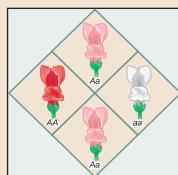
- BOX 21-7** *cis*-Regulatory Sequences in Animal Development and Evolution, 759  
**BOX 22-1** Bistability and Hysteresis, 782

- BOX 12-1** Application of Site-Specific Recombination to Genetic Engineering, 386  
**BOX 14-4** Defects in Pre-mRNA Splicing Cause Human Disease, 497  
**BOX 14-5** Deaminases and HIV, 503  
**BOX 15-5** Antibiotics Arrest Cell Division by Blocking Specific Steps in Translation, 552  
**BOX 15-7** A Frontline Drug in Tuberculosis Therapy Targets SsrA Tagging, 565  
**BOX 18-3** Blocking Virulence by Silencing Pathways of Intercellular Communication, 635  
**BOX 19-3** Histone Modifications, Transcription Elongation, and Leukemia, 670  
**BOX 19-6** Transcriptional Repression and Human Disease, 696  
**BOX 20-3** microRNAs and Human Disease, 727  
**BOX 21-1** Formation of iPS Cells, 734  
**BOX 21-5** Stem Cell Niche, 755

- BOX 13-1** Consensus Sequences, 436  
**BOX 15-6** Ribosome and Polysome Profiling, 561  
**BOX 19-1** The Two-Hybrid Assay, 664  
**BOX 19-2** The ChIP-Chip and ChIP-Seq Assays Are the Best Method for Identifying Enhancers, 666

P A R T      1

# HISTORY



O U T L I N E

---

**CHAPTER 1**  
The Mendelian View  
of the World, 5

•  
**CHAPTER 2**  
Nucleic Acids Convey Genetic  
Information, 21

**U**NLIKE THE REST OF THIS BOOK, the two chapters that make up Part 1 contain material largely unchanged from earlier editions. We nevertheless keep these chapters because the material remains as important as ever. Specifically, Chapters 1 and 2 provide an historical account of how the field of genetics and the molecular basis of genetics were established. Key ideas and experiments are described.

Chapter 1 addresses the founding events in the history of genetics. We discuss everything from Mendel's famous experiments on peas, which uncovered the basic laws of heredity, to the one gene encodes one enzyme hypothesis of Garrod. Chapter 2 describes the revolutionary development of molecular biology that was started with Avery's discovery that DNA was the genetic material, and continued with James Watson and Francis Crick's proposal that the structure of DNA is a double helix, and the elucidation of the genetic code and the "central dogma" (DNA "makes" RNA which "makes" protein). Chapter 2 concludes with a discussion of recent developments stemming from the complete sequencing of the genomes of many organisms and the impact this sequencing has on modern biology.

### PHOTOS FROM THE COLD SPRING HARBOR LABORATORY ARCHIVES

---

**Vernon Ingram, Marshall W. Nirenberg, and Matthias Staehelin, 1963 Symposium on Synthesis and Structure of Macromolecules.** Ingram demonstrated that genes control the amino acid sequence of proteins; the mutation causing sickle-cell anemia produces a single amino acid change in the hemoglobin protein (Chapter 2). Nirenberg was key in unraveling the genetic code, using protein synthesis directed by artificial RNA templates *in vitro* (Chapters 2 and 16). For this achievement, he shared in the 1968 Nobel Prize in Physiology or Medicine. Staehelin worked on the small RNA molecules, tRNAs, which translate the genetic code into amino acid sequences of proteins (Chapters 2 and 16).

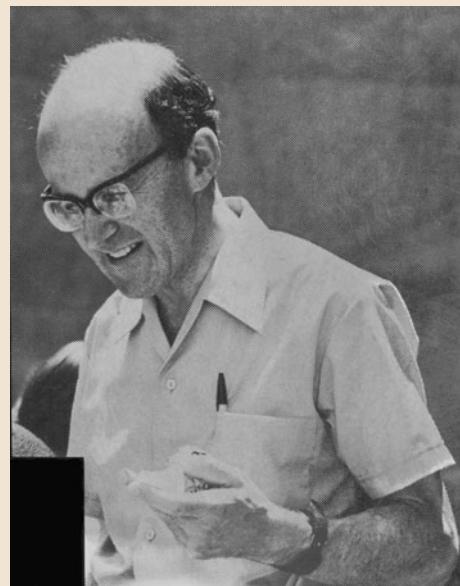


**Melvin Calvin, Francis Crick, George Gamow, and James Watson, 1963 Symposium on Synthesis and Structure of Macromolecules.** Calvin won the 1961 Nobel Prize in Chemistry for his work on CO<sub>2</sub> assimilation by plants. For their proposed structure of DNA, Crick and Watson shared in the 1962 Nobel Prize in Physiology or Medicine (Chapters 2 and 4). Gamow, a physicist attracted to the problem of the genetic code (Chapters 2 and 16), founded an informal group of like-minded scientists called the RNA Tie Club. (He is wearing the club tie, which he designed, in this picture.)





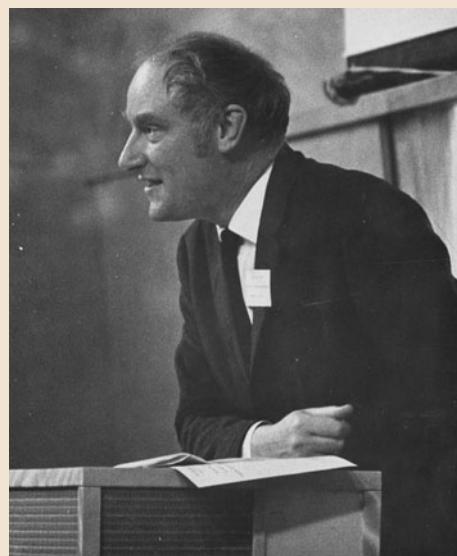
**Raymond Appleyard, George Bowen, and Martha Chase, 1953 Symposium on Viruses.** Appleyard and Bowen, both phage geneticists, are here shown with Chase, who, in 1952, together with Alfred Hershey, did the simple experiment that finally convinced most people that the genetic material is DNA (Chapter 2).



**Max Perutz, 1971 Symposium on Structure and Function of Proteins at the Three-Dimensional Level.** Perutz shared, with John Kendrew, the 1962 Nobel Prize for Chemistry; using X-ray crystallography, and after 25 years of effort, they were the first to solve the atomic structures of proteins—hemoglobin and myoglobin, respectively (Chapter 6).



**Sydney Brenner and James Watson, 1975 Symposium on The Synapse.** Brenner, shown here with Watson, contributed to the discoveries of mRNA and the nature of the genetic code (Chapters 2 and 16); his share of a Nobel Prize, in 2002, however, was for establishing the worm, *Caenorhabditis elegans*, as a model system for the study of developmental biology (Appendix 1).



**Francis Crick, 1963 Symposium on Synthesis and Structure of Macromolecules.** In addition to his role in solving the structure of DNA, Crick was an intellectual driving force in the development of molecular biology during the field's critical early years. His "adaptor hypothesis" (published in the RNA Tie Club newsletter) predicted the existence of molecules required to translate the genetic code of RNA into the amino acid sequence of proteins. Only later were tRNAs found to do just that (Chapter 15).



**Seymour Benzer, 1975 Symposium on The Synapse.** Using phage genetics, Benzer defined the smallest unit of mutation, which turned out later to be a single nucleotide (Chapter 1 and Appendix 1). This same work also provided an experimental definition of the gene—which he called a cistron—using functional complementation tests. Later, his studies focused on behavior, using the fruit fly as a model.



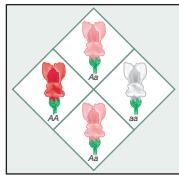
**Calvin Bridges, 1934 Symposium on Aspects of Growth.** Bridges (shown reading the newspaper) was part of T.H. Morgan's famous "fly group" that pioneered the development of the fruit fly *Drosophila* as a model genetic organism (Chapter 1 and Appendix 1). With him is John T. Buchholtz, a plant geneticist who was a summer visitor at CSHL at the time, and who, in 1941, became President of the Botanical Society of America.



**Charles Yanofsky, 1966 Symposium on The Genetic Code.** Yanofsky (right), together with Sydney Brenner, proved colinearity of the gene—that is, that successive groups of nucleotides encoded successive amino acids in the protein product (Chapter 2). He later discovered the first example of transcriptional regulation by RNA structure in his detailed analysis of attenuation at the tryptophan operon of *Escherichia coli* (Chapter 20). He is pictured here talking to Michael Chamberlin, who studied transcription initiation by RNA polymerase.



**Edwin Chargaff, 1947 Symposium on Nucleic Acids and Nucleoproteins.** The eminent nucleic acid biochemist Chargaff's famous ratios—that the amount of adenine in a DNA sample matched that of thymine, and the amount of cytosine matched that of guanine—were later understood in the context of Watson and Crick's DNA double helix structure. Perhaps frustrated that he had never come up with base pairs himself, he became a bitter critic of molecular biology, an occupation he described as "essentially the practice of biochemistry without a license."



# The Mendelian View of the World

IT IS EASY TO CONSIDER HUMAN BEINGS UNIQUE among living organisms. We alone have developed complicated languages that allow meaningful and complex interplay of ideas and emotions. Great civilizations have developed and changed our world's environment in ways inconceivable for any other form of life. There has always been a tendency, therefore, to think that something special differentiates humans from every other species. This belief has found expression in the many forms of religion through which we seek the origin of and explore the reasons for our existence and, in so doing, try to create workable rules for conducting our lives. Little more than a century ago, it seemed natural to think that, just as every human life begins and ends at a fixed time, the human species and all other forms of life must also have been created at a fixed moment.

This belief was first seriously questioned almost 150 years ago, when Charles Darwin and Alfred R. Wallace proposed their theories of evolution, based on the selection of the most fit. They stated that the various forms of life are not constant but continually give rise to slightly different animals and plants, some of which adapt to survive and multiply more effectively. At the time of this theory, they did not know the origin of this continuous variation, but they did correctly realize that these new characteristics must persist in the progeny if such variations are to form the basis of evolution.

At first, there was a great furor against Darwin, most of it coming from people who did not like to believe that humans and the rather obscene-looking apes could have a common ancestor, even if this ancestor had lived some 10 million years ago. There was also initial opposition from many biologists who failed to find Darwin's evidence convincing. Among these was the famous naturalist Jean L. Agassiz, then at Harvard, who spent many years writing against Darwin and Darwin's champion, Thomas H. Huxley, the most successful of the popularizers of evolution. But by the end of the 19th century, the scientific argument was almost complete; both the current geographic distribution of plants and animals and their selective occurrence in the fossil records of the geologic past were explicable only by postulating that continuously evolving groups of organisms had descended from a common ancestor. Today, evolution is an accepted fact for everyone except a fundamentalist minority, whose objections are based not on reasoning but on doctrinaire adherence to religious principles.

An immediate consequence of Darwinian theory is the realization that life first existed on our Earth more than 4 billion years ago in a simple

## OUTLINE

### Mendel's Discoveries, 6



### Chromosomal Theory of Heredity, 8



### Gene Linkage and Crossing Over, 9



### Chromosome Mapping, 11



### The Origin of Genetic Variability through Mutations, 13



### Early Speculations about What Genes Are and How They Act, 15



### Preliminary Attempts to Find a Gene–Protein Relationship, 16



### Visit Web Content for Structural Tutorials and Interactive Animations

form, possibly resembling the bacteria—the simplest variety of life known today. The existence of such small bacteria tells us that the essence of the living state is found in very small organisms. Evolutionary theory further suggests that the basic principles of life apply to all living forms.

## MENDEL'S DISCOVERIES

Gregor Mendel's experiments traced the results of breeding experiments (genetic crosses) between strains of peas differing in well-defined characteristics, like seed shape (round or wrinkled), seed color (yellow or green), pod shape (inflated or wrinkled), and stem length (long or short). His concentration on well-defined differences was of great importance; many breeders had previously tried to follow the inheritance of more gross qualities, like body weight, and were unable to discover any simple rules about their transmission from parents to offspring (see Box 1-1, Mendelian Laws).

### The Principle of Independent Segregation

After ascertaining that each type of parental strain bred true—that is, produced progeny with particular qualities identical to those of the parents—Mendel performed a number of crosses between parents (P) differing in single characteristics (such as seed shape or seed color). All the progeny ( $F_1$  = first filial generation) had the appearance of *one* parent only. For example, in a cross between peas having yellow seeds and peas having green seeds, all the progeny had yellow seeds. The trait that appears in the  $F_1$  progeny is called **dominant**, whereas the trait that does not appear in  $F_1$  is called **recessive**.

#### ► ADVANCED CONCEPTS

##### Box 1-1 Mendelian Laws

The most striking attribute of a living cell is its ability to transmit hereditary properties from one cell generation to another. The existence of heredity must have been noticed by early humans, who witnessed the passing of characteristics, like eye or hair color, from parents to offspring. Its physical basis, however, was not understood until the first years of the 20th century, when, during a remarkable period of creative activity, the chromosomal theory of heredity was established.

Hereditary transmission through the sperm and egg became known by 1860, and in 1868 Ernst Haeckel, noting that sperm consists largely of nuclear material, postulated that the nucleus is responsible for heredity. Almost 20 years passed before the chromosomes were singled out as the active factors, because the details of mitosis, meiosis, and fertilization had to be worked out first. When this was accomplished, it could be seen that, unlike other cellular constituents, the chromosomes are equally divided between daughter cells. Moreover, the complicated chromosomal changes that reduce the sperm and egg chromosome number to the haploid number during meiosis became understandable as nec-

essary for keeping the chromosome number constant. These facts, however, merely suggested that chromosomes carry heredity.

Proof came at the turn of the century with the discovery of the basic rules of heredity. The concepts were first proposed by Gregor Mendel in 1865 in a paper entitled "Experiments in Plant Hybridization" given to the Natural Science Society at Brno. In his presentation, Mendel described in great detail the patterns of transmission of traits in pea plants, his conclusions of the principles of heredity, and their relevance to the controversial theories of evolution. The climate of scientific opinion, however, was not favorable, and these ideas were completely ignored, despite some early efforts on Mendel's part to interest the prominent biologists of his time. In 1900, 16 years after Mendel's death, three plant breeders working independently on different systems confirmed the significance of Mendel's forgotten work. Hugo de Vries, Karl Correns, and Erich von Tschermak-Seysenegg, all doing experiments related to Mendel's, reached similar conclusions before they knew of Mendel's work.

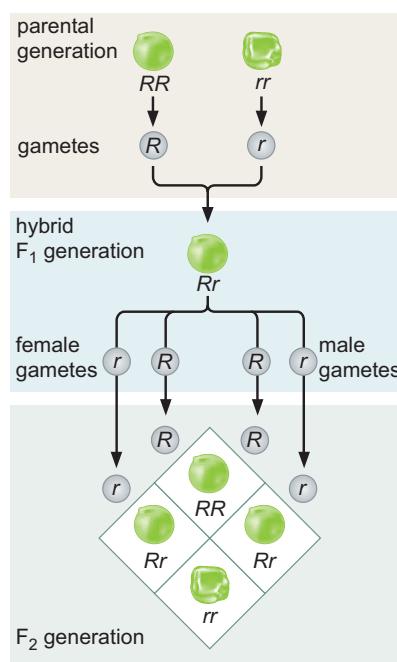
The meaning of these results became clear when Mendel set up genetic crosses between  $F_1$  offspring. These crosses gave the important result that the recessive trait reappeared in approximately 25% of the  $F_2$  progeny, whereas the dominant trait appeared in 75% of these offspring. For each of the seven traits he followed, the ratio in  $F_2$  of dominant to recessive traits was always approximately 3:1. When these experiments were carried to a third ( $F_3$ ) progeny generation, all the  $F_2$  peas with recessive traits bred true (produced progeny with the recessive traits). Those with dominant traits fell into two groups: one third bred true (produced only progeny with the dominant trait); the remaining two-thirds again produced mixed progeny in a 3:1 ratio of dominant to recessive.

Mendel correctly interpreted his results as follows (Fig. 1-1): the various traits are controlled by pairs of factors (which we now call **genes**), one factor derived from the male parent, the other from the female. For example, pure-breeding strains of round peas contain two versions (or **alleles**) of the roundness gene ( $RR$ ), whereas pure-breeding wrinkled strains have two copies of the wrinkledness ( $rr$ ) allele. The round-strain gametes each have one gene for roundness ( $R$ ); the wrinkled-strain gametes each have one gene for wrinkledness ( $r$ ). In a cross between  $RR$  and  $rr$ , fertilization produces an  $F_1$  plant with both alleles ( $Rr$ ). The seeds look round because  $R$  is dominant over  $r$ . We refer to the appearance or physical structure of an individual as its **phenotype**, and to its genetic composition as its **genotype**. Individuals with identical phenotypes may possess different genotypes; thus, to determine the genotype of an organism, it is frequently necessary to perform genetic crosses for several generations. The term **homozygous** refers to a gene pair in which both the maternal and paternal genes are identical (e.g.,  $RR$  or  $rr$ ). In contrast, those gene pairs in which paternal and maternal genes are different (e.g.,  $Rr$ ) are called **heterozygous**.

One or several letters or symbols may be used to represent a particular gene. The dominant allele of the gene may be indicated by a capital letter ( $R$ ), by a superscript + ( $r^+$ ), or by a + standing alone. In our discussions here, we use the first convention in which the dominant allele is represented by a capital letter and the recessive allele by the lowercase letter.

It is important to notice that a given gamete contains only one of the two copies (one allele) of the genes present in the organism it comes from (e.g., either  $R$  or  $r$ , but never both) and that the two types of gametes are produced in equal numbers. Thus, there is a 50:50 chance that a given gamete from an  $F_1$  pea will contain a particular gene ( $R$  or  $r$ ). This choice is purely random. We do not expect to find exact 3:1 ratios when we examine a limited number of  $F_2$  progeny. The ratio will sometimes be slightly higher and other times slightly lower. But as we look at increasingly larger samples, we expect that the ratio of peas with the dominant trait to peas with the recessive trait will approximate the 3:1 ratio more and more closely.

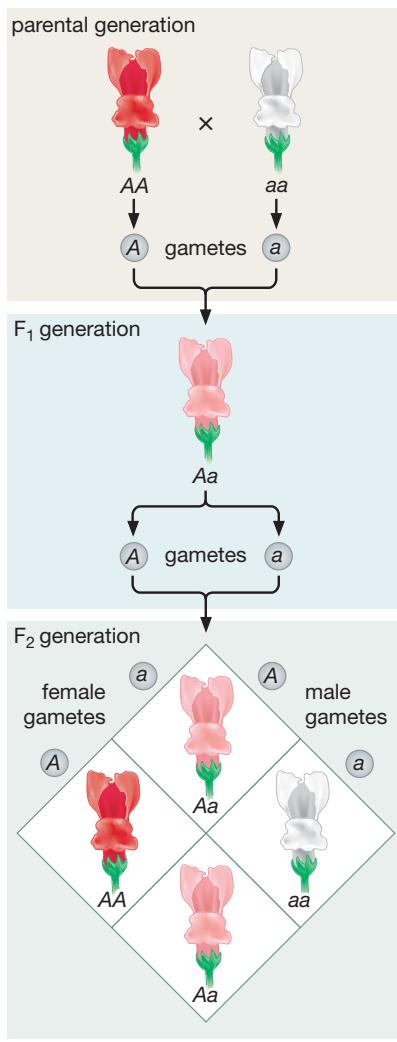
The reappearance of the recessive characteristic in the  $F_2$  generation indicates that recessive alleles are neither modified nor lost in the  $F_1$  ( $Rr$ ) generation, but that the dominant and recessive genes are independently transmitted and so are able to segregate independently during the formation of sex cells. This **principle of independent segregation** is frequently referred to as Mendel's first law.



**FIGURE 1-1** How Mendel's first law (independent segregation) explains the 3:1 ratio of dominant to recessive phenotypes among the  $F_2$  progeny.  $R$  represents the dominant gene and  $r$  the recessive gene. The round seed represents the dominant phenotype, the wrinkled seed the recessive phenotype.

## Some Alleles Are neither Dominant nor Recessive

In the crosses reported by Mendel, one member of each gene pair was clearly dominant to the other. Such behavior, however, is not universal. Sometimes the heterozygous phenotype is intermediate between the two homozygous



**FIGURE 1-2** The inheritance of flower color in the snapdragon. One parent is homozygous for red flowers ( $AA$ ) and the other homozygous for white flowers ( $aa$ ). No dominance is present, and the heterozygous  $F_1$  flowers are pink. The 1:2:1 ratio of red, pink, and white flowers in the  $F_2$  progeny is shown by appropriate coloring.

phenotypes. For example, the cross between a pure-breeding red snapdragon (*Antirrhinum*) and a pure-breeding white variety gives  $F_1$  progeny of the intermediate pink color. If these  $F_1$  progeny are crossed among themselves, the resulting  $F_2$  progeny contain red, pink, and white flowers in the proportion of 1:2:1 (Fig. 1-2). Thus, it is possible here to distinguish heterozygotes from homozygotes by their phenotype. We also see that Mendel's laws do not depend on whether one allele of a gene pair is dominant over the other.

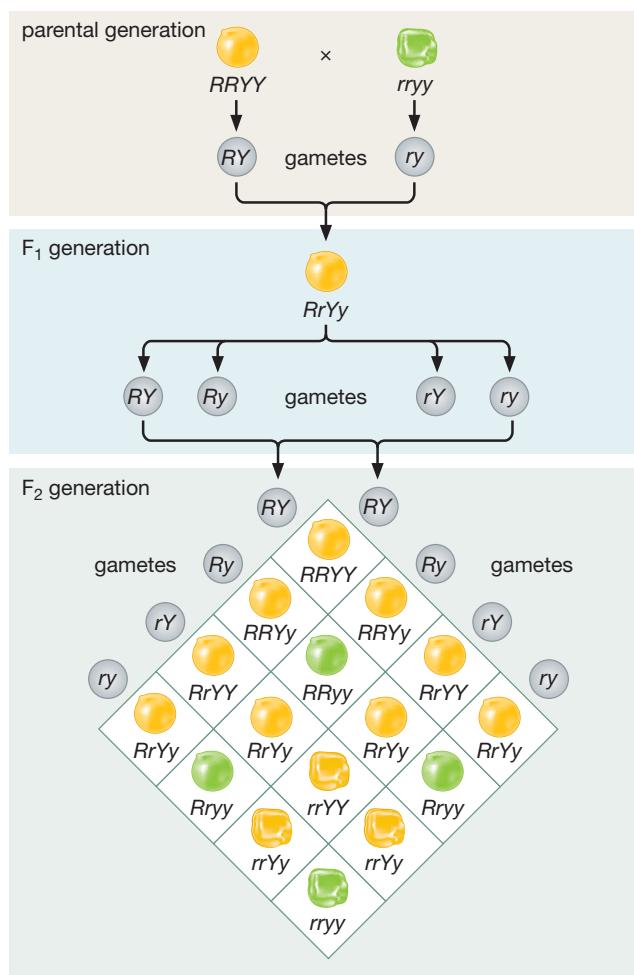
### Principle of Independent Assortment

Mendel extended his breeding experiments to peas differing by more than one characteristic. As before, he started with two strains of peas, each of which bred pure when mated with itself. One of the strains had round yellow seeds; the other, wrinkled green seeds. Since round and yellow are dominant over wrinkled and green, the entire  $F_1$  generation produced round yellow seeds. The  $F_1$  generation was then crossed within itself to produce a number of  $F_2$  progeny, which were examined for seed appearance (phenotype). In addition to the two original phenotypes (round yellow; wrinkled green), two new types (**recombinants**) emerged: wrinkled yellow and round green.

Again Mendel found he could interpret the results by the postulate of genes, if he assumed that each gene pair was independently transmitted to the gamete during sex-cell formation. This interpretation is shown in Figure 1-3. Any one gamete contains only one type of allele from each gene pair. Thus, the gametes produced by an  $F_1$  ( $RrYy$ ) will have the composition  $RY$ ,  $Ry$ ,  $rY$ , or  $ry$ , but never  $Rr$ ,  $Yy$ ,  $YY$ , or  $RR$ . Furthermore, in this example, all four possible gametes are produced with equal frequency. There is no tendency of genes arising from one parent to stay together. As a result, the  $F_2$  progeny phenotypes appear in the ratio nine round yellow, three round green, three wrinkled yellow, and one wrinkled green as depicted in the Punnett square, named after the British mathematician who introduced it (in the lower part of Fig. 1-3). This **principle of independent assortment** is frequently called Mendel's second law.

### CHROMOSOMAL THEORY OF HEREDITY

A principal reason for the original failure to appreciate Mendel's discovery was the absence of firm facts about the behavior of chromosomes during meiosis and mitosis. This knowledge was available, however, when Mendel's laws were confirmed in 1900 and was seized upon in 1903 by American biologist Walter S. Sutton. In his classic paper "The Chromosomes in Heredity," Sutton emphasized the importance of the fact that the diploid chromosome group consists of two morphologically similar sets and that, during meiosis, every gamete receives only one chromosome of each homologous pair. He then used this fact to explain Mendel's results by assuming that genes are parts of the chromosome. He postulated that the yellow- and green-seed genes are carried on a certain pair of chromosomes and that the round- and wrinkled-seed genes are carried on a different pair. This hypothesis immediately explains the experimentally observed 9:3:3:1 segregation ratios. Although Sutton's paper did not prove the chromosomal theory of heredity, it was immensely important, for it brought together for the first time the independent disciplines of genetics (the study of breeding experiments) and cytology (the study of cell structure).



**FIGURE 1-3** How Mendel's second law (independent assortment) operates. In this example, the inheritance of yellow ( $Y$ ) and green ( $y$ ) seed color is followed together with the inheritance of round ( $R$ ) and wrinkled ( $r$ ) seed shapes. The  $R$  and  $Y$  alleles are dominant over  $r$  and  $y$ . The genotypes of the various parents and progeny are indicated by letter combinations, and four different phenotypes are distinguished by appropriate shading.

## GENE LINKAGE AND CROSSING OVER

Mendel's principle of independent assortment is based on the fact that genes located on different chromosomes behave independently during meiosis. Often, however, two genes do not assort independently because they are located on the same chromosome (**linked genes**; see Box 1-2, Genes Are Linked to Chromosomes). Many examples of nonrandom assortment were found as soon as a large number of mutant genes became available for breeding analysis. In every well-studied case, the number of linked groups was identical to the haploid chromosome number. For example, there are four groups of linked genes in *Drosophila* and four morphologically distinct chromosomes in a haploid cell.

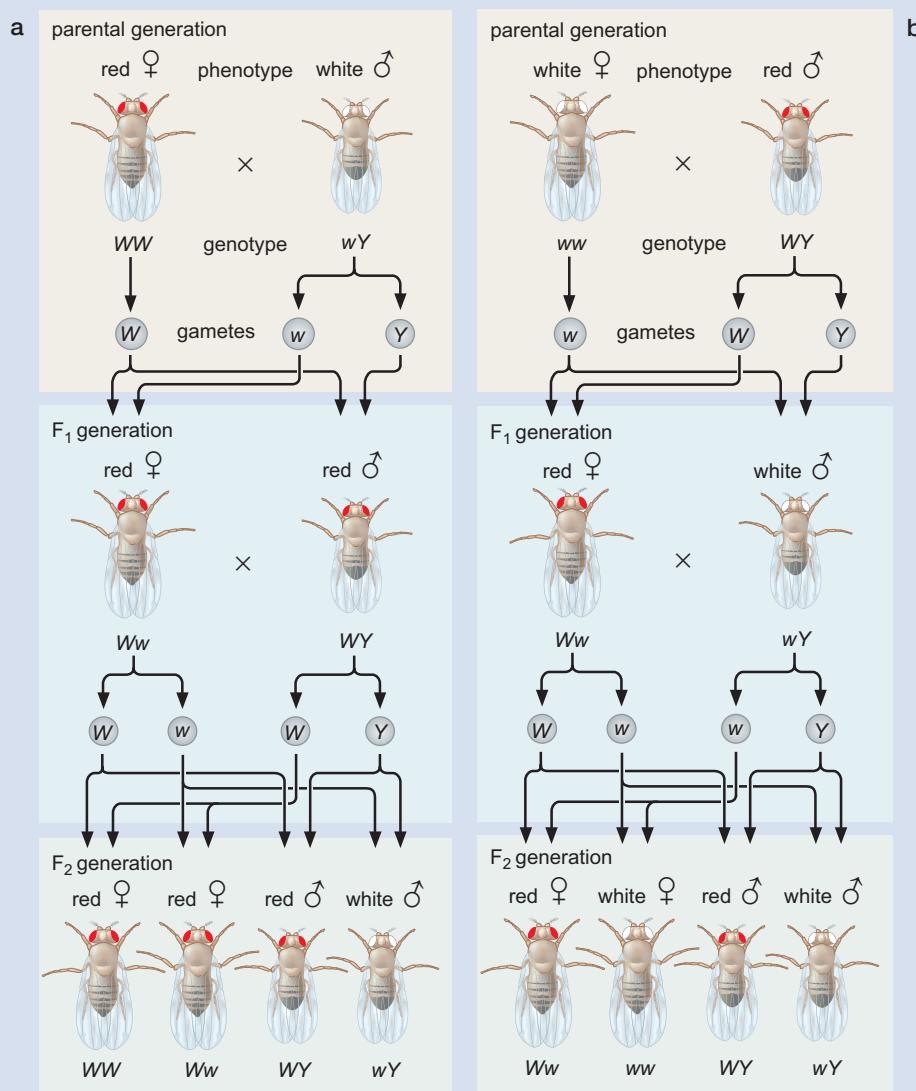
Linkage, however, is in effect never complete. The probability that two genes on the same chromosome will remain together during meiosis ranges from just less than 100% to nearly 50%. This variation in linkage suggests that there must be a mechanism for exchanging genes on homologous chromosomes. This mechanism is called **crossing over**. Its cytological basis was first described by Belgian cytologist F.A. Janssens. At the start of meiosis, through the process of **synapsis**, the homologous chromosomes form pairs with their long axes parallel. At this stage, each chromosome has duplicated to form two chromatids. Thus, synapsis brings together four chromatids (a tetrad), which coil about one another. Janssens postulated that, possibly

**Box 1-2** Genes Are Linked to Chromosomes

Initially, all breeding experiments used genetic differences already existing in nature. For example, Mendel used seeds obtained from seed dealers, who must have obtained them from farmers. The existence of alternative forms of the same gene (alleles) raises the question of how they arose. One obvious hypothesis states that genes can change (mutate) to give rise to new genes (**mutant genes**). This hypothesis was first seriously tested, beginning in 1908, by the great American biologist Thomas Hunt Morgan and his young collaborators, geneticists Calvin B. Bridges, Hermann J. Muller, and Alfred H. Sturtevant. They worked with the tiny fly *Drosophila melanogaster*. The first mutant found was a male with white eyes instead of the normal red eyes. The white-eyed variant appeared spontaneously in a

culture bottle of red-eyed flies. Because essentially all *Drosophila* found in nature have red eyes, the gene leading to red eyes was referred to as the **wild-type gene**; the gene leading to white eyes was called a mutant gene (allele).

The white-eye mutant gene was immediately used in breeding experiments (Box 1-2 Fig. 1), with the striking result that the behavior of the allele completely paralleled the distribution of an X chromosome (i.e., was sex-linked). This finding immediately suggested that this gene might be located on the X chromosome, together with those genes controlling sex. This hypothesis was quickly confirmed by additional genetic crosses using newly isolated mutant genes. Many of these additional mutant genes also were sex-linked.

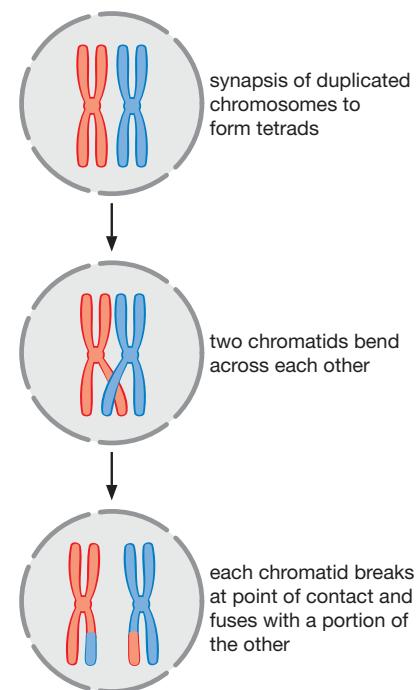


**BOX 1-2 FIGURE 1** The inheritance of a sex-linked gene in *Drosophila*. Genes located on sex chromosomes can express themselves differently in male and female progeny, because if there is only one X chromosome present, recessive genes on this chromosome are always expressed. Here are two crosses, both involving a recessive gene (*w*, for white eye) located on the X chromosome. (a) The male parent is a white-eyed (*wY*) fly, and the female is homozygous for red eye (*WW*). (b) The male has red eyes (*WY*) and the female white eyes (*ww*). The letter *Y* stands here not for an allele, but for the Y chromosome, present in male *Drosophila* in place of a homologous X chromosome. There is no gene on the Y chromosome corresponding to the *w* or *W* gene on the X chromosome.

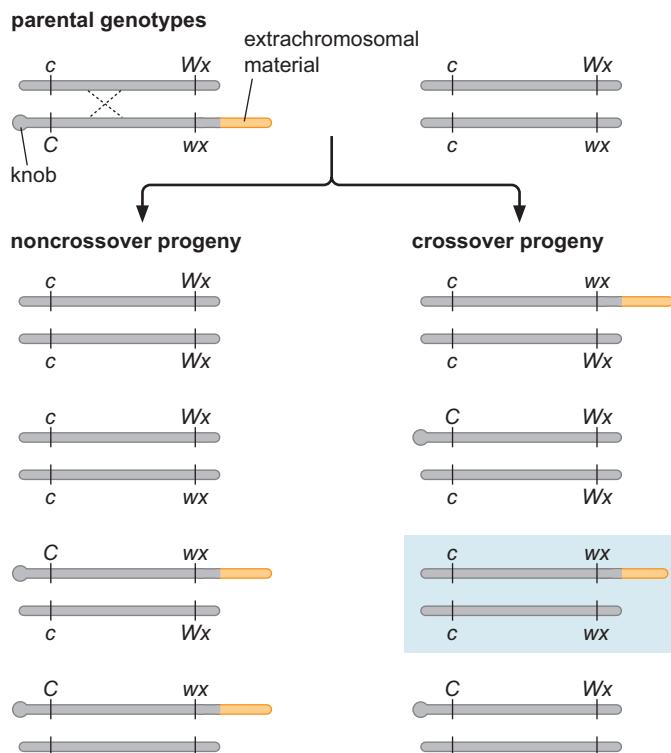
because of tension resulting from this coiling, two of the chromatids might sometimes break at a corresponding place on each. These events could create four broken ends, which might rejoin crossways, so that a section of each of the two chromatids would be joined to a section of the other (Fig. 1-4). In this manner, recombinant chromatids might be produced that contain a segment derived from each of the original homologous chromosomes. Formal proof of Janssens's hypothesis that chromosomes physically interchange material during synapsis came more than 20 years later, when in 1931, Barbara McClintock and Harriet B. Creighton, working at Cornell University with the corn plant *Zea mays*, devised an elegant cytological demonstration of chromosome breakage and rejoining (Fig. 1-5).

## CHROMOSOME MAPPING

Thomas Hunt Morgan and his students, however, did not await formal cytological proof of crossing over before exploiting the implication of Janssens's hypothesis. They reasoned that genes located close together on a chromosome would assort with one another much more regularly (close linkage) than genes located far apart on a chromosome. They immediately saw this as a way to locate (map) the relative positions of genes on chromosomes and thus to produce a **genetic map**. The way they used the frequencies of the various recombinant classes is very straightforward. Consider the segregation of three genes all located on the same chromosome. The arrangement of the genes can be determined by means of three crosses, in each of which two genes are followed (two-factor crosses). A cross between *AB* and *ab* yields four progeny types: the two parental genotypes (*AB* and *ab*) and two recombinant genotypes (*Ab* and *aB*). A cross between *AC* and *ac* similarly gives two parental combinations as well as the *Ac* and *ac*



**FIGURE 1-4** Janssens's hypothesis of crossing over.



**FIGURE 1-5** Demonstration of physical exchanges between homologous chromosomes. In most organisms, pairs of homologous chromosomes have identical shapes. Occasionally, however, the two members of a pair are not identical; one is marked by the presence of extrachromosomal material or compacted regions that reproducibly form knob-like structures. McClintock and Creighton found one such pair and used it to show that crossing over involves actual physical exchanges between the paired chromosomes. In the experiment shown here, the homozygous *c, wx* progeny had to arise by crossing over between the *C* and *wx* loci. When such *c, wx* offspring were cytologically examined, knob chromosomes were seen, showing that a knobless *Wx* region had been physically replaced by a knobbed *wx* region. The colored box in the figure identifies the chromosomes of the homozygous *c, wx* offspring.

**FIGURE 1-6** Assignment of the tentative order of three genes on the basis of three two-factor crosses.

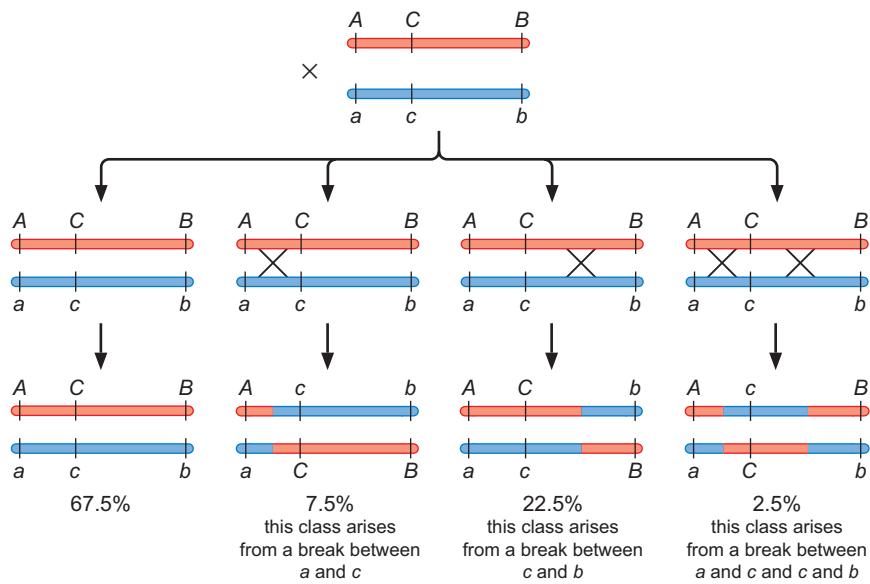


recombinants, whereas a cross between *BC* and *bc* produces the parental types and the recombinants *Bc* and *bC*. Each cross will produce a specific ratio of parental to recombinant progeny. Consider, for example, the fact that the first cross gives 30% recombinants, the second cross 10%, and the third cross 25%. This tells us that genes *a* and *c* are closer together than *a* and *b* or *b* and *c* and that the genetic distances between *a* and *b* and *b* and *c* are more similar. The gene arrangement that best fits these data is *a-c-b* (Fig. 1-6).

The correctness of gene order suggested by crosses of two gene factors can usually be unambiguously confirmed by three-factor crosses. When the three genes used in the preceding example are followed in the cross *ABC* × *abc*, six recombinant genotypes are found (Fig. 1-7). They fall into three groups of reciprocal pairs. The rarest of these groups arises from a double crossover. By looking for the least frequent class, it is often possible to instantly confirm (or deny) a postulated arrangement. The results in Figure 1-7 immediately confirm the order hinted at by the two-factor crosses. Only if the order is *a-c-b* does the fact that the rare recombinants are *AcB* and *aCb* make sense.

The existence of multiple crossovers means that the amount of recombination between the outside markers *a* and *b* (*ab*) is usually less than the sum of the recombination frequencies between *a* and *c* (*ac*) and *c* and *b* (*cb*). To obtain a more accurate approximation of the distance between the outside markers, we calculate the probability (*ac* × *cb*) that when a crossover occurs between *c* and *b*, a crossover also occurs between *a* and *c*, and vice versa (*cb* × *ac*). This probability subtracted from the sum of the frequencies expresses more accurately the amount of recombination. The simple formula

$$ab = ac + cb - 2(ac)(cb)$$



**FIGURE 1-7** The use of three-factor crosses to assign gene order. The least frequent pair of reciprocal recombinants must arise from a double crossover. The percentages listed for the various classes are the theoretical values expected for an infinitely large sample. When finite numbers of progeny are recorded, the exact values will be subject to random statistical fluctuations.

is applicable in all cases where the occurrence of one crossover does not affect the probability of another crossover. Unfortunately, accurate mapping is often disturbed by *interference* phenomena, which can either increase or decrease the probability of correlated crossovers.

Using such reasoning, the Columbia University group headed by Morgan had by 1915 assigned locations to more than 85 mutant genes in *Drosophila* (Table 1-1), placing each of them at distinct spots on one of the four linkage groups, or chromosomes. Most importantly, all the genes on a given chromosome were located on a line. The gene arrangement was strictly linear and never branched. The genetic map of one of the chromosomes of *Drosophila* is shown in Figure 1-8. Distances between genes on such a map are measured in **map units**, which are related to the frequency of recombination between the genes. Thus, if the frequency of recombination between two genes is found to be 5%, the genes are said to be separated by five map units. Because of the high probability of double crossovers between widely spaced genes, such assignments of map units can be considered accurate only if recombination between closely spaced genes is followed.

Even when two genes are at the far ends of a very long chromosome, they assort together at least 50% of the time because of multiple crossovers. The two genes will be separated if an odd number of crossovers occurs between them, but they will end up together if an even number occurs between them. Thus, in the beginning of the genetic analysis of *Drosophila*, it was often impossible to determine whether two genes were on different chromosomes or at the opposite ends of one long chromosome. Only after large numbers of genes had been mapped was it possible to demonstrate convincingly that the number of linkage groups equalled the number of cytologically visible chromosomes. In 1915, Morgan, with his students Alfred H. Sturtevant, Hermann J. Muller, and Calvin B. Bridges, published their definitive book *The Mechanism of Mendelian Heredity*, which first announced the general validity of the chromosomal basis of heredity. We now rank this concept, along with the theories of evolution and the cell, as a major achievement in our quest to understand the nature of the living world.

## THE ORIGIN OF GENETIC VARIABILITY THROUGH MUTATIONS

---

It now became possible to understand the hereditary variation that is found throughout the biological world and that forms the basis of the theory of evolution. Genes are normally copied exactly during chromosome duplication. Rarely, however, changes (**mutations**) occur in genes to give rise to altered forms, most—but not all—of which function less well than the wild-type alleles. This process is necessarily rare; otherwise, many genes would be changed during every cell cycle, and offspring would not ordinarily resemble their parents. There is, instead, a strong advantage in there being a small but finite mutation rate; it provides a constant source of new variability, necessary to allow plants and animals to adapt to a constantly changing physical and biological environment.

Surprisingly, however, the results of the Mendelian geneticists were not avidly seized upon by the classical biologists, then the authorities on the evolutionary relations between the various forms of life. Doubts were raised about whether genetic changes of the type studied by Morgan and his students were sufficient to permit the evolution of radically new structures,

**TABLE 1-1** The 85 Mutant Genes Reported in *Drosophila melanogaster* in 1915

Name	Region Affected	Name	Region Affected
<b>Group 1</b>			
<i>Abnormal</i>	Abdomen	<i>Lethal, 13</i>	Body, death
<i>Bar</i>	Eye	<i>Miniature</i>	Wing
<i>Bifid</i>	Venetation	<i>Notch</i>	Venetation
<i>Bow</i>	Wing	<i>Reduplicated</i>	Eye color
<i>Cherry</i>	Eye color	<i>Ruby</i>	Leg
<i>Chrome</i>	Body color	<i>Rudimentary</i>	Wing
<i>Cleft</i>	Venetation	<i>Sable</i>	Body color
<i>Club</i>	Wing	<i>Shifted</i>	Venetation
<i>Depressed</i>	Wing	<i>Short</i>	Wing
<i>Dotted</i>	Thorax	<i>Skee</i>	Wing
<i>Eosin</i>	Eye color	<i>Spoon</i>	Wing
<i>Facet</i>	Ommatidia	<i>Spot</i>	Body color
<i>Forked</i>	Spine	<i>Tan</i>	Antenna
<i>Furrowed</i>	Eye	<i>Truncate</i>	Wing
<i>Fused</i>	Venetation	<i>Vermilion</i>	Eye color
<i>Green</i>	Body color	<i>White</i>	Eye color
<i>Jaunty</i>	Wing	<i>Yellow</i>	Body color
<i>Lemon</i>	Body color		
<b>Group 2</b>			
<i>Antlered</i>	Wing	<i>Jaunty</i>	Wing
<i>Apterous</i>	Wing	<i>Limited</i>	Abdominal band
<i>Arc</i>	Wing	<i>Little crossover</i>	Chromosome 2
<i>Balloon</i>	Venetation	<i>Morula</i>	Ommatidia
<i>Black</i>	Body color	<i>Olive</i>	Body color
<i>Blistered</i>	Wing	<i>Plexus</i>	Venetation
<i>Comma</i>	Thorax mark	<i>Purple</i>	Eye color
<i>Confluent</i>	Venetation	<i>Speck</i>	Thorax mark
<i>Cream II</i>	Eye color	<i>Strap</i>	Wing
<i>Curved</i>	Wing	<i>Streak</i>	Pattern
<i>Dachs</i>	Leg	<i>Trefoil</i>	Pattern
<i>Extra vein</i>	Venetation	<i>Truncate</i>	Wing
<i>Fringed</i>	Wing	<i>Vestigial</i>	Wing
<b>Group 3</b>			
<i>Band</i>	Pattern	<i>Pink</i>	Eye color
<i>Beaded</i>	Wing	<i>Rough</i>	Eye
<i>Cream III</i>	Eye color	<i>Safranin</i>	Eye color
<i>Deformed</i>	Eye	<i>Sepia</i>	Eye color
<i>Dwarf</i>	Size of body	<i>Sooty</i>	Body color
<i>Ebony</i>	Body color	<i>Spineless</i>	Spine
<i>Giant</i>	Size of body	<i>Spread</i>	Wing
<i>Kidney</i>	Eye	<i>Trident</i>	Pattern
<i>Low crossing over</i>	Chromosome 3	<i>Truncate</i>	Wing
<i>Maroon</i>	Eye color	<i>Whitehead</i>	Pattern
<i>Peach</i>	Eye color	<i>White ocelli</i>	Simple eye
<b>Group 4</b>			
<i>Bent</i>	Wing	<i>Eyeless</i>	Eye

The mutations fall into four linkage groups. Because four chromosomes were cytologically observed, this indicated that the genes are situated on the chromosomes. Notice that mutations in various genes can act to alter a single character, such as body color, in different ways.

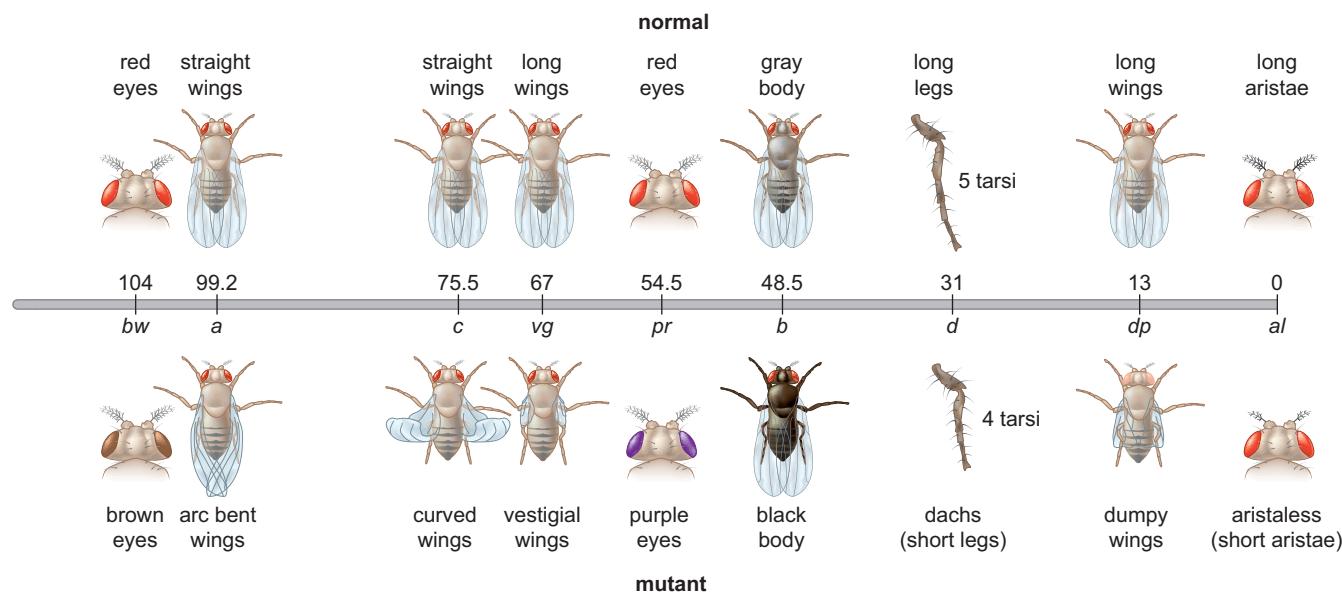


FIGURE 1-8 The genetic map of chromosome 2 of *Drosophila melanogaster*.

like wings or eyes. Instead, these biologists believed that there must also occur more powerful “macromutations,” and that it was these events that allowed great evolutionary advances.

Gradually, however, doubts vanished, largely as a result of the efforts of the mathematical geneticists Sewall Wright, Ronald A. Fisher, and John Burden Sanderson Haldane. They showed that, considering the great age of Earth, the relatively low mutation rates found for *Drosophila* genes, together with only mild selective advantages, would be sufficient to allow the gradual accumulation of new favorable attributes. By the 1930s, biologists began to reevaluate their knowledge of the origin of species and to understand the work of the mathematical geneticists. Among these new Darwinians were biologist Julian Huxley (a grandson of Darwin’s original publicist, Thomas Huxley), geneticist Theodosius Dobzhansky, paleontologist George Gaylord Simpson, and ornithologist Ernst Mayr. In the 1940s all four wrote major works, each showing from his special viewpoint how Mendelianism and Darwinism were indeed compatible.

## EARLY SPECULATIONS ABOUT WHAT GENES ARE AND HOW THEY ACT

Almost immediately after the rediscovery of Mendel’s laws, geneticists began to speculate about both the chemical structure of the gene and the way it acts. No real progress could be made, however, because the chemical identity of the genetic material remained unknown. Even the realization that both nucleic acids and proteins are present in chromosomes did not really help, since the structure of neither was at all understood. The most fruitful speculations focused attention on the fact that genes must be, in some sense, self-duplicating. Their structure must be exactly copied every time one chromosome becomes two. This fact immediately raised the profound chemical question of how a complicated molecule could be precisely copied to yield exact replicas.

Some physicists also became intrigued with the gene, and when quantum mechanics burst on the scene in the late 1920s, the possibility arose that in order to understand the gene, it would first be necessary to master the subtleties of the most advanced theoretical physics. Such thoughts, however, never really took root, since it was obvious that even the best physicists or theoretical chemists would not concern themselves with a substance whose structure still awaited elucidation. There was only one fact that they might ponder: Muller's and L.J. Stadler's independent 1927 discoveries that X-rays induce mutations. Because there is a greater possibility that an X-ray will hit a larger gene than a smaller gene, the frequency of mutations induced in a given gene by a given X-ray dose yields an estimate of the size of this gene. But even here, so many special assumptions were required that virtually no one, not even Muller and Stadler themselves, took the estimates very seriously.

## PRELIMINARY ATTEMPTS TO FIND A GENE–PROTEIN RELATIONSHIP

---

The most fruitful early endeavors to find a relationship between genes and proteins examined the ways in which gene changes affect which proteins are present in the cell. At first these studies were difficult, because no one knew anything about the proteins that were present in structures such as the eye or the wing. It soon became clear that genes with simple metabolic functions would be easier to study than genes affecting gross structures. One of the first useful examples came from a study of a hereditary disease affecting amino acid metabolism. Spontaneous mutations occur in humans affecting the ability to metabolize the amino acid phenylalanine. When individuals homozygous for the mutant trait eat food containing phenylalanine, their inability to convert the amino acid to tyrosine causes a toxic level of phenylpyruvic acid to build up in the bloodstream. Such diseases, examples of “inborn errors of metabolism,” suggested to English physician Archibald E. Garrod, as early as 1909, that the wild-type gene is responsible for the presence of a particular enzyme, and that in a homozygous mutant, the enzyme is congenitally absent.

Garrod’s general hypothesis of a gene–enzyme relationship was extended in the 1930s by work on flower pigments by Haldane and Rose Scott-Moncrieff in England, studies on the hair pigment of the guinea pig by Wright in the United States, and research on the pigments of insect eyes by A. Kuhn in Germany and by Boris Ephrussi and George W. Beadle, working first in France and then in California. In all cases, the evidence revealed that a particular gene affected a particular step in the formation of the respective pigment whose absence changed, say, the color of a fly’s eyes from red to ruby. However, the lack of fundamental knowledge about the structures of the relevant enzymes ruled out deeper examination of the gene–enzyme relationship, and no assurance could be given either that most genes control the synthesis of proteins (by then it was suspected that all enzymes were proteins) or that all proteins are under gene control.

As early as 1936, it became apparent to the Mendelian geneticists that future experiments of the sort successful in elucidating the basic features of Mendelian genetics were unlikely to yield productive evidence about how genes act. Instead, it would be necessary to find biological objects more suitable for chemical analysis. They were aware, moreover, that contemporary knowledge of nucleic acid and protein chemistry was completely inadequate for a fundamental chemical attack on even the most suitable

biological systems. Fortunately, however, the limitations in chemistry did not deter them from learning how to do genetic experiments with chemically simple molds, bacteria, and viruses. As we shall see, the necessary chemical facts became available almost as soon as the geneticists were ready to use them.

## SUMMARY

---

Heredity is controlled by chromosomes, which are the cellular carriers of genes. Hereditary factors were first discovered and described by Mendel in 1865, but their importance was not realized until the start of the 20th century. Each gene can exist in a variety of different forms called alleles. Mendel proposed that a hereditary factor (now known to be a gene) for each hereditary trait is given by each parent to each of its offspring. The physical basis for this behavior is the distribution of homologous chromosomes during meiosis: one (randomly chosen) of each pair of homologous chromosomes is distributed to each haploid cell. When two genes are on the same chromosome, they tend to be inherited together (linked). Genes affecting different characteristics are sometimes inherited independently of each other, because they are located on different chromosomes. In any case, linkage is seldom complete because homologous chromosomes attach to each other during meiosis and often break at identical spots and rejoin crossways (crossing over). Crossing over transfers genes initially located on a paternally derived chromosome onto gene groups originating from the maternal parent.

Different alleles from the same gene arise by inheritable changes (mutations) in the gene itself. Normally, genes are extremely stable and are copied exactly during chromosome duplication; mutation occurs only rarely and usually has harmful consequences. Mutation does, however, play a positive role, because the accumulation of rare favorable mutations provides the basis for genetic variability that is presupposed by the theory of evolution.

For many years, the structure of genes and the chemical ways in which they control cellular characteristics were a mystery. As soon as large numbers of spontaneous mutations had been described, it became obvious that a one gene—one characteristic relationship does not exist and that all complex characteristics are under the control of many genes. The most sensible idea, postulated by Garrod in 1909, was that genes affect the synthesis of enzymes. However, the tools of Mendelian geneticists—organisms such as the corn plant, the mouse, and even the fruit fly *Drosophila*—were not suitable for detailed chemical investigations of gene–protein relations. For this type of analysis, work with much simpler organisms was to become indispensable.

## BIBLIOGRAPHY

---

- Ayala F.J. and Kiger J.A., Jr. 1984. *Modern genetics*, 2nd ed. Benjamin Cummings, Menlo Park, California.
- Beadle G.W. and Ephrussi B. 1937. Development of eye color in *Drosophila*: Diffusible substances and their inter-relations. *Genetics* **22**: 76–86.
- Carlson E.A. 1966. *The gene: A critical history*. Saunders, Philadelphia.
- . 1981. *Genes, radiation, and society: The life and work of H.J. Muller*. Cornell University Press, Ithaca, New York.
- Caspari E. 1948. Cytoplasmic inheritance. *Adv Genet* **2**: 1–66.
- Correns C. 1937. *Nicht Mendelnde Vererbung* (ed F. von Wettstein). Borntraeger, Berlin.
- Dobzhansky T. 1941. *Genetics and the origin of species*, 2nd ed. Columbia University Press, New York.
- Fisher R.A. 1930. *The genetical theory of natural selection*. Clarendon Press, Oxford.
- Garrod A.E. 1908. Inborn errors of metabolism. *Lancet* **2**: 1–7, 73–79, 142–148, 214–220.
- Haldane J.B.S. 1932. *The courses of evolution*. Harper & Row, New York.
- Huxley J. 1943. *Evolution: The modern synthesis*. Harper & Row, New York.
- Lea D.E. 1947. *Actions of radiations on living cells*. Macmillan, New York.
- Mayr E. 1942. *Systematics and the origin of species*. Columbia University Press, New York.
- . 1982. *The growth of biological thought: Diversity, evolution, and inheritance*. Harvard University Press, Cambridge, Massachusetts.
- McClintock B. 1951. Chromosome organization and gene expression. *Cold Spring Harbor Symp. Quant. Biol.* **16**: 13–57.
- . 1984. The significance of responses of genome to challenge. *Science* **226**: 792–800.
- McClintock B. and Creighton H.B. 1931. A correlation of cytological and genetical crossing over in *Zea mays*. *Proc. Natl. Acad. Sci.* **17**: 492–497.
- Moore J. 1972a. *Heredity and development*, 2nd ed. Oxford University Press, Oxford.
- . 1972b. *Readings in heredity and development*. Oxford University Press, Oxford.
- Morgan T.H. 1910. Sex-linked inheritance in *Drosophila*. *Science* **32**: 120–122.
- Morgan T.H., Sturtevant A.H., Muller H.J., and Bridges C.B. 1915. *The mechanism of Mendelian heredity*. Holt, Rinehart & Winston, New York.
- Muller H.J. 1927. Artificial transmutation of the gene. *Science* **46**: 84–87.
- Olby R.C. 1966. *Origins of Mendelism*. Constable and Company Ltd., London.

- Peters J.A. 1959. *Classic papers in genetics*. Prentice-Hall, Englewood Cliffs, New Jersey.
- Rhoades M.M. 1946. Plastid mutations. *Cold Spring Harbor Symp. Quant. Biol.* **11**: 202–207.
- Sager R. 1972. *Cytoplasmic genes and organelles*. Academic Press, New York.
- Scott-Moncrieff R. 1936. A biochemical survey of some Mendelian factors for flower color. *J. Genetics* **32**: 117–170.
- Simpson G.G. 1944. *Tempo and mode in evolution*. Columbia University Press, New York.
- Sonneborn T.M. 1950. The cytoplasm in heredity. *Heredity* **4**: 11–36.
- Stadler L.J. 1928. Mutations in barley induced by X-rays and radium. *Science* **110**: 543–548.
- Sturtevant A.H. 1913. The linear arrangement of six sex-linked factors in *Drosophila* as shown by mode of association. *J. Exp. Zool.* **14**: 39–45.
- Sturtevant A.H. and Beadle G.W. 1962. *An introduction to genetics*. Dover, New York.
- Sutton W.S. 1903. The chromosome in heredity. *Biol. Bull.* **4**: 231–251.
- Wilson E.B. 1925. *The cell in development and heredity*, 3rd ed. Macmillan, New York.
- Wright S. 1931. Evolution in Mendelian populations. *Genetics* **16**: 97–159.
- . 1941. The physiology of the gene. *Physiol. Rev.* **21**: 487–527.

## QUESTIONS

## MasteringBiology®

For instructor-assigned tutorials and problems, go to MasteringBiology.

For answers to even-numbered questions, see Appendix 2: Answers.

**Question 1.** You are comparing two alleles of Gene X. What defines the two alleles as distinct alleles?

**Question 2.** True or false. Explain your choice. One gene possesses only two alleles.

**Question 3.** True or false. Explain your choice. One trait is always determined by one gene.

**Question 4.** True or false. Explain your choice. For a given gene, one can always define the alleles as dominant or recessive.

**Question 5.** You want to identify dominant/recessive relationship for skin color for a new frog species that you found in the rain forest. Assume that one autosomal gene controls skin color in this species. All of the frogs that you found for that species are bright blue or yellow. A bright blue female and bright blue male frog mate and produce all bright blue progeny. A yellow female and yellow male frog mate and produce a mix of bright blue and yellow progeny. Identify each trait (bright blue skin color and yellow skin color) as dominant or recessive. Explain your choices. Identify the genotype for each parent in the two crosses. Use the letter *B* to refer to the gene conferring skin color.

### Question 6.

- After crossing true-breeding pea plants with yellow seeds to true-breeding pea plants with green seeds as Mendel did, what phenotype do you expect for the pea plants in the F<sub>1</sub> generation if yellow seeds are dominant to green seeds?
- You self-cross the F<sub>1</sub> generation. Give the expected phenotypic ratio of the F<sub>2</sub> generation.
- Give the expected genotypic ratio of the F<sub>2</sub> generation.
- Give the expected ratio of heterozygotes to homozygotes in the F<sub>2</sub> generation.

**Question 7.** Mendel studied seven distinct traits for pea plants. By luck six of the traits were on different chromosomes, and two traits were separated by a great distance on one chromosome. If Mendel selected two traits controlled by linked genes in his

initial studies, which law would be affected (Mendel's first or second law)? Explain your choice.

**Question 8.** You want to map the positions of three genes (*X*, *Y*, and *Z*) all found on one chromosome in *Drosophila*. Each gene has one dominant allele and one recessive allele. You perform the three different two-factor crosses (Cross 1: *XY* and *xy*, Cross 2: *YZ* and *yz*, and Cross 3: *XZ* and *xz*). Assume all crosses are between diploid flies homozygous for the alleles of these genes. You observe 7% recombinants in the first cross, 20% recombinants in the second cross, and 13% recombinants in the third cross. Draw a map placing the genes in the proper order and give the distance between each gene in map units (m.u.).

**Question 9.** You want to confirm your ordering for Question 8 using a three-factor cross (cross *XYZ/xyz* and *xyz/xyz*). Your least common recombinants are *xYZ* and *Xyz*. Does this confirm your order from Question 8? Explain why or why not.

**Question 10.** You again want to map the positions of three genes (*L*, *M*, and *N*) in *Drosophila*. Each gene has one dominant allele and one recessive allele. You perform the three different two-factor crosses (Cross 1: *LM* and *lm*, Cross 2: *MN* and *mn*, and Cross 3: *LN* and *ln*). Assume all crosses are between diploid flies homozygous for the alleles of these genes. You observe 5% recombinants in the first cross, 50% recombinants in the second cross, and 50% recombinants in the third cross. Based on the data given, what can you determine for the gene order and distance between the genes?

**Question 11.** Following up on the observations in Question 10, you complete new crosses using gene *O*. You observe 30% recombination for a cross between *MO* and *mo*, 35% recombination for a cross between *LO* and *lo*, and 25% recombination for a cross between *NO* and *no*. Assume all crosses are between diploid flies homozygous for the alleles of these genes. Given the information from Questions 10 and 11, draw a map placing the genes in the proper order and give the distance between each gene in map units.

**Question 12.** Define mutation. The cell has many mechanisms to prevent mutations. Explain how a very low mutation rate could be advantageous over the prevention of all mutations in an organism.

**Question 13.** Differentiate between chromosomes and chromatids.

**Question 14.** You are mapping the 6th chromosome of the sheep blowfly *Lucilia cuprina* and want to test how your calculations compare to a published map. In a recent cross, you studied the mutations *tri*, *pk*, and *y* that display thickened vein junctions, pink body color, and yellow eyes, respectively. From a cross between a male homozygous for the three mutations and a heterozygous female (*tri pk y/+++*), you record the counts for the progeny. In the published map, the distance between *y* and *pk* is 23.0 m.u., the distance between *pk* and *tri* is 18.4 m.u., and the distance between *y* and *tri* is 41.4 m.u. Based on the published map and given values below, calculate the expected values for observed progeny that represent either a single or double crossover. Remember that your observed values are data that include some statistical fluctuations.

Total progeny counted: 1000

Total recombinants that represent a double crossover: 15

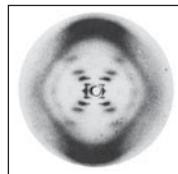
Published map information from Weller and Foster (1993. *Genome* **36**: 495–506).

**Question 15.** You are studying a new species of bird. You know the species has sex chromosomes similar to chicken. Males carry two Z chromosomes, whereas females carry one Z chromosome and one W chromosome. Because the genome has not been sequenced yet, you will perform crosses to gain more genetic information. You are interested in the eye color of the birds. You obtain true-breeding birds with black or green eyes. You cross a black-eyed male to a green-eyed female. Assume the trait is determined by one gene.

- A. Considering a dominant/recessive relationship, you want to determine if black is recessive to green or if green is recessive to black. How could you use the phenotypes for the F<sub>1</sub> and F<sub>2</sub> progeny to help you answer this question?
- B. If the trait is sex-linked, refine your answer to part A with respect to the dominant/recessive relationship of a sex-linked trait on the Z chromosome for the F<sub>1</sub> generation.
- C. Assume that black is dominant to green. You cross a black-eyed male from the F<sub>1</sub> generation to a black-eyed female from the F<sub>1</sub> generation. If the trait is sex-linked, predict the genotypic and phenotypic ratios for the F<sub>2</sub> generation.

*This page intentionally left blank*

CHAPTER 2



## Nucleic Acids Convey Genetic Information

THAT SPECIAL MOLECULES MIGHT CARRY genetic information was appreciated by geneticists long before the problem claimed the attention of chemists. By the 1930s, geneticists began speculating as to what sort of molecules could have the kind of stability that the gene demanded, yet be capable of permanent, sudden change to the mutant forms that must provide the basis of evolution. Until the mid-1940s, there appeared to be no direct way to attack the chemical essence of the gene. It was known that chromosomes possessed a unique molecular constituent, deoxyribonucleic acid (DNA). Despite this, there was no way to show that DNA carried genetic information, as opposed to serving merely as a molecular scaffold for a still undiscovered class of proteins especially tailored to carry genetic information. It was generally assumed that genes would be composed of amino acids because, at that time, they appeared to be the only biomolecules with sufficient complexity to convey genetic information.

It therefore made sense to approach the nature of the gene by asking how genes function within cells. In the early 1940s, research on the mold *Neurospora*, spearheaded by George W. Beadle and Edward Tatum, was generating increasingly strong evidence supporting the 30-year-old hypothesis of Archibald E. Garrod that genes work by controlling the synthesis of specific enzymes (the one gene–one enzyme hypothesis). Thus, given that all known enzymes had, by this time, been shown to be proteins, the key problem was the way genes participate in the synthesis of proteins. From the very start of serious speculation, the simplest hypothesis was that genetic information within genes determines the order of the 20 different amino acids within the polypeptide chains of proteins.

In attempting to test this proposal, intuition was of little help even to the best biochemists, because there is no logical way to use enzymes as tools to determine the order of each amino acid added to a polypeptide chain. Such schemes would require, for the synthesis of a single type of protein, as many ordering enzymes as there are amino acids in the respective protein. But because all enzymes known at that time were themselves proteins (we now know that RNA can also act as an enzyme), still additional ordering enzymes would be necessary to synthesize the ordering enzymes. This situation clearly poses a paradox, unless we assume a fantastically interrelated series of syntheses in which a given protein has many different enzymatic

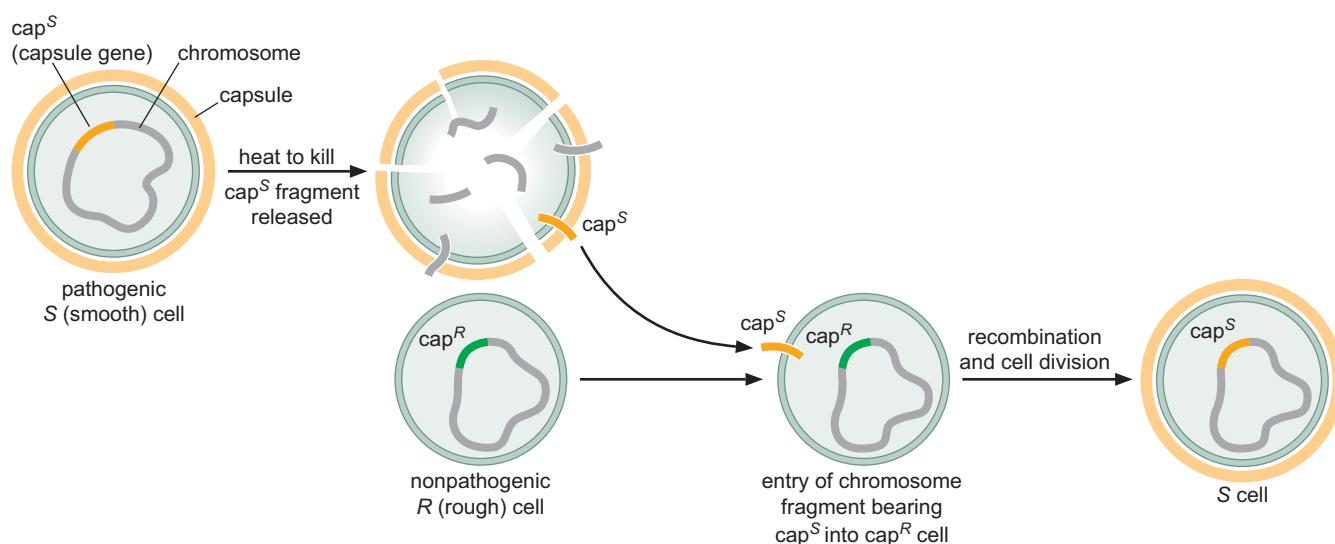
### OUTLINE

- Avery's Bombshell: DNA Can Carry Genetic Specificity, 22
  - The Double Helix, 24
- The Genetic Information within DNA Is Conveyed by the Sequence of Its Four Nucleotide Building Blocks, 30
  - The Central Dogma, 33
- Establishing the Direction of Protein Synthesis, 38
  - The Era of Genomics, 40
- Visit Web Content for Structural Tutorials and Interactive Animations

specificities. With such an assumption, it might be possible (and then only with great difficulty) to visualize a workable cell. It did not seem likely, however, that most proteins would be found to carry out multiple tasks. In fact, all the current knowledge pointed to the opposite conclusion of one protein, one function.

### AVERY'S BOMBSHELL: DNA CAN CARRY GENETIC SPECIFICITY

The idea that DNA might be the key genetic molecule emerged most unexpectedly from studies on pneumonia-causing bacteria. In 1928 English microbiologist Frederick Griffith made the startling observation that nonvirulent strains of the bacteria became virulent when mixed with their heat-killed pathogenic counterparts. That such **transformations** from nonvirulence to virulence represented hereditary changes was shown by using descendants of the newly pathogenic strains to transform still other nonpathogenic bacteria. This raised the possibility that, when pathogenic cells are killed by heat, their genetic components remain undamaged. Moreover, once liberated from the heat-killed cells, these components can pass through the cell wall of the living recipient cells and undergo subsequent genetic recombination with the recipient's genetic apparatus (Fig. 2-1). Subsequent research has confirmed this genetic interpretation. Pathogenicity reflects the action of the capsule gene, which codes for a key enzyme involved in the synthesis of the carbohydrate-containing capsule that surrounds most pneumonia-causing bacteria. When the *S* (smooth) allele of the capsule gene is present, a capsule is formed around the cell that is necessary for pathogenesis (the formation of a capsule also gives a smooth appearance to the colonies formed from these cells). When the *R*



**FIGURE 2-1** Transformation of a genetic characteristic of a bacterial cell (*Streptococcus pneumoniae*) by addition of heat-killed cells of a genetically different strain. Here we show an *R* cell receiving a chromosomal fragment containing the capsule gene from a heat-treated *S* cell. Since most *R* cells receive other chromosomal fragments, the efficiency of transformation for a given gene is usually less than 1%.

(rough) allele of this gene is present, no capsule is formed, the respective cells are not pathogenic, and the colonies these cells are round around the edges.

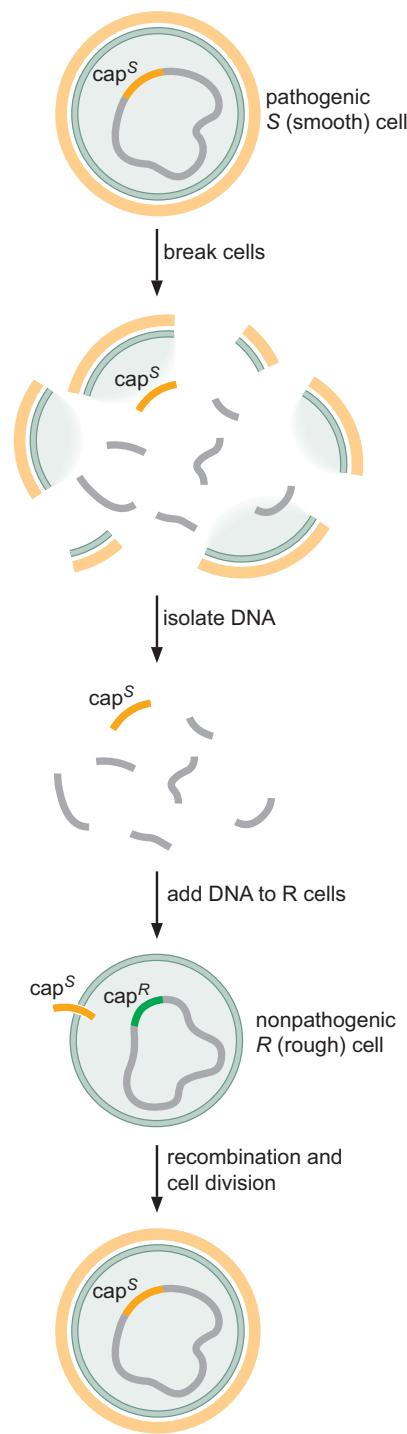
Within several years after Griffith's original observation, extracts of the killed bacteria were found capable of inducing hereditary transformations, and a search began for the chemical identity of the transforming agent. At that time, the vast majority of biochemists still believed that genes were proteins. It therefore came as a great surprise when in 1944, after some 10 years of research, U.S. microbiologist Oswald T. Avery and his colleagues at the Rockefeller Institute in New York, Colin M. MacLeod and Maclyn McCarty, made the momentous announcement that the active genetic principle was DNA (Fig. 2-2). Supporting their conclusion were key experiments showing that the transforming activity of their highly purified active fractions was destroyed by deoxyribonuclease, a recently purified enzyme that specifically degrades DNA molecules to their nucleotide building blocks but has no effect on the integrity of protein molecules or RNA. In contrast, the addition of either ribonuclease (which degrades RNA) or various proteolytic enzymes (which degrade proteins) had no influence on the transforming activity.

### Viral Genes Are Also Nucleic Acids

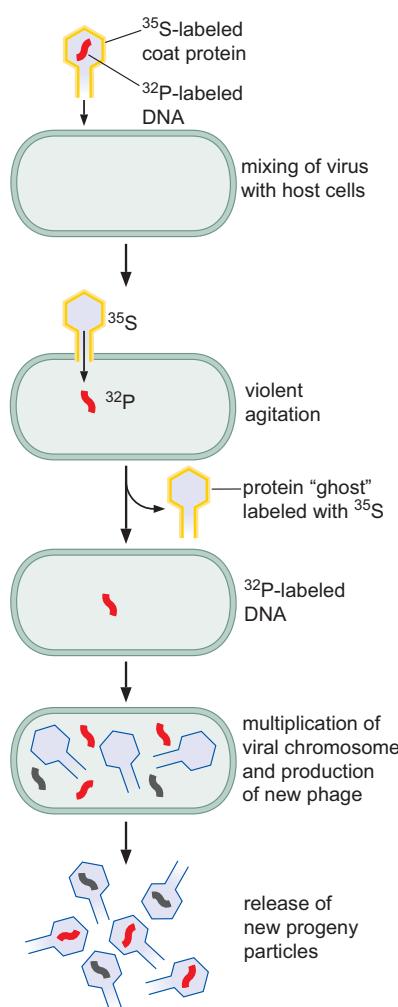
Equally important confirmatory evidence came from chemical studies with viruses and virus-infected cells. By 1950 it was possible to obtain a number of essentially pure viruses and to determine which types of molecules were present in them. This work led to the very important generalization that all viruses contain nucleic acid. Because there was at that time a growing realization that viruses contain genetic material, the question immediately arose as to whether the nucleic acid component was the carrier of viral genes. A crucial test of the question came from isotopic study of the multiplication of T2, a bacterial virus (typically called a **bacteriophage, or phage**) composed of a DNA core and a protective shell built up by the aggregation of a number of different protein molecules. In these experiments, performed in 1952 by Alfred D. Hershey and Martha Chase working at Cold Spring Harbor Laboratory in Long Island, New York, the protein coat was labeled with the radioactive isotope  $^{35}\text{S}$  and the DNA core with the radioactive isotope  $^{32}\text{P}$ . The labeled virus was then used to follow the fates of the phage protein and nucleic acid as phage multiplication proceeded, particularly to see which labeled atoms from the parental phage entered the host cell and later appeared in the progeny phage.

Clear-cut results emerged from these experiments; much of the parental nucleic acid and none of the parental protein was detected in the progeny phage (Fig. 2-3). Moreover, it was possible to show that little of the parental protein even enters the bacteria; instead, it stays attached to the outside of the bacterial cell, performing no function after the DNA component has passed inside. This point was neatly shown by violently agitating infected bacteria after the entrance of the DNA; the protein coats were shaken off without affecting the ability of the bacteria to form new phage particles.

With some viruses it is now possible to do an even more convincing experiment. For example, purified DNA from the mouse polyoma virus can enter mouse cells and initiate a cycle of viral multiplication producing many thousands of new polyoma particles. The primary function of viral protein is thus to protect and transport its genetic/nucleic acid component in its movement from one cell to another.



**FIGURE 2-2** Isolation of a chemically pure transforming agent. (Adapted, with permission, from Stahl F.W. 1964. *The mechanics of inheritance*, Fig. 2.3. © Pearson Education, Inc.)



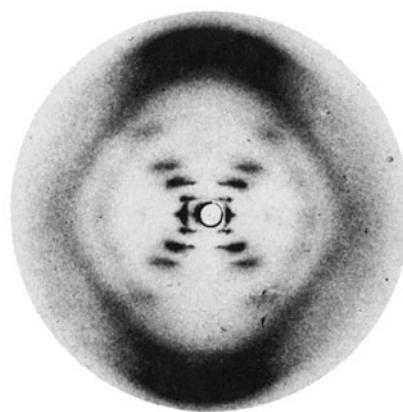
**FIGURE 2-3** Demonstration that only the DNA component of the bacteriophage T2 carries the genetic information and that the protein coat serves only as a protective shell.

## THE DOUBLE HELIX

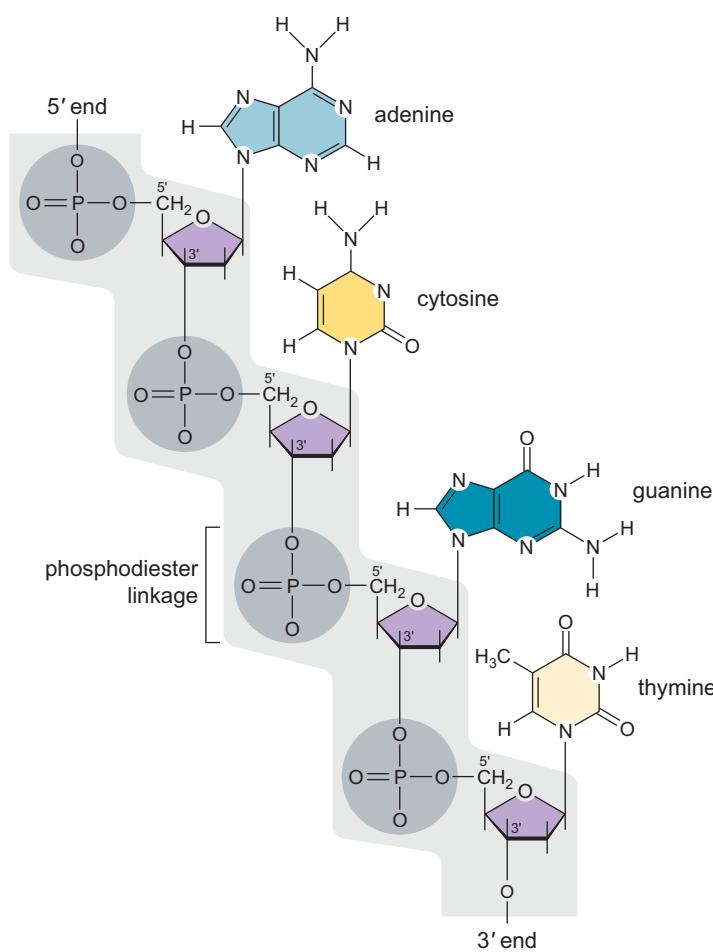
While work was proceeding on the X-ray analysis of protein structure, a smaller number of scientists were trying to solve the X-ray diffraction pattern of DNA. The first diffraction patterns were taken in 1938 by William Astbury using DNA supplied by Ola Hammarsten and Torbjörn Caspersson. It was not until the early 1950s that high-quality X-ray diffraction photographs were taken by Maurice Wilkins and Rosalind Franklin (Fig. 2-4). These photographs suggested not only that the underlying DNA structure was helical but that it was composed of more than one polynucleotide chain—either two or three. At the same time, the covalent bonds of DNA were being unambiguously established. In 1952 a group of organic chemists working in the laboratory of Alexander Todd showed that 3'-5' phosphodiester bonds regularly link together the nucleotides of DNA (Fig. 2-5).

In 1951, because of interest in Linus Pauling's  $\alpha$  helix protein motif (which we shall consider in Chapter 6), an elegant theory of diffraction of helical molecules was developed by William Cochran, Francis H. Crick, and Vladimir Vand. This theory made it easy to test possible DNA structures on a trial-and-error basis. The correct solution, a complementary double helix (see Chapter 4), was found in 1953 by Crick and James D. Watson, then working in the laboratory of Max Perutz and John Kendrew in Cambridge, United Kingdom. Their arrival at the correct answer depended largely on finding the stereochemically most favorable configuration compatible with the X-ray diffraction data of Wilkins and Franklin.

In the double helix, the two DNA chains are held together by hydrogen bonds (a weak noncovalent chemical bond; see Chapter 3) between pairs of bases on the opposing strands (Fig. 2-6). This base pairing is very specific: the purine adenine only base-pairs to the pyrimidine thymine, whereas the purine guanine only base-pairs to the pyrimidine cytosine. In double-helical DNA, the number of A residues must be equal to the



**FIGURE 2-4** The key X-ray photograph involved in the elucidation of the DNA structure. This photograph, taken by Rosalind Franklin at King's College, London, in the winter of 1952–1953, confirmed the guess that DNA was helical. The helical form is indicated by the crossways pattern of X-ray reflections (photographically measured by darkening of the X-ray film) in the center of the photograph. The very heavy black regions at the top and bottom reveal that the 3.4-Å-thick purine and pyrimidine bases are regularly stacked next to each other, perpendicular to the helical axis. (Printed, with permission, from Franklin R.E. and Gosling R.G. 1953. *Nature* 171: 740–741. © Macmillan.)



**FIGURE 2-5** A portion of a DNA polynucleotide chain, showing the 3' → 5' phosphodiester linkages that connect the nucleotides. Phosphate groups connect the 3' carbon of one nucleotide with the 5' carbon of the next.

number of T residues, whereas the number of G and C residues must likewise be equal (see Box 2-1, Chargaff's Rules). As a result, the sequence of the bases of the two chains of a given double helix have a complementary relationship, and the sequence of any DNA strand exactly defines that of its partner strand.

The discovery of the double helix initiated a profound revolution in the way many geneticists analyzed their data. The gene was no longer a mysterious entity, the behavior of which could be investigated only by genetic experiments. Instead, it quickly became a real molecular object about which chemists could think objectively, as they did about smaller molecules such as pyruvate and ATP. Most of the excitement, however, came not merely from the fact that the structure was solved, but also from the nature of the structure. Before the answer was known, there had always been the worry that it would turn out to be dull, revealing nothing about how genes replicate and function. Fortunately, the answer was immensely exciting. The two intertwined strands of complementary structures suggested that one strand serves as the specific surface (template) upon which the other strand is made (Fig. 2-6). If this hypothesis were true, then the fundamental problem of gene replication, about which geneticists had puzzled for so many years, was, in fact, conceptually solved.



**FIGURE 2-6** The replication of DNA. The newly synthesized strands are shown in orange.

## ► KEY EXPERIMENTS

### Box 2-1 Chargaff's Rules

Biochemist Erwin Chargaff used a technique called “paper chromatography” to analyze the nucleotide composition of DNA. By 1949 his data showed not only that the four different nucleotides are not present in equal amounts, but also that the exact ratios of the four nucleotides vary from one species to another (Box 2-1 Table 1). These findings opened up the possibility that it is the precise arrangement of nucleotides within a DNA molecule that confers its genetic specificity.

Chargaff’s experiments also showed that the relative ratios of the four bases were not random. The number of adenine (A)

residues in all DNA samples was equal to the number of thymine (T) residues, and the number of guanine (G) residues equaled the number of cytosine (C) residues. In addition, regardless of the DNA source, the ratio of purines to pyrimidines was always approximately 1 (purines = pyrimidines). The fundamental significance of the A = T and G = C relationships (Chargaff’s rules) could not emerge, however, until serious attention was given to the three-dimensional structure of DNA.

**BOX 2-1 TABLE 1** Data Leading to the Formulation of Chargaff’s Rules

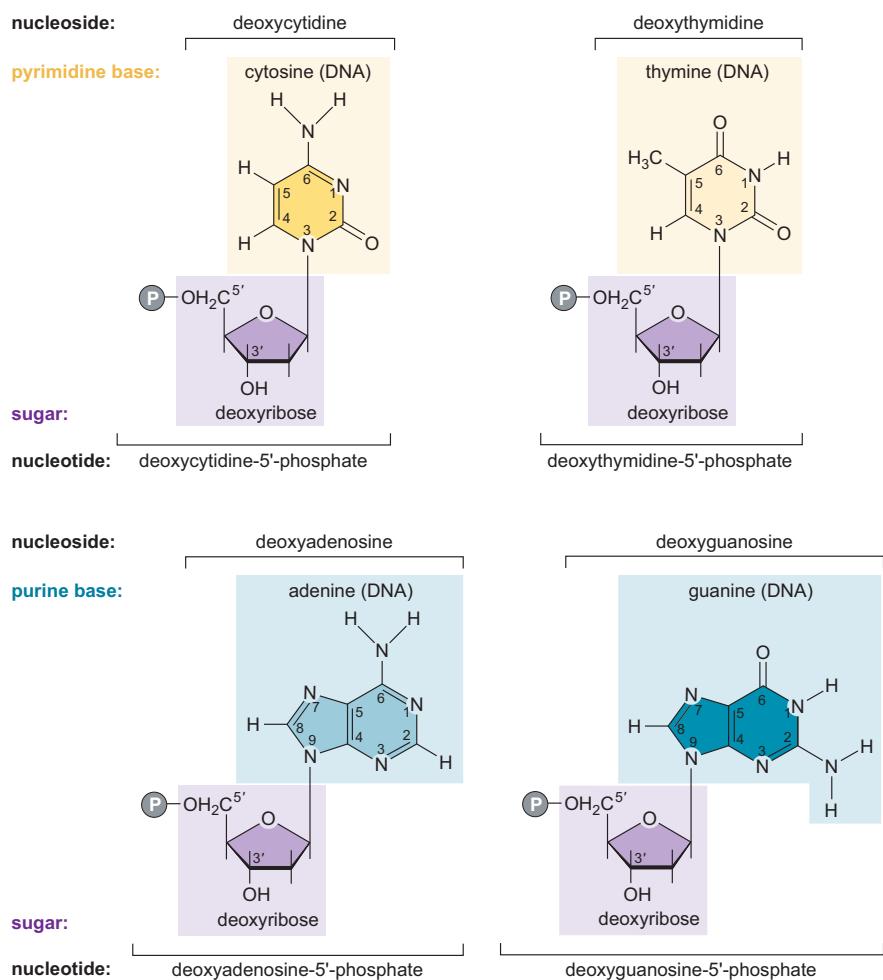
Source	Adenine to Guanine	Thymine to Cytosine	Adenine to Thymine	Guanine to Cytosine	Purines to Pyrimidines
Ox	1.29	1.43	1.04	1.00	1.1
Human	1.56	1.75	1.00	1.00	1.0
Hen	1.45	1.29	1.06	0.91	0.99
Salmon	1.43	1.43	1.02	1.02	1.02
Wheat	1.22	1.18	1.00	0.97	0.99
Yeast	1.67	1.92	1.03	1.20	1.0
<i>Hemophilus influenzae</i>	1.74	1.54	1.07	0.91	1.0
<i>Escherichia coli</i> K2	1.05	0.95	1.09	0.99	1.0
Avian tubercle bacillus	0.4	0.4	1.09	1.08	1.1
<i>Serratia marcescens</i>	0.7	0.7	0.95	0.86	0.9
<i>Bacillus schatz</i>	0.7	0.6	1.12	0.89	1.0

After Chargaff E. et al. 1949. *J. Biol. Chem.* 177: 405.

## Finding the Polymerases That Make DNA

Rigorous proof that a single DNA chain is the template that directs the synthesis of a complementary DNA chain had to await the development of test-tube (*in vitro*) systems for DNA synthesis. These came much faster than anticipated by molecular geneticists, whose world until then had been far removed from that of the biochemist well versed in the procedures needed for enzyme isolation. Leading this biochemical assault on DNA replication was U.S. biochemist Arthur Kornberg, who by 1956 had demonstrated DNA synthesis in cell-free extracts of bacteria. Over the next several years, Kornberg went on to show that a specific polymerizing enzyme was needed to catalyze the linking together of the building-block precursors of DNA. Kornberg’s studies revealed that the nucleotide building blocks for DNA are energy-rich precursors (dATP, dGTP, dCTP, and dTTP; Fig. 2-7). Further studies identified a single polypeptide, DNA polymerase I (DNA Pol I), that was capable of catalyzing the synthesis of new DNA strands. It links the nucleotide precursors by 3'-5' phosphodiester bonds (Fig. 2-8). Furthermore, it works only in the presence of DNA, which is needed to order the four nucleotides in the polynucleotide product.

DNA Pol I depends on a DNA template to determine the sequence of the DNA it is synthesizing. This was first demonstrated by allowing the enzyme

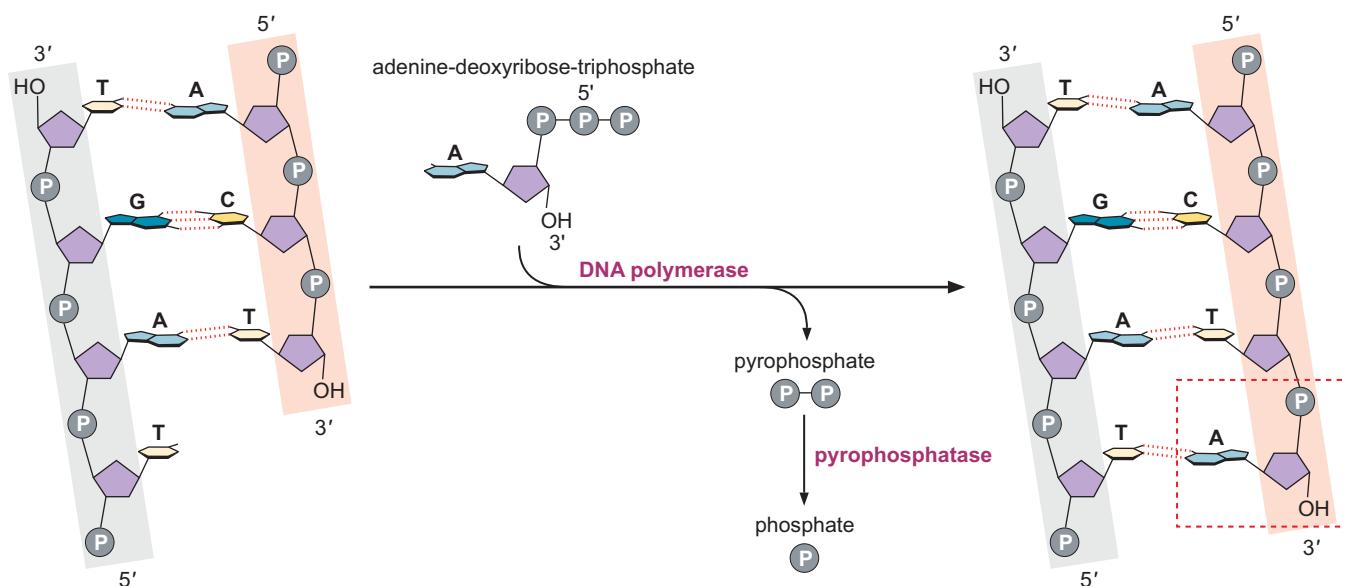


**FIGURE 2-7** The nucleotides of DNA. The structures of the different components of each of the four nucleotides are shown.

to work in the presence of DNA molecules that contained varying amounts of A:T and G:C base pairs. In every case, the enzymatically synthesized product had the base ratios of the template DNA (Table 2-1). During this cell-free synthesis, no synthesis of proteins or any other molecular class occurs, unambiguously eliminating any non-DNA compounds as intermediate carriers of genetic specificity. Thus, there is no doubt that DNA is the direct template for its own formation.

### Experimental Evidence Favors Strand Separation during DNA Replication

Simultaneously with Kornberg's research, in 1958 Matthew Meselson and Franklin W. Stahl, then at the California Institute of Technology, carried out an elegant experiment in which they separated daughter DNA molecules and, in so doing, showed that the two strands of the double helix permanently separate from each other during DNA replication (Fig. 2-9). Their success was due in part to the use of the heavy isotope  $^{15}\text{N}$  as a tag to differentially label the parental and daughter DNA strands. Bacteria grown in a medium containing the heavy isotope  $^{15}\text{N}$  have denser DNA than bacteria grown under normal conditions with  $^{14}\text{N}$ . Also contributing

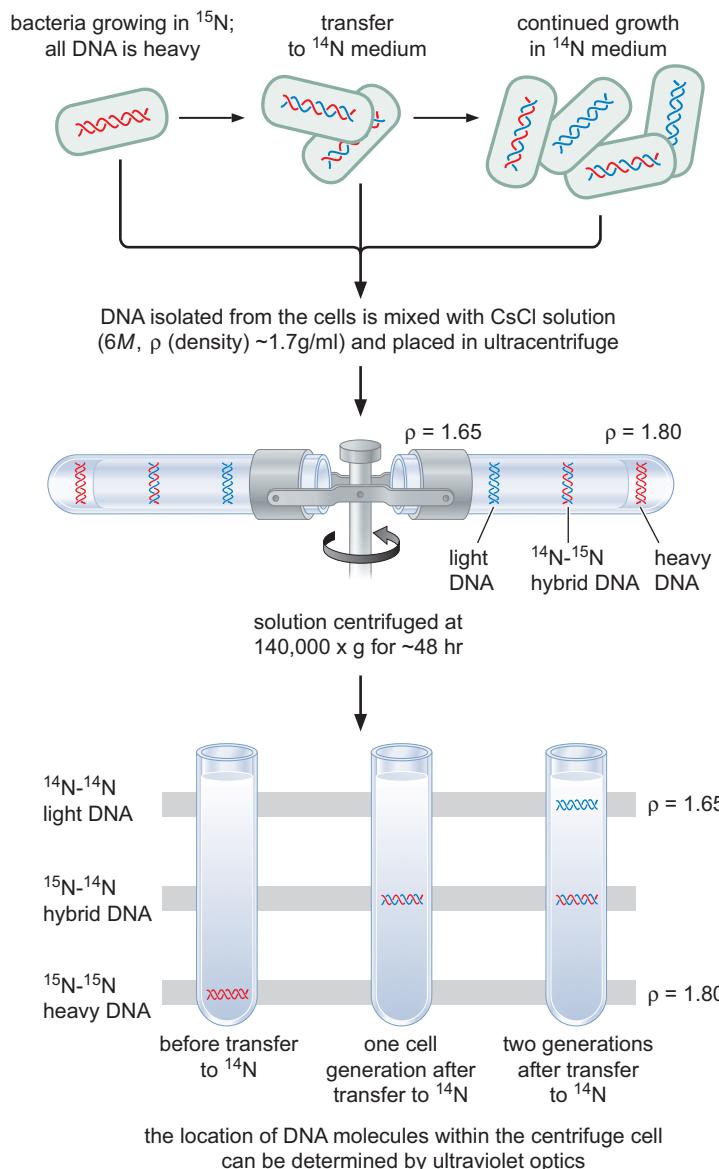


**FIGURE 2-8** Enzymatic synthesis of a DNA chain catalyzed by DNA polymerase I. This image shows the addition of a nucleotide to a growing DNA strand as catalyzed by DNA polymerase. Although the DNA polymerase can catalyze DNA synthesis by itself, in the cell the released pyrophosphate molecule is rapidly converted to two phosphates by an enzyme called pyrophosphatase, making the forward reaction of nucleotide addition even more favorable.

to the success of the experiment was the development of procedures for separating heavy DNA from light DNA in density gradients of heavy salts like cesium chloride. When high centrifugal forces are applied, the solution becomes more dense at the bottom of the centrifuge tube (which, when spinning, is the farthest from the axis of rotation). When the correct initial solution density is chosen, the individual DNA molecules will move to the central region of the centrifuge tube, where their density equals that of the salt solution. In this situation, DNA molecules in which both strands are composed of entirely  $^{15}\text{N}$  precursors (heavy-heavy or HH DNA) will form a band at a higher density (closer to the bottom of the tube) than DNA molecules in which both strands are composed entirely of  $^{14}\text{N}$  precursors (light-light or LL DNA). If bacteria containing heavy DNA are transferred to a light medium (containing  $^{14}\text{N}$ ) and allowed to grow, the precursor nucleotides available for use in DNA synthesis will be light; hence, DNA synthesized after transfer will be distinguishable from DNA made before transfer.

**TABLE 2-1** A Comparison of the Base Composition of Enzymatically Synthesized DNAs and Their DNA Templates

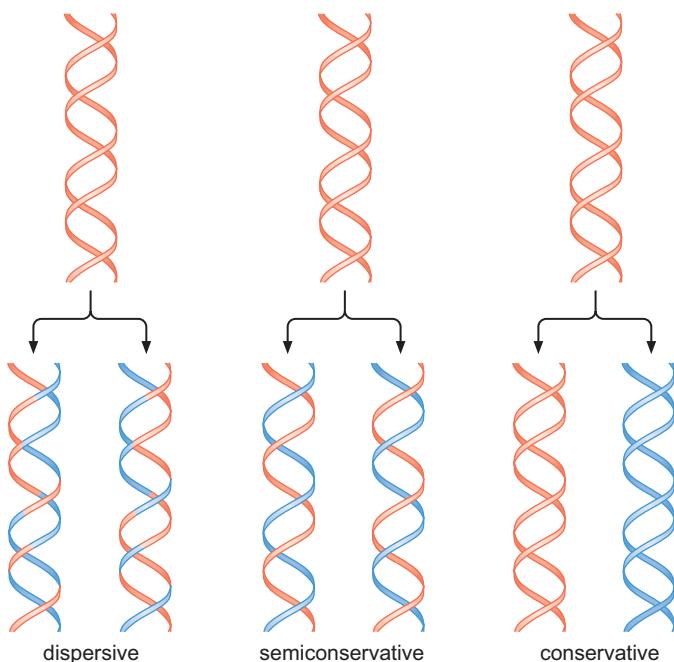
Source of DNA Template	Base Composition of the Enzymatic Product				$\frac{\text{A} + \text{T}}{\text{G} + \text{C}}$	$\frac{\text{A} + \text{T}}{\text{G} + \text{C}}$
	Adenine	Thymine	Guanine	Cytosine	In Product	In Template
<i>Micrococcus lysodeikticus</i> (a bacterium)	0.15	0.15	0.35	0.35	0.41	0.39
<i>Aerobacter aerogenes</i> (a bacterium)	0.22	0.22	0.28	0.28	0.80	0.82
<i>Escherichia coli</i>	0.25	0.25	0.25	0.25	1.00	0.97
Calf thymus	0.29	0.28	0.21	0.22	1.32	1.35
Phage T2	0.32	0.32	0.18	0.18	1.78	1.84



**FIGURE 2-9** Use of a cesium chloride (CsCl) density gradient to demonstrate the separation of complementary strands during DNA replication.

If DNA replication involves strand separation, definite predictions can be made about the density of the DNA molecules found after various growth intervals in a light medium. After one generation of growth, all the DNA molecules should contain one heavy strand and one light strand and thus be of intermediate density (heavy–light or HL DNA). This result is exactly what Meselson and Stahl observed. Likewise, after two generations of growth, half the DNA molecules were light and half hybrid, just as strand separation predicts. It is important to note that during isolation from the bacteria the DNA was broken into small fragments, which ensured that the vast majority of the DNA was either fully replicated or not replicated at all. If the entire bacterial genome were maintained intact, then there would have been many intermediate-density molecules (neither HH, HL, nor LL) that were only partially replicated.

Thus, Meselson and Stahl's experiments showed that DNA **replication** is a semiconservative process in which the single strands of the double helix remain intact (are conserved) during a replication process that distributes one parental strand into each of the two daughter molecules (thus the



**FIGURE 2-10** Three possible mechanisms for DNA replication. When the structure of DNA was discovered, several models were proposed to explain how it was replicated; three are illustrated here. The experiments proposed by Meselson and Stahl clearly distinguished among these models, demonstrating that DNA was replicated semiconservatively.

“semi” in semiconservative). These experiments ruled out two other models at the time: the conservative and the dispersive replication schemes (Fig. 2-10). In the conservative model, both of the parental strands were proposed to remain together and the two new strands of DNA would form an entirely new DNA molecule. In this model, fully light DNA would be formed after one cell generation. In the dispersive model, which was favored by many at the time, the DNA strands were proposed to be broken as frequently as every ten base pairs and used to prime the synthesis of similarly short regions of DNA. These short DNA fragments would subsequently be joined to form complete DNA strands. In this complex model, all DNA strands would be composed of both old and new DNA (thus nonconservative) and fully light DNA would only be observed after many generations of growth.

### THE GENETIC INFORMATION WITHIN DNA IS CONVEYED BY THE SEQUENCE OF ITS FOUR NUCLEOTIDE BUILDING BLOCKS

The finding of the double helix had effectively ended any controversy about whether DNA was the primary genetic substance. Even before strand separation during DNA replication was experimentally verified, the main concern of molecular geneticists had turned to how the genetic information of DNA functions to order amino acids during protein synthesis (see Box 2-2, Evidence That Genes Control Amino Acid Sequences in Proteins). With all DNA chains capable of forming double helices, the essence of their genetic specificity had to reside in the linear sequences of their four nucleotide building blocks. Thus, as information-containing entities, DNA molecules were by then properly regarded as very long words (as we shall see later, they are now best considered very long sentences) built up from a four-letter alphabet (A, G, C, and T). Even with only four letters, the number of potential DNA sequences ( $4^N$ , where  $N$  is the number of letters in the sequence) is

very, very large for even the smallest of DNA molecules; a virtually infinite number of different genetic messages can exist. Now we know that a typical bacterial gene is made up of approximately 1000 base pairs. The number of potential genes of this size is  $4^{1000}$ , a number that is orders of magnitude larger than the number of known genes in any organism.

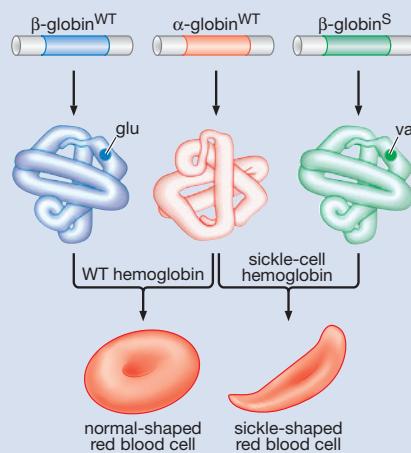
## ► KEY EXPERIMENTS

### Box 2-2 Evidence That Genes Control Amino Acid Sequences in Proteins

The first experimental evidence that genes (DNA) control amino acid sequences arose from the study of the hemoglobin present in humans suffering from the genetic disease sickle-cell anemia. If an individual has the *S* allele of the  $\beta$ -globin gene (which encodes one of the two polypeptides that together form hemoglobin) present in both homologous chromosomes (*SS*), a severe anemia results, characterized by the red blood cells having a sickle-cell shape. If only one of the two alleles of the  $\beta$ -globin gene are of the *S* form (+*S*), the anemia is less severe and the red blood cells appear almost normal in shape. The type of hemoglobin in red blood cells correlates with the genetic pattern. In the *SS* case, the hemoglobin is abnormal, characterized by a solubility different from that of normal hemoglobin, whereas in the +*S* condition, half the hemoglobin is normal and half abnormal.

Wild-type hemoglobin molecules are constructed from two kinds of polypeptide chains:  $\alpha$  chains and  $\beta$  chains (see Box 2-2 Fig. 1). Each chain has a molecular weight of about 16,100 daltons (D). Two  $\alpha$  chains and two  $\beta$  chains are present in each molecule, giving hemoglobin a molecular weight of about 64,400 D. The  $\alpha$  chains and  $\beta$  chains are controlled by distinct genes so that a single mutation will affect either the  $\alpha$  chain or the  $\beta$  chain, but not both. In 1957, Vernon M. Ingram at Cambridge University showed that sickle hemoglobin differs from normal hemoglobin by the change

of one amino acid in the  $\beta$  chain: at position 6, the glutamic acid residue found in wild-type hemoglobin is replaced by valine. Except for this one change, the entire amino acid sequence is identical in normal and mutant hemoglobin. Because this change in amino acid sequence was observed only in patients with the *S* allele of the  $\beta$ -globin gene, the simplest hypothesis is that the *S* allele of the gene encodes the change in the  $\beta$ -globin gene. Subsequent studies of amino acid sequences in hemoglobin isolated from other forms of anemia completely supported this proposal; sequence analysis showed that each specific anemia is characterized by a single amino acid replacement at a unique site along the polypeptide chain (Box 2-2 Fig. 2).



**BOX 2-2 FIGURE 1** Formation of wild-type and sickle-cell hemoglobin. (Source of hemoglobin structures: Illustration, Irving Geis. Rights owned by Howard Hughes Medical Institute. Not to be reproduced without permission.)

$\alpha$ chain									
position	1	2	16	30	57	58	68	141	
amino acid	Val	Leu	Lys <sup>+</sup>	Glu <sup>-</sup>	Gly	His <sup>+</sup>	AspN	Arg	
Hb variant	Hb I								
	Asp <sup>-</sup>								
	Hb G Honolulu								
	GluN								
	Hb Norfolk								
Hb M Boston	Asp <sup>-</sup>								
	Tyr								
Hb G Philadelphia	Lys <sup>+</sup>								

$\beta$ chain											
position	1	2	3	6	7	26	63	67	125	150	
amino acid	Val	His <sup>+</sup>	Leu	Glu <sup>-</sup>	Glu <sup>-</sup>	Glu <sup>-</sup>	His <sup>+</sup>	Val	Glu	His <sup>+</sup>	
Hb variant	Hb S										
	Val										
	Hb C										
	Lys <sup>+</sup>										
	Hb G San José										
Hb E	Gly										
	Lys <sup>+</sup>										
	Hb M Saskatoon										
	Tyr										
Hb Zürich	Arg <sup>+</sup>										
	Hb M Milwaukee-1										
	Glu <sup>-</sup>										
Hb D $\beta$ Punjab	GluN										

**BOX 2-2 FIGURE 2** A summary of some established amino acid substitutions in human hemoglobin variants.

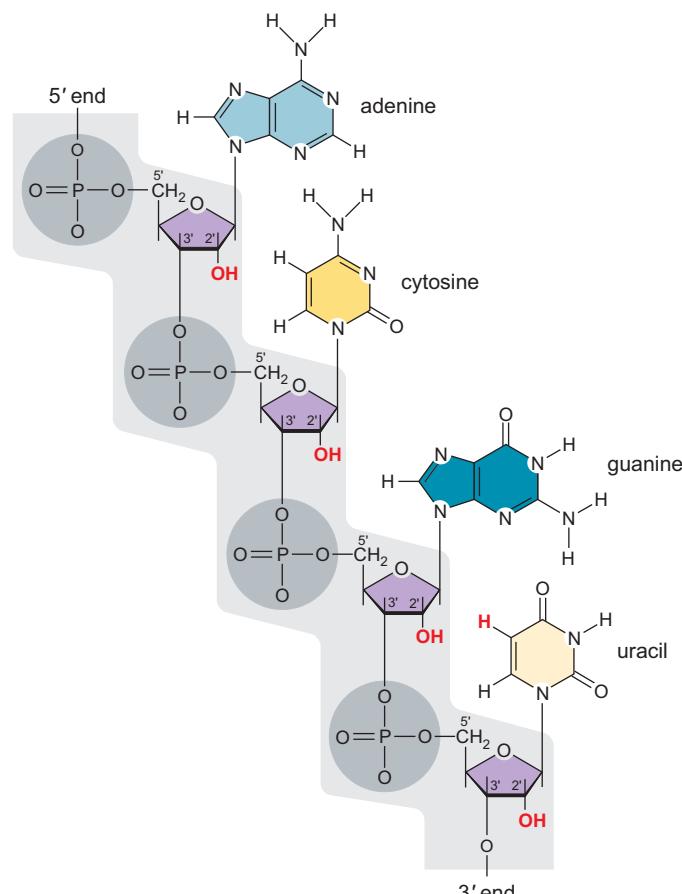
## DNA Cannot Be the Template That Directly Orders Amino Acids during Protein Synthesis

Although DNA must carry the information for ordering amino acids, it was quite clear that the double helix itself could not be the template for protein synthesis. Experiments showing that protein synthesis occurs at sites where DNA is absent ruled out a direct role for DNA. Protein synthesis in all eukaryotic cells occurs in the cytoplasm, which is separated by the nuclear membrane from the chromosomal DNA.

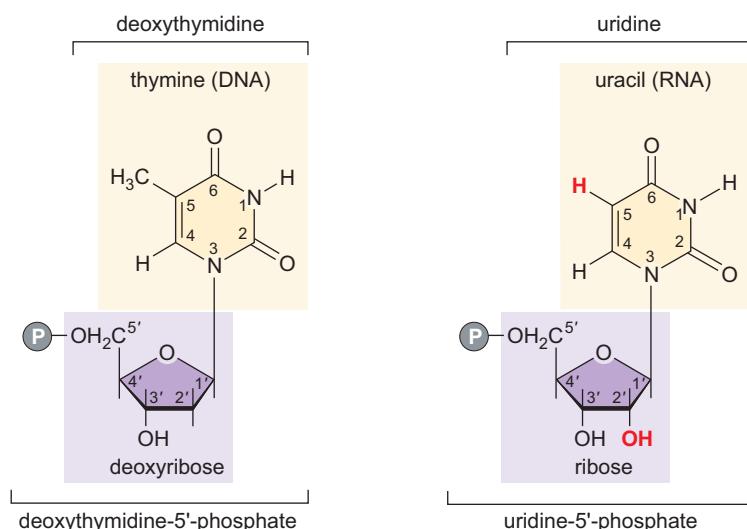
Therefore, at least for eukaryotic cells, a second information-containing molecule had to exist that obtains its genetic specificity from DNA. This molecule would then move to the cytoplasm to function as the template for protein synthesis. Attention from the start focused on the still functionally obscure second class of nucleic acids, RNA. Torbjörn Caspersson and Jean Brachet had found RNA to reside largely in the cytoplasm; and it was easy to imagine single DNA strands, when not serving as templates for complementary DNA strands, acting as templates for complementary RNA chains.

## RNA Is Chemically Very Similar to DNA

Mere inspection of RNA structure shows how it can be exactly synthesized on a DNA template. Chemically, it is very similar to DNA. It, too, is a long, unbranched molecule containing four types of nucleotides linked together by 3'-5' phosphodiester bonds (Fig. 2-11). Two differences in its chemical



**FIGURE 2-11** A portion of a polyribonucleotide (RNA) chain. Elements in red are distinct from DNA.

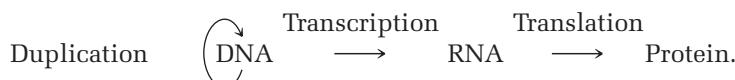


**FIGURE 2-12** Distinctions between the nucleotides of RNA and DNA. A nucleotide of DNA is shown next to a nucleotide of RNA. All RNA nucleotides have the sugar ribose (instead of deoxyribose for DNA), which has a hydroxyl group on the 2' carbon (shown in red). In addition, RNA has the pyrimidine base uracil instead of thymine. Uracil has a hydrogen at the 5 position of the pyrimidine ring (shown in red) rather than the methyl group found in that position for thymine. The three other bases that occur in DNA and RNA are identical.

groups distinguish RNA from DNA. The first is a minor modification of the sugar component (Fig. 2-12). The sugar of DNA is deoxyribose, whereas RNA contains ribose, identical to deoxyribose except for the presence of an additional OH (hydroxyl) group on the 2' carbon. The second difference is that RNA contains no thymine but instead contains the closely related pyrimidine uracil. Despite these differences, however, polyribonucleotides have the potential for forming complementary helices of the DNA type. Neither the additional hydroxyl group nor the absence of the methyl group found in thymine but not in uridine affects RNA's ability to form double-helical structures held together by base pairing. Unlike DNA, however, RNA is typically found in the cell as a single-stranded molecule. If double-stranded RNA helices are formed, they most often are composed of two parts of the same single-stranded RNA molecule.

## THE CENTRAL DOGMA

By the fall of 1953, the working hypothesis was adopted that chromosomal DNA functions as the template for RNA molecules, which subsequently move to the cytoplasm, where they determine the arrangement of amino acids within proteins. In 1956 Francis Crick referred to this pathway for the flow of genetic information as the **central dogma**:



Here the arrows indicate the directions proposed for the transfer of genetic information. The arrow encircling DNA signifies that DNA is the template for its self-replication. The arrow between DNA and RNA indicates that RNA synthesis (called **transcription**) is directed by a DNA template. Correspondingly, the synthesis of proteins (called **translation**) is directed by an RNA template. Most importantly, the last two arrows were presented as unidirectional; that is, RNA sequences are never determined by protein templates nor was DNA then imagined ever to be made on RNA templates. The idea that proteins never serve as templates for RNA has stood the test of time. However, as we will see in Chapter 12, RNA chains sometimes do

act as templates for DNA chains of complementary sequence. Such reversals of the normal flow of information are very rare events compared with the enormous number of RNA molecules made on DNA templates. Thus, the central dogma as originally proclaimed more than 50 years ago still remains essentially valid.

### The Adaptor Hypothesis of Crick

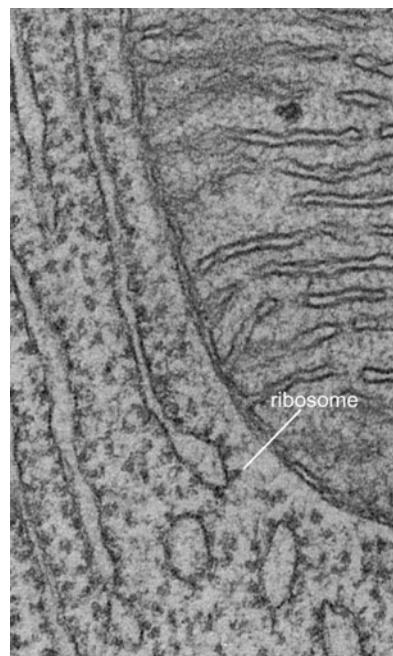
At first it seemed simplest to believe that the RNA templates for protein synthesis were folded up to create cavities on their outer surfaces specific for the 20 different amino acids. The cavities would be so shaped that only one given amino acid would fit, and in this way RNA would provide the information to order amino acids during protein synthesis. By 1955, however, Crick became disenchanted with this conventional wisdom, arguing that it would never work. In the first place, the specific chemical groups on the four bases of RNA (A, U, G, and C) should mostly interact with water-soluble groups. Yet, the specific side groups of many amino acids (e.g., leucine, valine, and phenylalanine) strongly prefer interactions with water-insoluble (hydrophobic) groups. In the second place, even if somehow RNA could be folded so as to display some hydrophobic surfaces, it seemed at the time unlikely that an RNA template would be used to discriminate accurately between chemically very similar amino acids like glycine and alanine or valine and isoleucine, both pairs differing only by the presence of single methyl ( $\text{CH}_3$ ) groups. Crick thus proposed that prior to incorporation into proteins, amino acids are first attached to specific adaptor molecules, which in turn possess unique surfaces that can bind specifically to bases on the RNA templates.

### Discovery of Transfer RNA

The discovery of how proteins are synthesized required the development of cell-free extracts capable of making proteins from amino acid precursors as directed by added RNA molecules. These were first effectively developed beginning in 1953 by Paul C. Zamecnik and his collaborators. Key to their success were the recently available radioactively tagged amino acids, which they used to mark the trace amounts of newly made proteins, as well as high-quality, easy-to-use, preparative ultracentrifuges for fractionation of their cellular extracts. Early on, the cellular site of protein synthesis was pinpointed to be the ribosomes, small RNA-containing particles in the cytoplasm of all cells engaged in protein synthesis (Fig. 2-13).

Several years later, Zamecnik, by then collaborating with Mahlon B. Hoagland, went on to make the seminal discovery that prior to their incorporation into proteins, amino acids are first attached to what we now call **transfer RNA (tRNA)** molecules. Transfer RNA accounts for some 10% of all cellular RNA (Fig. 2-14).

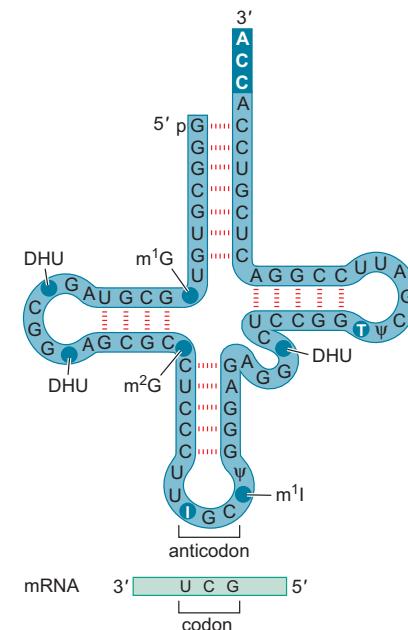
To nearly everyone except Crick, this discovery was totally unexpected. He had, of course, previously speculated that his proposed “adaptors” might be short RNA chains, because their bases would be able to base-pair and “read” the appropriate groups on the RNA molecules that served as the templates for protein synthesis. As we shall relate later in greater detail (Chapter 15), the transfer RNA molecules of Zamecnik and Hoagland are in fact the adaptor molecules postulated by Crick. Each transfer RNA contains a sequence of adjacent bases (the anticodon) that bind specifically during protein synthesis to successive groups of bases (codons) along the RNA template.



**FIGURE 2-13** Electron micrograph of ribosomes attached to the endoplasmic reticulum. This electron micrograph (105,000x) shows a portion of a pancreatic cell. The upper right portion shows a portion of the mitochondrion and the lower left shows a large number of ribosomes (small circles of electron density) attached to the endoplasmic reticulum. Some ribosomes exist free in the cytoplasm; others are attached to the membranous endoplasmic reticulum. (Courtesy of K.R. Porter.)

## The Paradox of the Nonspecific-Appearing Ribosomes

About 85% of cellular RNA is found in ribosomes, and because its absolute amount is greatly increased in cells engaged in large-scale protein synthesis (e.g., pancreas cells and rapidly growing bacteria), **ribosomal RNA (rRNA)** was initially thought to be the template for ordering amino acids. But once the ribosomes of *Escherichia coli* were carefully analyzed, several disquieting features emerged. First, all *E. coli* ribosomes, as well as those from all other organisms, are composed of two unequally sized subunits, each containing RNA, that either stick together or fall apart in a reversible manner, depending on the surrounding ion concentration. Second, all the rRNA chains within the small subunits are of similar chain lengths (about 1500 bases in *E. coli*), as are the rRNA chains of the large subunits (about 3000 bases). Third, the base composition of both the small and large rRNA chains is approximately the same (high in G and C) in all known bacteria, plants, and animals, despite wide variations in the AT/GC ratios of their respective DNA. This was not to be expected if the rRNA chains were in fact a large collection of different RNA templates derived from a large number of different genes. Thus, neither the small nor large class of rRNA had the feel of template RNA.



## Discovery of Messenger RNA (mRNA)

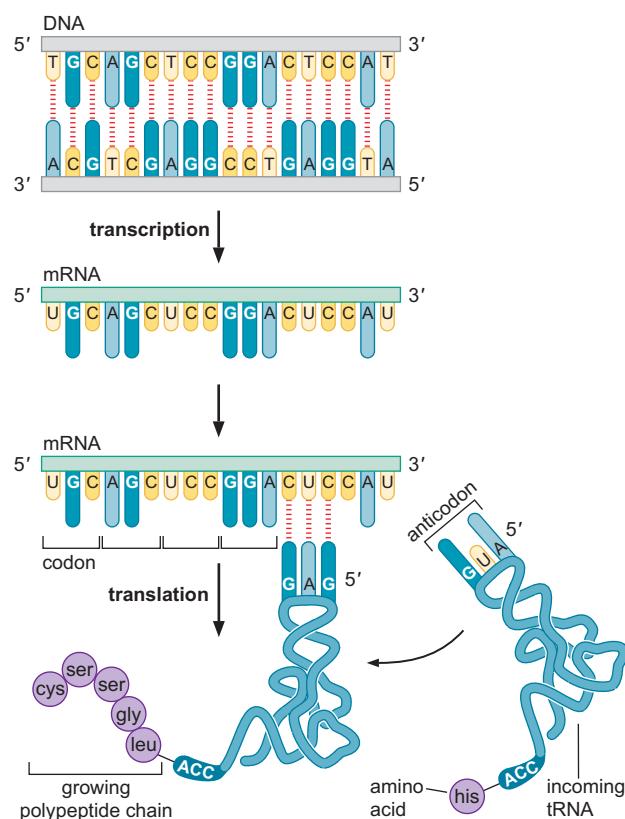
Cells infected with phage T4 provided the ideal system to find the true template. Following infection by this virus, cells stop synthesizing *E. coli* RNA; the only RNA synthesized is transcribed off the T4 DNA. Most strikingly, not only does T4 RNA have a base composition very similar to T4 DNA, but it does not bind to the ribosomal proteins that normally associate with rRNA to form ribosomes. Instead, after first attaching to previously existing ribosomes, T4 RNA moves across their surface to bring its bases into positions where they can bind to the appropriate tRNA–amino acid precursors for protein synthesis (Fig. 2-15). In so acting, T4 RNA orders the amino acids and is thus the long-sought-for RNA template for protein synthesis. Because it carries the information from DNA to the ribosomal sites of protein synthesis, it is called **messenger RNA (mRNA)**. The observation of T4 RNA binding to *E. coli* ribosomes, first made in the spring of 1960, was soon followed with evidence for a separate messenger class of RNA within uninfected *E. coli* cells, thereby definitively ruling out a template role for any rRNA. Instead, in ways that we discuss more extensively in Chapter 15, the rRNA components of ribosomes, together with some 50 different ribosomal proteins that bind to them, serve as the factories for protein synthesis, functioning to bring the tRNA–amino acid precursors into positions where they can read off the information provided by the mRNA templates.

Only a few percent of total cellular RNA is mRNA. This RNA shows the expected large variations in length and nucleotide composition required to encode the many different proteins found in a given cell. Hence, it is easy to understand why mRNA was first overlooked. Because only a small segment of mRNA is attached at a given moment to a ribosome, a single mRNA molecule can simultaneously be read by several ribosomes. Most ribosomes are found as parts of **polyribosomes** (groups of ribosomes translating the same mRNA), which can include more than 50 members (Fig. 2-16).

**FIGURE 2-14** Yeast alanine tRNA structure, as determined by Robert W. Holley and his associates. The anticodon in this tRNA recognizes the codon for alanine in the mRNA. Several modified nucleosides exist in the structure: ψ = pseudouridine, T = ribothymidine, DHU = 5,6-dihydrouridine, I = inosine, m<sup>1</sup>G = 1-methylguanosine, m<sup>1</sup>I = 1-methylinosine, and m<sup>2</sup>G = N,N-dimethylguanosine.

## Enzymatic Synthesis of RNA upon DNA Templates

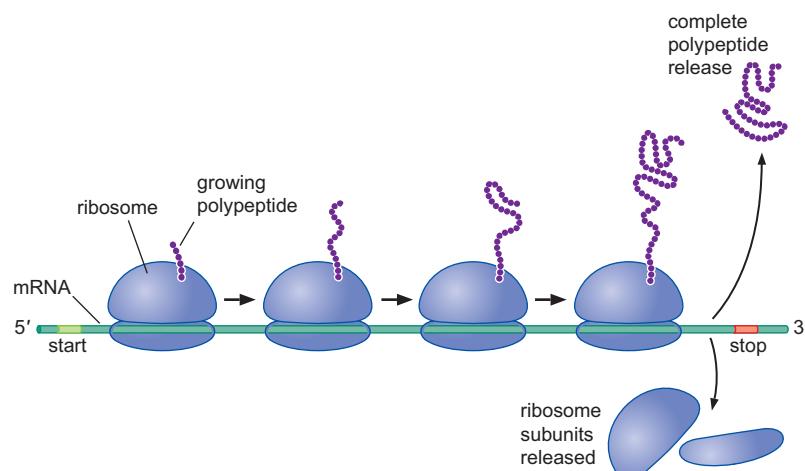
As mRNA was being discovered, the first of the enzymes that synthesize (or transcribe) RNA using DNA templates was being independently iso-



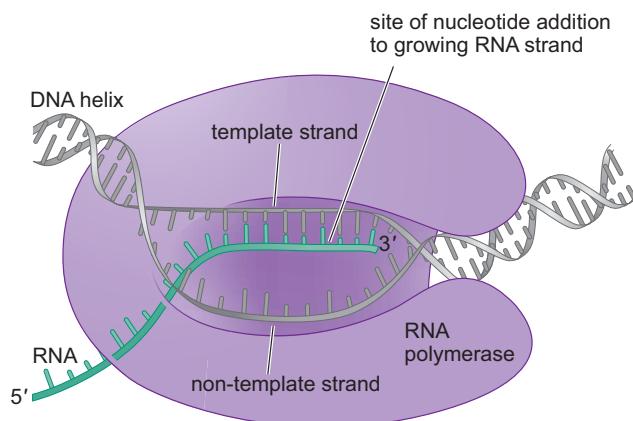
**FIGURE 2-15** Transcription and translation. The nucleotides of mRNA are assembled to form a complementary copy of one strand of DNA. Each group of three is a codon that is complementary to a group of three nucleotides in the anticodon region of a specific tRNA molecule. When base pairing occurs, an amino acid carried at the other end of the tRNA molecule is added to the growing protein chain.

lated in the labs of biochemists Jerard Hurwitz and Samuel B. Weiss. Called **RNA polymerases**, these enzymes function only in the presence of DNA, which serves as the template upon which single-stranded RNA chains are made, and use the nucleotides ATP, GTP, CTP, and UTP as precursors (Fig. 2-17). These enzymes make RNA using appropriate segments of chromosomal DNA as their templates. Direct evidence that DNA lines up the correct ribonucleotide precursors came from seeing how the RNA base composition varied with the addition of DNA molecules of different AT/GC ratios. In every enzymatic synthesis, the RNA AU/GC ratio was roughly similar to the DNA AT/GC ratio (Table 2-2).

During transcription, only one of the two strands of DNA is used as a template to make RNA. This makes sense, because the messages carried by the



**FIGURE 2-16** Diagram of a polyribosome. Each ribosome attaches at a start signal at the 5' end of an mRNA chain and synthesizes a polypeptide as it proceeds along the molecule. Several ribosomes may be attached to one mRNA molecule at one time; the entire assembly is called a polyribosome.



**FIGURE 2-17** Enzymatic synthesis of RNA upon a DNA template, catalyzed by RNA polymerase.

two strands, being complementary but not identical, are expected to code for completely different polypeptides. The synthesis of RNA always proceeds in a fixed direction, beginning at the 5' end and concluding with the 3'-end nucleotide (see Fig. 2-17).

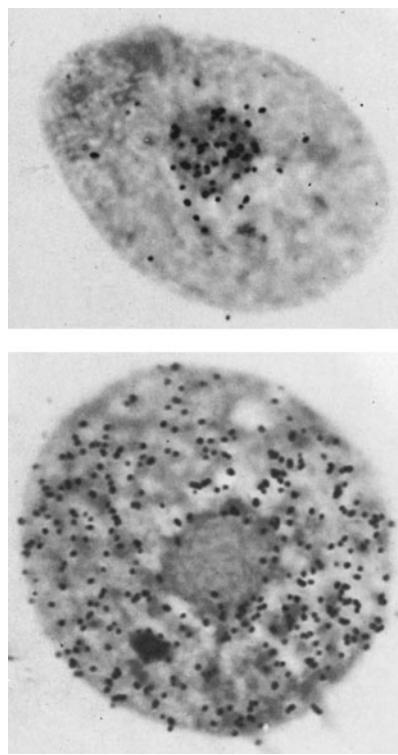
By this time, there was firm evidence for the postulated movement of RNA from the DNA-containing nucleus to the ribosome-containing cytoplasm of eukaryotic cells. By briefly exposing cells to radioactively labeled precursors, then adding a large excess of unlabeled ribonucleotides (a “pulse chase” experiment), mRNA synthesized during a short time window was labeled. These studies showed that mRNA is synthesized in the nucleus. Within an hour, most of this RNA had left the nucleus and was observed in the cytoplasm (Fig. 2-18).

### Establishing the Genetic Code

Given the existence of 20 amino acids but only four bases, groups of several nucleotides must somehow specify a given amino acid. Groups of two, however, would specify only 16 ( $4 \times 4$ ) amino acids. So from 1954, the start of serious thinking about what the genetic code might be like, most attention was given to how triplets (groups of three) might work, even though they obviously would provide more permutations ( $4 \times 4 \times 4$ ) than needed if each amino acid was specified by only a single triplet. The assumption of colinearity was then very important. It held that successive groups of nucleotides along a DNA chain code for successive amino acids along a given polypeptide chain. An elegant mutational analysis on bacterial proteins, carried out in the early 1960s by Charles Yanofsky and Sydney Brenner,

**TABLE 2-2** Comparison of the Base Composition of Enzymatically Synthesized RNAs with the Base Composition of Their Double-Helical DNA Templates

Source of DNA Template	Composition of the RNA Bases				$\frac{A+U}{G+C}$	$\frac{A+T}{G+C}$
	Adenine	Uracil	Guanine	Cytosine		
T2	0.31	0.34	0.18	0.17	1.86	1.84
Calf thymus	0.31	0.29	0.19	0.21	1.50	1.35
<i>Escherichia coli</i>	0.24	0.24	0.26	0.26	0.92	0.97
<i>Micrococcus lysodeikticus</i> (a bacterium)	0.17	0.16	0.33	0.34	0.49	0.39



**FIGURE 2-18** Demonstration that RNA is synthesized in the nucleus and moves to the cytoplasm. (Top) Autoradiograph of a cell (*Tetrahymena*) exposed to radioactive cytidine for 15 min. Superimposed on a photograph of a thin section of the cell is a photograph of an exposed silver emulsion. Each dark spot represents the origin of an electron emitted from a  $^3\text{H}$  (tritium) atom that has been incorporated into RNA. Almost all the newly made RNA is found within the nucleus. (Bottom) Autoradiograph of a similar cell exposed to radioactive cytidine for 12 min and then allowed to grow for 88 min in the presence of non-radioactive cytidine. Practically all the label incorporated into RNA in the first 12 min has left the nucleus and moved into the cytoplasm. (Courtesy of D.M. Prescott, University of Colorado Medical School; reproduced, with permission, from Prescott D.M. 1964. *Progr. Nucleic Acid Res. Mol. Biol.* 3: 35. © Elsevier.)

**TABLE 2-3** The Genetic Code

		second position								
		U	C	A	G					
first position	U	UUU UUC UUA UUG	Phe Leu	UCU UCC UCA UCG	Ser	UAU UAC UAA UAG	Tyr stop stop	UGU UGC UGA UGG	Cys stop Trp	U C A G
	C	CUU CUC CUA CUG	Leu	CCU CCC CCA CCG	Pro	CAU CAC CAA CAG	His Gln	CGU CGC CGA CGG	Arg	U C A G
	A	AUU AUC AUA AUG	Ile Met	ACU ACC ACA ACG	Thr	AAU AAC AAA AAG	Asn Lys	AGU AGC AGA AGG	Ser Arg	U C A G
	G	GUU GUC GUA GUG	Val	GCU GCC GCA GCG	Ala	GAU GAC GAA GAG	Asp Glu	GGU GGC GGA GGG	Gly	U C A G

showed that colinearity does in fact exist. Equally important were the genetic analyses by Brenner and Crick, which in 1961 first established that groups of three nucleotides are used to specify individual amino acids.

But which specific groups of three bases (codons) determine which specific amino acids could only be learned by biochemical analysis. The major breakthrough came when Marshall Nirenberg and Heinrich Matthaei, then working together, observed in 1961 that the addition of the synthetic polynucleotide poly U (UUUUU . . .) to a cell-free system capable of making proteins leads to the synthesis of polypeptide chains containing only the amino acid phenylalanine. The nucleotide groups UUU thus must specify phenylalanine. Use of increasingly more complex polynucleotides as synthetic messenger RNAs rapidly led to the identification of more and more codons. Particularly important in completing the code was the use of polynucleotides like AGUAGU, put together by organic chemist Har Gobind Khorana. These further defined polynucleotides were critical to test more specific sets of codons. Completion of the code in 1966 revealed that 61 out of the 64 possible permuted groups corresponded to amino acids, with most amino acids being encoded by more than one nucleotide triplet (Table 2-3).

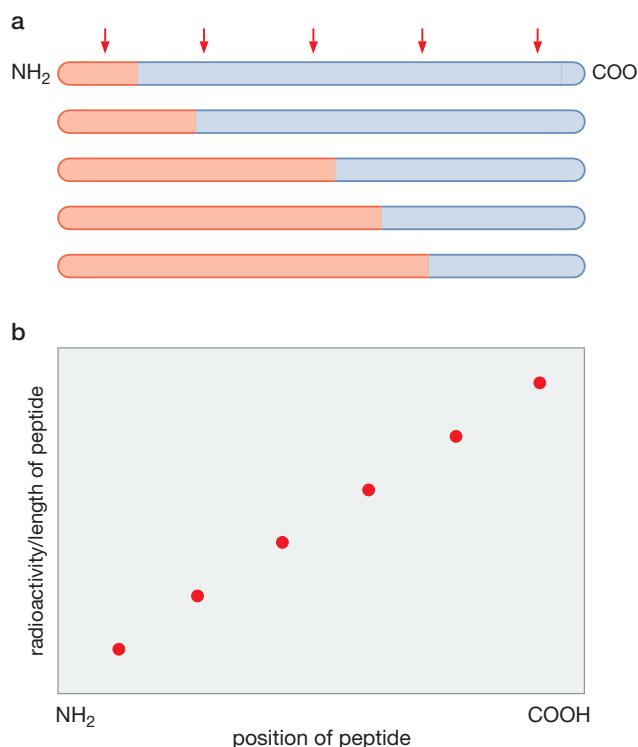
## ESTABLISHING THE DIRECTION OF PROTEIN SYNTHESIS

The nature of the genetic code, once determined, led to further questions about how a polynucleotide chain directs the synthesis of a polypeptide. As we have seen here and shall discuss in detail in Chapter 9, polynucleotide chains (both DNA and RNA) are synthesized by adding to the 3' end of the growing strand (growth in the 5' → 3' direction). But what about the growing

polypeptide chain? Is it assembled in an amino-terminal to carboxy-terminal direction, or the opposite?

This question was answered using a cell-free system for protein synthesis similar to the one used to identify the amino acid codons. Instead of providing synthetic mRNAs, however, the investigators provided  $\beta$ -globin mRNA to direct protein synthesis. A few minutes after initiation of protein synthesis, the cell-free system was treated with a radioactive amino acid for a few seconds (less than the time required to synthesize a complete globin chain) after which protein synthesis was immediately stopped. A brief radioactive labeling regime of this kind is known as **pulse-labeling**. Next, the  $\beta$ -globin chains that had *completed* their growth during the period of the pulse-labeling were separated from incomplete chains by gel electrophoresis (Chapter 7). Thus, all proteins analyzed would have completed their synthesis in the presence of radiolabeled precursors. The full-length polypeptides were then treated with an enzyme, the protease trypsin, that cleaves proteins at particular sites in the polypeptide chain, thereby generating a series of peptide fragments. In the final step of the experiment, the amount of radioactivity that had been incorporated into each peptide fragment was measured (Fig. 2-19).

Because in this experiment all proteins finished their synthesis in the presence of radioactive precursors, the peptides last to be synthesized will have the highest density of radiolabeled precursors (Fig. 2-19a). Conversely, peptides with the least amount of radioactive amino acid (normalized to the size of the peptide) would be derived from regions of the  $\beta$ -globin protein that were the first to be synthesized. The investigators observed that radioactive labeling was lowest for peptides from the amino-terminal region of globin and greatest for peptides from the carboxy-terminal region (Fig. 2-19b). This led to the conclusion that the direction of protein synthesis is from the amino terminus to the carboxyl terminus. In other words, new amino acids are added to the carboxyl terminus of the growing polypeptide chain.



**FIGURE 2-19** Incorporation of radioactively labeled amino acids into a growing polypeptide chain. (a) Distribution of radioactivity (shown in blue) among completed chains after a short period of labeling. The sites of trypsin cleavage of the  $\beta$ -globin protein are indicated by the red arrows. (b) Incorporation of label normalized to the length of each peptide is plotted as a function of position of the peptide within the completed chain.

### Start and Stop Signals Are Also Encoded within DNA

Initially, it was guessed that translation of an mRNA molecule would commence at one end and finish when the entire mRNA message had been read into amino acids. But, in fact, translation both starts and stops at internal positions. Thus, signals must be present within DNA (and its mRNA products) to initiate and terminate translation. The stop signals were the first to be worked out. Three separate codons (UAA, UAG, and UGA), first known as **nonsense codons**, do not direct the addition of a particular amino acid. Instead, these codons serve as translational stop signals (sometimes called stop codons). The way translational start signals are encoded is more complicated. The amino acid methionine initiates all polypeptide chains, but the triplet (AUG) that codes for these initiating methionines also codes for methionine residues that are found at internal protein positions. In prokaryotes, the AUG codons that start new polypeptide chains are preceded by specific purine-rich blocks of nucleotides that serve to attach mRNA to ribosomes (see Chapter 15). In eukaryotes, the position of the AUG relative to the beginning of the mRNA is the critical determinant, with the first AUG always being selected as the start site of translation.

## THE ERA OF GENOMICS

---

With the elucidation of the central dogma, it became clear by the mid-1960s how the genetic blueprint contained in the nucleotide sequence could determine phenotype. This meant that profound insights into the nature of living things and their evolution would be revealed from DNA sequences. In recent years the advent of rapid, automated DNA sequencing methods has led to the determination of complete genome sequences for hundreds of organisms. Even the human genome, a single copy of which is composed of more than 3 billion base pairs, has been elucidated and shown to contain more than 20,000 genes. The sequencing of the genomes of many organisms has made the comparative analysis of genome sequences very useful. By comparing the predicted amino acid sequences encoded by similar genes from different organisms one can frequently identify important regions of a protein. For example, the amino acids in DNA polymerases that are critical for binding the incoming nucleotide or that directly catalyze nucleotide addition are well conserved in the DNA polymerases from many different organisms. Similarly, amino acids that are important to DNA polymerase function in bacteria but not in eukaryotic cells will be conserved only in the amino acid sequences predicted by the genome sequences from bacteria.

Comparison of different genomes can also offer insights into DNA sequences that do not encode proteins. The identification of sequences that direct gene expression, DNA replication, chromosome segregation, and recombination can all be facilitated by comparing genome sequences. Because these regulatory sequences tend to diverge more rapidly, these comparisons are often made between closely related species (such as between different bacteria or between humans and other primates). The value of comparisons between closely related species has led to efforts to sequence the genomes of organisms closely related to well-studied model organisms such as the fruit fly *Drosophila melanogaster*, the yeast *Saccharomyces cerevisiae*, or multiple primates.

Comparative genomics between different individuals of the same organism has the potential to identify mutations that lead to disease. For example, recent efforts have developed methods to rapidly compare the sequences of

a small subset of the human genome among many different individuals in an effort to identify disease genes. Finally, it is possible to envision a day when comparative genome analysis will reveal basic insights into the origins of complex behavior in humans, such as the acquisition of language, as well as the mechanisms underlying the evolutionary diversification of animal body plans.

The purpose of the forthcoming chapters is to provide a firm foundation for understanding how DNA functions as the template for biological complexity. The chapters in Part 2 review the basic chemistry and biochemical structures that are relevant to the main themes of this book. The final chapter in Part 2 presents various laboratory techniques commonly used to investigate biological structures and problems. The initial chapters in Part 3, Maintenance of the Genome, describe the structure of the genetic material and its faithful duplication. The following chapters present the processes that provide a means for generating genetic variation as well as the repair of damaged parts of the genome. Part 4, Expression of the Genome, shows how the genetic instructions contained in DNA are converted into proteins. Part 5, Regulation, describes strategies for differential gene activity that are used to generate complexity within organisms (e.g., embryogenesis) and diversity among organisms (e.g., evolution). The last chapter in Part 5 on systems biology presents interdisciplinary approaches for investigating more complex levels of biological organization. And an appendix describes several model organisms that have served as important experimental systems to reveal general biological patterns across many different organisms.

## SUMMARY

---

The discovery that DNA is the genetic material can be traced to experiments performed by Griffith, who showed that nonvirulent strains of bacteria could be genetically transformed with a substance derived from a heat-killed pathogenic strain. Avery, McCarty, and MacLeod subsequently demonstrated that the transforming substance was DNA. Further evidence that DNA is the genetic material was obtained by Hershey and Chase in experiments with radiolabeled bacteriophage. Building on Chargaff's rules and Franklin and Wilkins' X-ray diffraction studies, Watson and Crick proposed a double-helical structure of DNA. In this model, two polynucleotide chains are twisted around each other to form a regular double helix. The two chains within the double helix are held together by hydrogen bonds between pairs of bases. Adenine is always joined to thymine, and guanine is always bonded to cytosine. The existence of the base pairs means that the sequence of nucleotides along the two chains are not identical, but complementary. The finding of this relationship suggested a mechanism for the replication of DNA in which each strand serves as a template for its complement. Proof for this hypothesis came from (a) the observation of Meselson and Stahl that the two strands of each double helix separate during each round of DNA replication, and (b) Kornberg's discovery of an enzyme that uses

single-stranded DNA as a template for the synthesis of a complementary strand.

As we have seen, according to the "central dogma" information flows from DNA to RNA to protein. This transformation is achieved in two steps. First, DNA is transcribed into an RNA intermediate (messenger RNA), and second, the mRNA is translated into protein. Translation of the mRNA requires RNA adaptor molecules called tRNAs. The key characteristic of the genetic code is that each triplet codon is recognized by a tRNA, which is associated with a cognate amino acid. Out of 64 ( $4 \times 4 \times 4$ ) potential codons, 61 are used to specify the 20 amino acid building blocks of proteins, whereas 3 are used to provide chain-terminating signals. Knowledge of the genetic code allows us to predict protein-coding sequences from DNA sequences. The advent of rapid DNA sequencing methods has ushered in a new era of genomics, in which complete genome sequences are being determined for a wide variety of organisms, including humans. Comparing genome sequences offers a powerful method to identify critical regions of the genome that encode not only important elements of proteins but also regulatory regions that control the expression of genes and the duplication of the genome.

## BIBLIOGRAPHY

---

- Brenner S., Jacob F., and Meselson M. 1961. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature* **190**: 576–581.
- Brenner S., Stretton A.O.W., and Kaplan S. 1965. Genetic code: The nonsense triplets for chain termination and their suppression. *Nature* **206**: 994–998.
- Cairns J., Stent G.S., and Watson J.D., eds 1966. *Phage and the origins of molecular biology*. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
- Chargaff E. 1951. Structure and function of nucleic acids as cell constituents. *Fed Proc* **10**: 654–659.
- Cold Spring Harbor Symposia on Quantitative Biology*. 1966. Vol. 31: *The genetic code*. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
- Crick F.H.C. 1955. On degenerate template and the adaptor hypothesis. A note for the RNA Tie Club, unpublished. Mentioned in Crick's 1957 discussion, pp. 25–26, in The structure of nucleic acids and their role in protein synthesis. *Biochem Soc Symp* no. 14, Cambridge University Press, Cambridge, England.
- . 1958. On protein synthesis. *Symp. Soc. Exp. Biol.* **12**: 548–555.
- . 1963. The recent excitement in the coding problem. *Prog Nucleic Acid Res.* **1**: 164–217.
- . 1988. *What mad pursuit: A personal view of scientific discovery*. Basic Books, New York.
- Crick F.H.C. and Watson J.D. 1954. The complementary structure of deoxyribonucleic acid. *Proc. Roy. Soc. A* **223**: 80–96.
- Echols H. and Gross C.A., eds 2001. *Operators and promoters: The story of molecular biology and its creators*. University of California Press, Berkeley, California.
- Franklin R.E. and Gosling R.G. 1953. Molecular configuration in sodium thymonuclease. *Nature* **171**: 740–741.
- Hershey A.D. and Chase M. 1952. Independent function of viral protein and nucleic acid on growth of bacteriophage. *J. Gen. Physiol.* **36**: 39–56.
- Hoagland M.B., Stephenson M.L., Scott J.F., Hecht L.I., and Zamecnik P.C. 1958. A soluble ribonucleic acid intermediate in protein synthesis. *J. Biol. Chem.* **231**: 241–257.
- Holley R.W., Apgar J., Everett G.A., Madison J.T., Marquise M., Merrill S.H., Penswick J.R., and Zamir A. 1965. Structure of a ribonucleic acid. *Science* **147**: 1462–1465.
- Ingram V.M. 1957. Gene mutations in human hemoglobin: The chemical difference between normal and sickle cell hemoglobin. *Nature* **180**: 326–328.
- Jacob F. and Monod J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**: 318–356.
- Judson H.F. 1996. *The eighth day of creation*, expanded edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
- Kornberg A. 1960. Biological synthesis of deoxyribonucleic acid. *Science* **131**: 1503–1508.
- Kornberg A. and Baker T.A. 1992. *DNA replication*. W.H. Freeman, New York.
- McCarty M. 1985. *The transforming principle: Discovering that genes are made of DNA*. Norton, New York.
- Meselson M. and Stahl F.W. 1958. The replication of DNA in *Escherichia coli*. *Proc. Natl. Acad. Sci.* **44**: 671–682.
- Nirenberg M.W. and Matthaei J.H. 1961. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci.* **47**: 1588–1602.
- Olby R. 1975. *The path to the double helix*. University of Washington Press, Seattle.
- Portugal F.H. and Cohen J.S. 1980. *A century of DNA: A history of the discovery of the structure and function of the genetic substance*. MIT Press, Cambridge, Massachusetts.
- Sarabhai A.S., Stretton A.O.W., Brenner S., and Bolte A. 1964. Colinearity of the gene with the polypeptide chain. *Nature* **201**: 13–17.
- Stent G.S. and Calendar R. 1978. *Molecular genetics: An introductory narrative*, 2nd ed. Freeman, San Francisco.
- Volkin E. and Astrachan L. 1956. Phosphorus incorporation in *E. coli* ribonucleic acid after infection with bacteriophage T2. *Virology* **2**: 146–161.
- Watson J.D. 1963. Involvement of RNA in synthesis of proteins. *Science* **140**: 17–26.
- . 1968. *The double helix*. Atheneum, New York.
- . 1980. *The double helix: A Norton critical edition* (ed. G.S. Stent). Norton, New York.
- . 2000. *A passion for DNA: Genes, genomes and society*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
- . 2002. *Genes, girls, and Gamow: After the double helix*. Knopf, New York.
- Watson J.D. and Crick F.H.C. 1953a. Genetical implications of the structure of deoxyribonucleic acid. *Nature* **171**: 964–967.
- . 1953b. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**: 737–738.
- Wilkins M.H.F., Stokes A.R., and Wilson H.R. 1953. Molecular structure of deoxypentose nucleic acid. *Nature* **171**: 738–740.
- Yanofsky C., Carlton B.C., Guest J.R., Helinski D.R., and Henning U. 1964. On the colinearity of gene structure and protein structure. *Proc. Natl. Acad. Sci.* **51**: 266–272.

## QUESTIONS

MasteringBiology®

For instructor-assigned tutorials and problems, go to MasteringBiology.

For answers to even-numbered questions, see Appendix 2: Answers.

**Question 1.** Avery, MacLeod, and McCarty concluded that DNA contained the genetic information to transform nonpathogenic *R* (rough) *Streptococcus pneumoniae* to pathogenic *S* (smooth) *S. pneumoniae*. Explain the experimental logic behind treatment of the active purified fraction with deoxyribonuclease, ribonuclease, or proteolytic enzymes.

**Question 2.** In the 1952 Hershey–Chase experiment (Fig. 2-3), explain why the protein is labeled with  $^{35}\text{S}$  and the DNA is labeled with  $^{32}\text{P}$ . Is it possible to do the reverse labeling (DNA with  $^{35}\text{S}$  and protein with  $^{32}\text{P}$ )?

**Question 3.**

- Considering the Chargaff's data in Box 2-1, explain how the human data help refute that thymine pairs with cytosine in double-stranded DNA.

- B. If Chargaff only collected the data for *E. coli* K2, would you be able to confidently refute that thymine pairs with cytosine in double-stranded DNA? Explain why or why not.

**Question 4.** Describe the structural differences between a base, nucleoside, and nucleotide (that are applicable for any given base).

**Question 5.** Review the data in Table 2-1. Justify how the experimental data for *Aerobacter aerogenes* support the hypothesis that a DNA polymerase uses a template to direct the synthesis of new DNA with a specific sequence.

**Question 6.**

- A. Using the same experimental setup as in the original Meselson and Stahl experiment (see Figs. 2-9 and 2-10), predict the bands (heavy, light, and/or intermediate) that you would observe after one round of replication if DNA polymerase replicated the bacterial genome by the dispersive model of replication.
- B. Using the same experimental setup as in the original Meselson and Stahl experiment, predict the bands (heavy, light, and/or intermediate) that you would observe after one round of replication if DNA polymerase replicated the bacterial genome by the conservative model of replication.
- C. How many rounds of replication would the original Meselson and Stahl experiment include to distinguish between the three models of replication (dispersive, conservative, semiconservative)? Explain your answer.

**Question 7.** Review Box 2-2. Explain how Vernon Ingram used the S allele and sickle cell anemia to provide evidence that genes encoded proteins.

**Question 8.** Describe several general properties of RNA that provided clues that RNA linked the genetic information from DNA to the amino acid sequence in proteins.

**Question 9.** Provide the justifications that the rRNA does not dictate the sequence of amino acids in protein synthesis.

**Question 10.** How is polyribosome formation advantageous for expression of a specific protein?

**Question 11.** Using the genetic code given in Table 2-3, name the amino acids(s) produced from the template AGUAGU using the cell-free translation system.

**Question 12.** Write out the steps in the Central Dogma. List where each step occurs in the cell. For each arrow, name the process that the arrow represents and the primary enzyme (complex) responsible for completing that step. Name the step(s) in which mRNA, tRNA, and rRNA have a role and briefly describe the role of each in the step(s) you name.

**Question 13.**

- A. Review the experiment and data depicted in Figure 2-19. In terms of the experimental setup, explain why the values for radioactivity per length of peptide are not equal for each full-length peptide isolated.
- B. Predict how the data shown in the graph would change if the trypsin cleavage step did not take place.
- C. Predict how the data shown in the graph would change if proteins were translated in the carboxy-terminal to amino-terminal direction. Explain why you would see that change.

**Question 14.** Tissières and Hopkins studied the relationship between DNA and protein synthesis. In one experiment, they measured the incorporation of amino acids into proteins in the presence of the enzyme deoxyribonuclease (DNase). They incubated a crude *E. coli* extract with varying concentrations of DNase for 10 min before adding the necessary components for the protein synthesis reaction including <sup>14</sup>C-labeled alanine. The amount of radioactivity incorporated is represented as cpm (counts per minute) in the data summarized below.

DNase ( $\mu\text{g/ml}$ )	0	1	5	10	20	50
cpm	813	334	372	364	386	426
% Inhibition		59	54	55	53	48

- A. What effect does the addition of DNase have on protein synthesis?
- B. From what you know about the central dogma, explain why the addition of DNase causes the effect on amino acid incorporation observed.

Data adapted from Tissières and Hopkins (1961. *Proc. Natl. Acad. Sci.* **47**: 2015–2023).

**Question 15.** Audrey Stevens measured the incorporation of <sup>32</sup>P-labeled ADP or <sup>32</sup>P-labeled ATP into RNA. He added various nucleotides mixtures to an *E. coli* extract that catalyzed RNA synthesis. The added components per reaction are listed in the table below. The observed radioactivity incorporated into the RNA is listed in terms of cpm (counts per minute).

Reaction components added	cpm incorporated
ATP	790
ATP, UTP, CTP, GTP	3920
ADP	690
ADP, UDP, CDP, GDP	1800

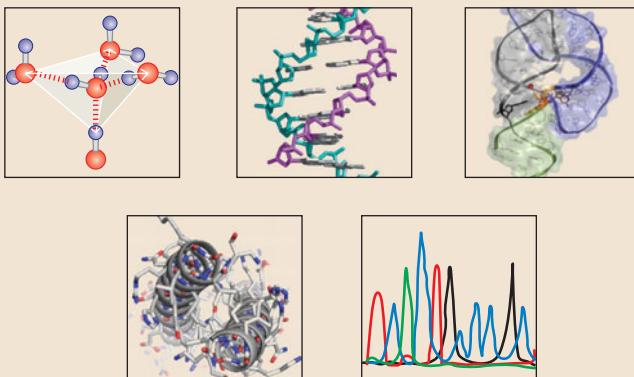
- A. Based on your knowledge from this chapter and the data given, what enzyme do you think is present in the *E. coli* extract?
- B. Why do you think the second reaction has the greatest cpm value compared to the other reactions?

Data adapted from Stevens (1960. *Biochem. Biophys. Res. Commun.* **3**: 92–96).

*This page intentionally left blank*

P A R T      2

# STRUCTURE AND STUDY OF MACROMOLECULES



## O U T L I N E

---

### CHAPTER 3

The Importance of Weak and Strong Chemical Bonds, 51



### CHAPTER 4

The Structure of DNA, 77



### CHAPTER 5

The Structure and Versatility of RNA, 107



### CHAPTER 6

The Structure of Proteins, 121



### CHAPTER 7

Techniques of Molecular Biology, 147

PART 2 IS DEDICATED TO THE STRUCTURE of macromolecules, the chemistry underlying those structures, and the methods by which those molecules are studied.

The basic chemistry presented in Chapter 3 focuses on the nature of chemical bonds—both weak and strong—and describes their roles in biology. Our discussion opens with weak chemical interactions, namely hydrogen bonds, and van der Waals and hydrophobic interactions. These forces mediate most interactions between macromolecules—between proteins or between proteins and DNA, for example. These weak bonds are critical for the activity and regulation of the majority of cellular processes. Thus, enzymes bind their substrates using weak chemical interactions; and transcriptional regulators bind sites on DNA to switch genes on and off using the same class of bonds.

Individual weak interactions are very weak indeed and thus dissociate quickly after forming. This reversibility is important for their roles in biology. Inside cells, molecules must interact dynamically (reversibly) or the whole system would seize up. At the same time, certain interactions must, at least in the short term, be stable. To accommodate these apparently conflicting demands, multiple weak interactions tend to be used together.

We then turn to strong bonds—the bonds that hold together the components that make up each macromolecule. Thus, proteins are made up of amino acids linked in a specific order by strong bonds, and DNA is made up of similarly linked nucleotides. (The atoms that make up the amino acids and nucleotides are also joined together by strong bonds.)

Chapter 4 explores the structure of DNA in atomic detail, from the chemistry of its bases and backbone to the base-pairing interactions and other forces that hold the two strands together. Thus, we see how both strong and weak bonds are critical in bestowing upon this molecule its properties and thus defining its functions. DNA is often topologically constrained, and Chapter 4 considers the biological effects of such constraints, together with enzymes that alter topology.

Chapter 5 explores the structure of RNA. Despite its similarity to DNA, RNA has its own distinctive structural features and properties, including the remarkable capacity to catalyze several cellular processes, a theme we shall return to in later chapters of the book. RNA's ability to both encode information (like DNA) and act enzymatically (like many proteins) afforded it a fascinating role in the early evolution of life, a matter we return to in Chapter 17.

In Chapter 6, we see how the strong and weak bonds together also give proteins distinctive three-dimensional shapes (and thereby specific functions). Thus, just as weak bonds mediate interactions between macromolecules, so too they act between, for example, nonadjacent amino acids within a given protein. In so doing, they determine how the primary chain of amino acids folds into a three-dimensional shape.

We also consider, in Chapter 6, how proteins interact with each other and with other macromolecules, in particular DNA and RNA; as we will see again and again in this book, the interactions between proteins and nucleic acids lie at the heart of most processes and regulation. We also look at how the function of a given protein can be regulated. One way is by changing the shape of the protein, a mechanism called allosteric regulation. Thus, in one conformation, a given protein may perform a specific enzymatic function or bind a specific target molecule. In another conformation, however, it may lose that ability. Such a change in shape can be triggered by the binding of another protein or a small molecule such as a sugar. In other cases, an allosteric effect can be induced by a covalent modification. For example,

attaching one or more phosphate groups to a protein can trigger a change in the shape of that protein. Another way a protein can be controlled is by regulating when it is brought into contact with a target molecule. In this way a given protein can be recruited to work on different target proteins in response to different signals.

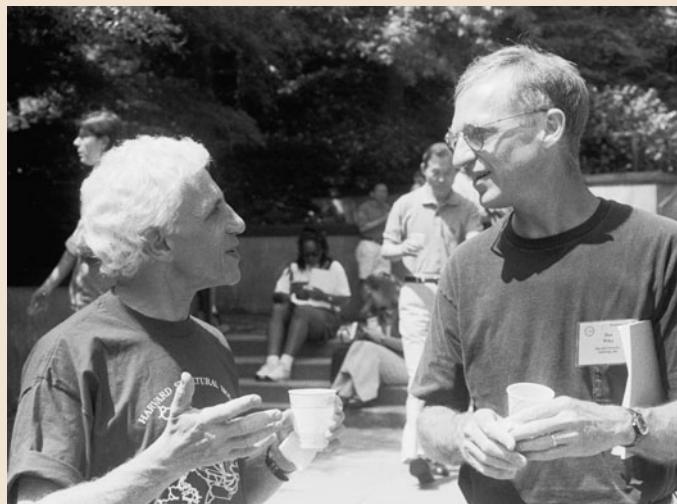
We end Part 2 with a chapter (Chapter 7) describing many of the fundamental techniques of molecular biology. These are the techniques that are widely used in studying nucleic acids and proteins, as encountered throughout this book. Additional methods and techniques—often more specialized for particular problems—are presented within individual chapters, but those collected in Chapter 7 are the core set used throughout molecular biology.

## PHOTOS FROM THE COLD SPRING HARBOR LABORATORY ARCHIVES

---



**Joan Steitz and Fritz Lipmann, 1969 Symposium on The Mechanism of Protein Synthesis.** Steitz's research focuses on the structure and function of RNA molecules, particularly those involved in RNA splicing (Chapter 14), and she was an author of the fourth edition of this book. Lipmann showed that the high-energy phosphate group in ATP is the source of energy that drives many biological processes (Chapter 3). For this he shared, with Hans Krebs, the 1953 Nobel Prize in Physiology or Medicine.



**Stephen Harrison and Don Wiley, 1999 Symposium on Signaling and Gene Expression in the Immune System.** For many years these two structural biologists shared a laboratory at Harvard, pursuing independent and sometimes overlapping projects. Both were interested in viral infection. Harrison's research group was the first to determine the atomic structure of an intact virus particle; Wiley was renowned for his work on influenza hemagglutinin and MHC complexes. Harrison, who wrote the new chapter on the structure of proteins for this edition (Chapter 6), also determined, in a collaboration with Mark Ptashne, the first structure of a sequence-specific protein:DNA complex.



**Werner Arber and Daniel Nathans, 1978 Symposium on DNA: Replication and Recombination.** These two shared, with Hamilton O. Smith, the 1978 Nobel Prize in Physiology or Medicine for the characterization of type II restriction enzymes and their application to the molecular analysis of DNA (Chapter 7). This was one of the key discoveries in the development of recombinant DNA technology in the early 1970s.



**Kary Mullis, 1986 Symposium on Molecular Biology of *Homo sapiens*.** Mullis (right) invented polymerase chain reaction (PCR), one of the central techniques of molecular biology (Chapter 7), for which he won the 1993 Nobel Prize in Chemistry. He is here pictured with Maxine Singer (center), best known as a writer and administrator who has written several books on genetics (often together with Paul Berg) and who received the National Medal of Science in 1992. On the left is Georgii Georgiev, the founding editor of the *Russian Journal of Developmental Biology*.



**Paul Berg, 1963 Symposium on Synthesis and Structure of Macromolecules.** Berg was a pioneer in the construction of recombinant DNA molecules in vitro, work reflected in his share of the 1980 Nobel Prize in Chemistry.



**Hamilton O. Smith, 1984 Symposium on Recombination at the DNA Level.** Smith shared, with Werner Arber and Daniel Nathans, the 1978 Nobel Prize in Physiology or Medicine for the discovery and characterization of type II restriction enzymes (Chapter 7). Later, Smith worked with Craig Venter and was involved in projects ranging from the sequencing of the first bacterial genome (*Haemophilus influenzae*) to the creation of synthetic genomes (Chapter 17).



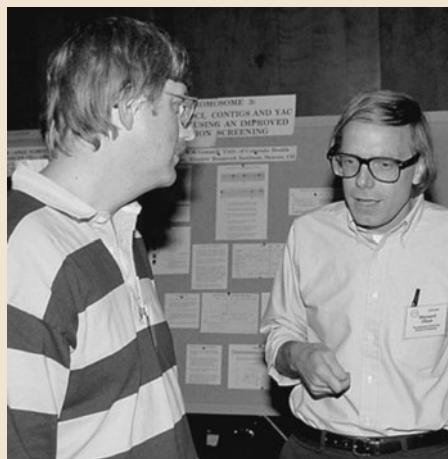
**Leroy Hood and J. Craig Venter, 1990 Genome Mapping and Sequencing Meeting.** Hood invented automated sequencing, building his first machine in the mid-1980s. It was Venter who later took first and greatest advantage of automated sequencing: by marrying the raw sequencing powers such machines offered with a shotgun strategy, he greatly accelerated the sequencing of whole genomes including that of the human (Chapter 7).



**Albert Keston, Sidney Udenfriend, and Frederick Sanger, 1949 Symposium on Amino Acids and Proteins.** Keston—*inventor of the test tape for detecting glucose*—and Udenfriend —*developer of screens for, and tests of, antimalarial drugs*—are here shown with Sanger, the only person to win two Nobel Prizes in Chemistry. The first, in 1958, was for developing a method to determine the amino acid sequence of a protein; the second, 22 years later, was for developing the primary method for sequencing DNA (Chapter 7). Beyond the obvious technological achievement, determining that a protein had a defined sequence revealed for the first time that it likely had a defined structure as well.



**Eric S. Lander (speaking), 1986 Symposium on Molecular Biology of *Homo sapiens*.** Lander was to become a leading figure in the public Human Genome Project and first author on the paper it produced reporting that sequence in 2001 (Chapter 7). As in the photo below of Gilbert and Botstein, Lander is here giving forth his views at the 1986 debate on whether it was worth trying to sequence the human genome. Beside him, David Page, whose work has focused on the structure, function, and evolution of the Y chromosome, appears thoughtful; in the foreground, Nancy Hopkins (a developmental biologist and an author on the fourth edition of this book)—and, in the background, James Watson—seem more amused.



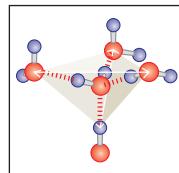
**Francis S. Collins and Maynard V. Olson, 1992 Genome Mapping and Sequencing Meeting.** Collins was one of the early “gene hunters,” finding first the much sought after cystic fibrosis gene in 1989. In 1993 he took over from James Watson as Director of the National Center for Human Genome Research and is today head of the National Institutes of Health. He is here seen listening to Olson in front of a poster about yeast artificial chromosomes (YACs), the vectors Olson had created a few years earlier and which allowed a 10-fold jump in the size of DNA fragments that could at the time be cloned (Chapter 7).



**Walter Gilbert and David Botstein, 1986 Symposium on Molecular Biology of *Homo sapiens*.** Gilbert, who invented a chemical method for sequencing DNA, is shown here with Botstein during the historic debate about whether it was feasible and sensible to attempt to sequence the human genome. Botstein, after working with phage for many years, contributed much to the development of the yeast *Saccharomyces cerevisiae* as a model eukaryote for molecular biologists; he was also an early figure in the emerging field of genomics (Chapter 7 and Appendix 1).

*This page intentionally left blank*

CHAPTER 3



# The Importance of Weak and Strong Chemical Bonds

## OUTLINE

THE MACROMOLECULES THAT PREOCCUPY us throughout this book—and those of most concern to molecular biologists—are proteins and nucleic acids. These are made of amino acids and nucleotides, respectively, and in both cases the constituents are joined by covalent bonds to make polypeptide (protein) and polynucleotide (nucleic acid) chains. Covalent bonds are strong, stable bonds and essentially never break spontaneously within biological systems. But weaker bonds also exist, and indeed are vital for life, partly because they can form and break under the physiological conditions present within cells. Weak bonds mediate the interactions between enzymes and their substrates, and between macromolecules—most strikingly, as we shall see in subsequent chapters, between proteins and DNA or RNA, or between proteins and other proteins. But equally important, weak bonds also mediate interactions between different parts of individual macromolecules, determining the shape of those molecules and hence their biological function. Thus, although a protein is a linear chain of covalently linked amino acids, its shape and function are determined by the stable three-dimensional (3D) structure it adopts. That shape is determined by a large collection of individually weak interactions that form between amino acids, which do not need to be adjacent in the primary sequence. Likewise, it is the weak, noncovalent bonds that hold the two chains of a DNA double helix together.

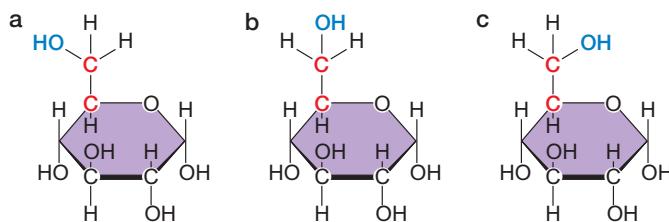
In the first part of the chapter, we consider the nature of chemical bonds and the concept of free energy—that is, energy that is released (or changed) during the formation of a chemical bond. We then concentrate on the weak bonds so vital to the proper function of all biological macromolecules. In particular, we describe what it is that gives weak bonds their weak character. In the last part of the chapter, we discuss high-energy bonds and consider the thermodynamic aspects of the peptide bond and the phosphodiester bond.

## CHARACTERISTICS OF CHEMICAL BONDS

A **chemical bond** is an attractive force that holds atoms together. Aggregates of finite size are called **molecules**. Originally, it was thought that only covalent

- Characteristics of Chemical Bonds, 51
    - The Concept of Free Energy, 54
  - Weak Bonds in Biological Systems, 55
    - High-Energy Bonds, 63
    - Molecules That Donate Energy Are Thermodynamically Unstable, 63
  - Enzymes Lower Activation Energies in Biochemical Reactions, 65
    - Free Energy in Biomolecules, 66
  - High-Energy Bonds in Biosynthetic Reactions, 67
    - Activation of Precursors in Group Transfer Reactions, 69
- Visit Web Content for Structural Tutorials and Interactive Animations

**FIGURE 3-1** Rotation about the C<sub>5</sub>—C<sub>6</sub> bond in glucose. This carbon–carbon bond is a single bond, and thus any of the three configurations, a, b, or c, may occur.



bonds hold atoms together in molecules; now, weaker attractive forces are known to be important in holding together many macromolecules. For example, the four polypeptide chains of hemoglobin are held together by the combined action of several weak bonds. Therefore, it is now customary also to call weak positive interactions “chemical bonds,” even though they are not strong enough, when present singly, to bind two atoms together effectively.

Chemical bonds are characterized in several ways. An obvious characteristic of a bond is its strength. Strong bonds almost never fall apart at physiological temperatures. This is why atoms united by covalent bonds always belong to the same molecule. Weak bonds are easily broken, and when they exist singly, they exist fleetingly. Only when present in ordered groups do weak bonds last a long time. The strength of a bond is correlated with its length, so that two atoms connected by a strong bond are always closer together than the same two atoms held together by a weak bond. For example, two hydrogen atoms bound covalently to form a hydrogen molecule (H:H) are 0.74 Å apart, whereas the same two atoms held together by van der Waals forces are 1.2 Å apart.

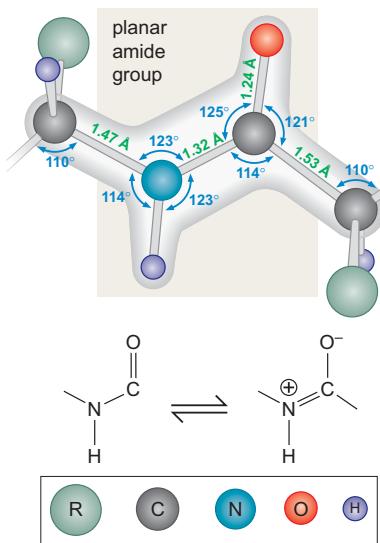
Another important characteristic is the maximum number of bonds that a given atom can make. The number of covalent bonds that an atom can form is called its **valence**. Oxygen, for example, has a valence of 2: It can never form more than two covalent bonds. But there is more variability in the case of van der Waals bonds, in which the limiting factor is purely steric. The number of possible bonds is limited only by the number of atoms that can touch each other simultaneously. The formation of hydrogen bonds is subject to more restrictions. A covalently bonded hydrogen atom usually participates in only one hydrogen bond, whereas an oxygen atom seldom participates in more than two hydrogen bonds.

The angle between two bonds originating from a single atom is called the **bond angle**. The angle between two specific covalent bonds is always approximately the same. For example, when a carbon atom has four single covalent bonds, they are directed tetrahedrally (bond angle = 109°). In contrast, the angles between weak bonds are much more variable.

Bonds differ also in the **freedom of rotation** they allow. Single covalent bonds permit free rotation of bound atoms (Fig. 3-1), whereas double and triple bonds are quite rigid. Bonds with partial double-bond character, such as the peptide bond, are also quite rigid. For that reason, the carbonyl (C=O) and imino (N=C) groups bound together by the peptide bond must lie in the same plane (Fig. 3-2). Much weaker, ionic bonds, on the other hand, impose no restrictions on the relative orientations of bonded atoms.

### Chemical Bonds Are Explainable in Quantum-Mechanical Terms

The nature of the forces, both strong and weak, that give rise to chemical bonds remained a mystery to chemists until the quantum theory of the atom (quantum mechanics) was developed in the 1920s. Then, for the first time, the various empirical laws regarding how chemical bonds are formed



**FIGURE 3-2** The planar shape of the peptide bond. Shown here is a portion of an extended polypeptide chain. Almost no rotation is possible about the peptide bond because of its partial double-bond character (see middle panel). All of the atoms in the shaded area must lie in the same plane. Rotation is possible, however, around the remaining two bonds, which make up the polypeptide configurations. (Adapted, with permission, from Pauling L. 1960. *The nature of the chemical bond and the structure of molecules and crystals: An introduction to modern structural chemistry*, 3rd ed., p. 495.) © Cornell University.

were put on a firm theoretical basis. It was realized that all chemical bonds, weak as well as strong, are based on electrostatic forces. Quantum mechanics provided explanations for covalent bonding by the sharing of electrons and also for the formation of weaker bonds.

### Chemical-Bond Formation Involves a Change in the Form of Energy

The spontaneous formation of a bond between two atoms always involves the release of some of the internal energy of the unbonded atoms and its conversion to another energy form. The stronger the bond, the greater is the amount of energy released upon its formation. The bonding reaction between two atoms A and B is thus described by



where AB represents the bonded aggregate. The rate of the reaction is directly proportional to the frequency of collision between A and B. The unit most often used to measure energy is the **calorie**, the amount of energy required to raise the temperature of 1 g of water from 14.5°C to 15.5°C. Because thousands of calories are usually involved in the breaking of a mole of chemical bonds, most energy changes within chemical reactions are expressed in kilocalories per mole (kcal/mol).

However, atoms joined by chemical bonds do not remain together forever, because there also exist forces that break chemical bonds. By far the most important of these forces arises from heat energy. Collisions with fast-moving molecules or atoms can break chemical bonds. During a collision, some of the kinetic energy of a moving molecule is given up as it pushes apart two bonded atoms. The faster a molecule is moving (the higher the temperature), the greater is the probability that, upon collision, it will break a bond. Hence, as the temperature of a collection of molecules is increased, the stability of their bonds decreases. The breaking of a bond is thus always indicated by the formula



The amount of energy that must be added to break a bond is exactly equal to the amount that was released upon formation of the bond. This equivalence follows from the first law of thermodynamics, which states that energy (except as it is interconvertible with mass) can be neither made nor destroyed.

### Equilibrium between Bond Making and Breaking

Every bond is thus a result of the combined actions of bond-making and bond-breaking forces. When an equilibrium is reached in a closed system, the number of bonds forming per unit time will equal the number of bonds breaking. Then the proportion of bonded atoms is described by the mass action formula:

$$K_{\text{eq}} = \frac{\text{conc}^{\text{AB}}}{\text{conc}^{\text{A}} \times \text{conc}^{\text{B}}}, \quad [\text{Equation 3-3}]$$

where  $K_{\text{eq}}$  is the **equilibrium constant**; and  $\text{conc}^{\text{A}}$ ,  $\text{conc}^{\text{B}}$ , and  $\text{conc}^{\text{AB}}$  are the concentrations of A, B, and AB, respectively, in moles per liter (mol/L). Whether we start with only free A and B, with only the molecule AB, or with a combination of AB and free A and B, at equilibrium the proportions of A, B, and AB will reach the concentrations given by  $K_{\text{eq}}$ .

## THE CONCEPT OF FREE ENERGY

---

There is always a change in the form of energy as the proportion of bonded atoms moves toward the **equilibrium concentration**. Biologically, the most useful way to express this energy change is through the physical chemist's concept of **free energy**, denoted by the symbol  $G$ , which honors the great 19th century physicist Josiah Gibbs. We shall not give a rigorous description of free energy in this text nor show how it differs from the other forms of energy. For this, the reader must refer to a chemistry text that discusses the second law of thermodynamics. It must suffice to say here that *free energy is energy that has the ability to do work*.

The second law of thermodynamics tells us that *a decrease in free energy ( $\Delta G$  is negative) always occurs in spontaneous reactions*. When equilibrium is reached, however, there is no further change in the amount of free energy ( $\Delta G = 0$ ). The equilibrium state for a closed collection of atoms is thus the state that contains the least amount of free energy.

The free energy lost as equilibrium is approached is either transformed into heat or used to increase the amount of entropy. We shall not attempt to define entropy here except to say that the amount of entropy is a measure of the amount of disorder. The greater the disorder, the greater is the amount of entropy. The existence of entropy means that many spontaneous chemical reactions (those with a net decrease in free energy) need not proceed with an evolution of heat. For example, when sodium chloride (NaCl) is dissolved in water, heat is absorbed rather than released. There is, nonetheless, a net decrease in free energy because of the increase in disorder of the sodium and chlorine ions as they move from a solid to a dissolved state.

### **$K_{\text{eq}}$ Is Exponentially Related to $\Delta G$**

Clearly, the stronger the bond, and hence the greater the change in free energy ( $\Delta G$ ) that accompanies its formation, the greater is the proportion of atoms that must exist in the bonded form. This common sense idea is quantitatively expressed by the physicochemical formula

$$\Delta G = -RT(\ln K_{\text{eq}}) \quad \text{or} \quad K_{\text{eq}} = e^{-\Delta G/RT}, \quad [\text{Equation 3-4}]$$

where  $R$  is the universal gas constant,  $T$  is the absolute temperature,  $\ln$  is the logarithm (of  $K_{\text{eq}}$ ) to the base  $e$ ,  $K_{\text{eq}}$  is the equilibrium constant, and  $e = 2.718$ .

Insertion of the appropriate values of  $R$  (1.987 cal/deg-mol) and  $T$  (298 at 25°C) tells us that  $\Delta G$  values as low as 2 kcal/mol can drive a bond-forming reaction to virtual completion if all reactants are present at molar concentrations (Table 3-1).

**TABLE 3-1** The Numerical Relationship between the Equilibrium Constant and  $\Delta G$  at 25°C

$K_{\text{eq}}$	$\Delta G$ (kcal/mol)
0.001	4.089
0.01	2.726
0.1	1.363
1.0	0
10.0	-1.363
100.0	-2.726
1000.0	-4.089

### **Covalent Bonds Are Very Strong**

The  $\Delta G$  values accompanying the formation of covalent bonds from free atoms, such as hydrogen or oxygen, are very large and negative in sign, usually -50 to -110 kcal/mol. Equation 3-4 tells us that  $K_{\text{eq}}$  of the bonding reaction will be correspondingly large, and thus the concentration of hydrogen or oxygen atoms existing unbound will be very small. For example, with a  $\Delta G$  value of -100 kcal/mol, if we start with 1 mol/L of the reacting atoms, only one in  $10^{40}$  atoms will remain unbound when equilibrium is reached.

## WEAK BONDS IN BIOLOGICAL SYSTEMS

The main types of weak bonds important in biological systems are the van der Waals bonds, hydrophobic bonds, hydrogen bonds, and ionic bonds. Sometimes, as we shall soon see, the distinction between a hydrogen bond and an ionic bond is arbitrary.

### Weak Bonds Have Energies between 1 and 7 kcal/mol

The weakest bonds are the van der Waals bonds. These have energies (1–2 kcal/mol) only slightly greater than the kinetic energy of heat motion. The energies of hydrogen and ionic bonds range between 3 and 7 kcal/mol.

In liquid solutions, almost all molecules form several weak bonds to nearby atoms. All molecules are able to form van der Waals bonds, whereas hydrogen and ionic bonds can form only between molecules that have a net charge (ions) or in which the charge is unequally distributed. Some molecules thus have the capacity to form several types of weak bonds. Energy considerations, however, tell us that molecules always have a greater tendency to form the stronger bond.

### Weak Bonds Are Constantly Made and Broken at Physiological Temperatures

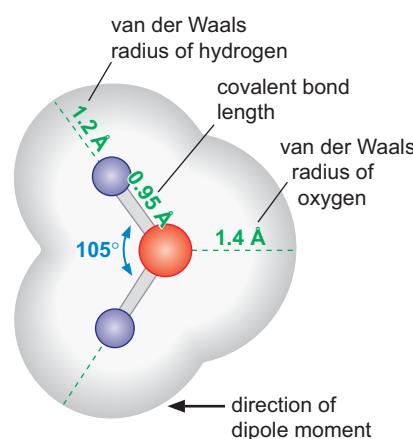
Weak bonds, at their weakest, have energies only slightly higher than the average energy of kinetic motion (heat) at 25°C (0.6 kcal/mol), but even the strongest of these bonds have only about 10 times that energy. Nevertheless, because there is a significant spread in the energies of kinetic motion, many molecules with sufficient kinetic energy to break the strongest weak bonds still exist at physiological temperatures.

## The Distinction between Polar and Nonpolar Molecules

All forms of weak interactions are based on attractions between electric charges. The separation of electric charges can be permanent or temporary, depending on the atoms involved. For example, the oxygen molecule ( $O_2$ ) has a symmetric distribution of electrons between its two oxygen atoms, therefore each of its two atoms is uncharged. In contrast, there is a non-uniform distribution of charge in water ( $H_2O$ ), in which the bond electrons are unevenly shared (Fig. 3-3). They are held more strongly by the oxygen atom, which thus carries a considerable negative charge, whereas the two hydrogen atoms together have an equal amount of positive charge. The center of the positive charge is on one side of the center of the negative charge. A combination of separated positive and negative charges is called an electric **dipole moment**. Unequal electron sharing reflects dissimilar affinities of the bonding atoms for electrons. Atoms that have a tendency to gain electrons are called **electronegative** atoms. **Electropositive** atoms have a tendency to give up electrons.

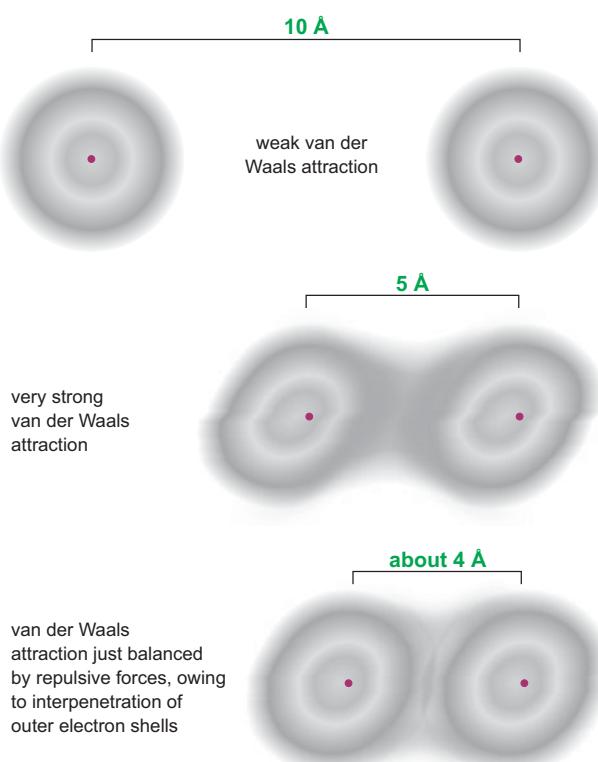
Molecules (such as  $H_2O$ ) that have a dipole moment are called **polar molecules**. **Nonpolar molecules** are those with no effective dipole moments. In methane ( $CH_4$ ), for example, the carbon and hydrogen atoms have similar affinities for their shared electron pairs, thus neither the carbon nor the hydrogen atom is noticeably charged.

The distribution of charge in a molecule can also be affected by the presence of nearby molecules, particularly if the affected molecule is polar. The



**FIGURE 3-3** The structure of a water molecule. For van der Waals radii, see Figure 3-5.

**FIGURE 3-4** Variation of van der Waals forces with distance. The atoms shown in this diagram are atoms of the inert rare gas argon. (Adapted from Pauling L. 1953. *General chemistry*, 2nd ed., p. 322. Courtesy Ava Helen and Linus Pauling Papers, Oregon State University Libraries.)



effect may cause a nonpolar molecule to acquire a slightly polar character. If the second molecule is not polar, its presence will still alter the nonpolar molecule, establishing a fluctuating charge distribution. Such induced effects, however, give rise to a much smaller separation of charge than is found in polar molecules, resulting in smaller interaction energies and correspondingly weaker chemical bonds.

### van der Waals Forces

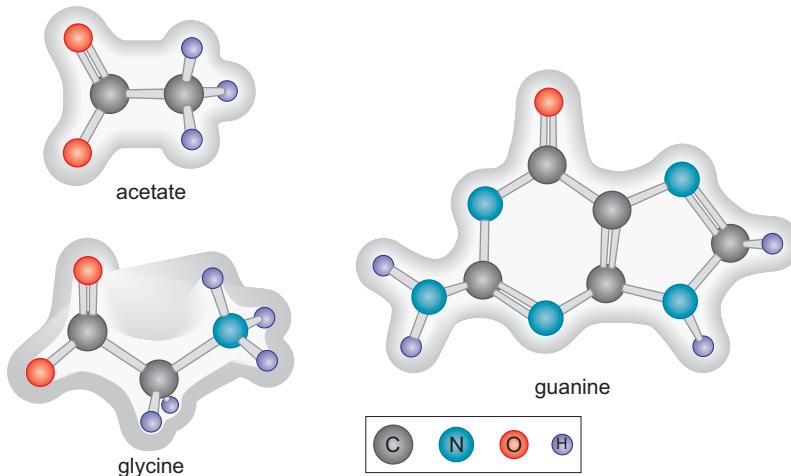
**van der Waals bonding** arises from a nonspecific attractive force originating when two atoms come close to each other. It is based not on the existence of permanent charge separations, but, rather, on the induced fluctuating charges caused by the nearness of molecules. It therefore operates between all types of molecules, nonpolar as well as polar. It depends heavily on the distance between the interacting groups, because the bond energy is inversely proportional to the sixth power of distance (Fig. 3-4).

There also exists a more powerful van der Waals *repulsive* force, which comes into play at even shorter distances. This repulsion is caused by the overlapping of the outer electron shells of the atoms involved. The van der Waals attractive and repulsive forces balance at a certain distance specific for each type of atom. This distance is the so-called **van der Waals radius** (Table 3-2; Fig. 3-5). The van der Waals bonding energy between two atoms separated by the sum of their van der Waals radii increases with the size of the respective atoms. For two average atoms, it is only  $\sim 1$  kcal/mol, which is just slightly more than the average thermal energy of molecules at room temperature (0.6 kcal/mol).

This means that van der Waals forces are an effective binding force at physiological temperatures only when several atoms in a given molecule are bound to several atoms in another molecule or another part of the same molecule. Then the energy of interaction is much greater than the

**TABLE 3-2** van der Waals Radii of the Atoms in Biological Molecules

Atom	van der Waals Radius (Å)
H	1.2
N	1.5
O	1.4
P	1.9
S	1.85
CH <sub>3</sub> group	2.0
Half thickness of aromatic molecule	1.7

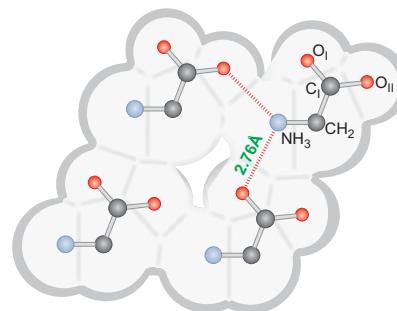


**FIGURE 3-5** Drawings of several molecules with the van der Waals radii of the atoms shown as shaded clouds.

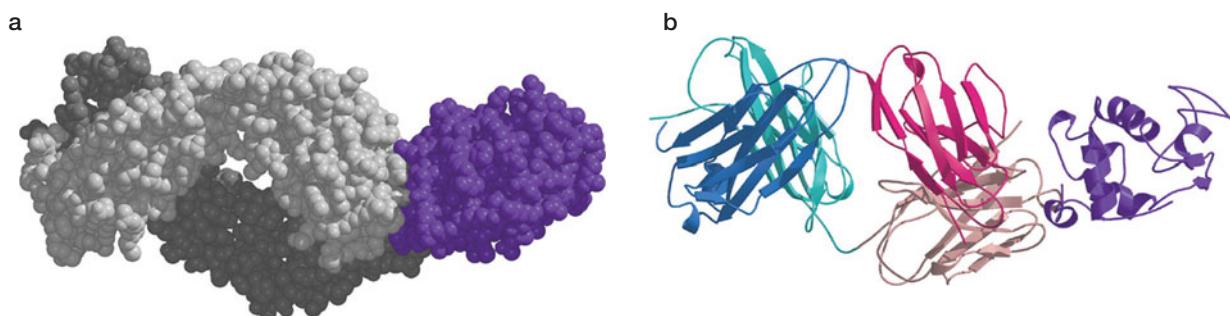
dissociating tendency resulting from random thermal movements. For several atoms to interact effectively, the molecular fit must be precise because the distance separating any two interacting atoms must not be much greater than the sum of their van der Waals radii (Fig. 3-6). The strength of interaction rapidly approaches zero when this distance is only slightly exceeded. Thus, the strongest type of van der Waals contact arises when a molecule contains a cavity exactly complementary in shape to a protruding group of another molecule, as is the case with an antigen and its specific antibody (Fig. 3-7). In this instance, the binding energies sometimes can be as large as 20–30 kcal/mol, so that antigen–antibody complexes seldom fall apart. The bonding pattern of polar molecules is rarely dominated by van der Waals interactions because such molecules can acquire a lower energy state (lose more free energy) by forming other types of bonds.

### Hydrogen Bonds

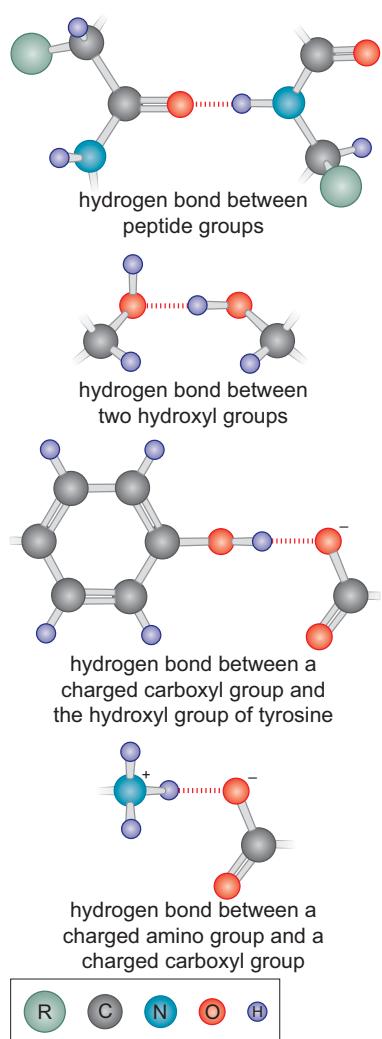
A **hydrogen bond** is formed between a donor hydrogen atom with some positive charge and a negatively charged acceptor atom (Fig. 3-8). For example, the hydrogen atoms of the amino ( $-\text{NH}_2$ ) group are attracted by the negatively charged keto ( $-\text{C}=\text{O}$ ) oxygen atoms. Sometimes, the hydrogen-



**FIGURE 3-6** The arrangement of molecules in a layer of a crystal formed by the amino acid glycine. The packing of the molecules is determined by the van der Waals radii of the groups, except for the  $\text{N}-\text{H} \cdots \text{O}$  contacts, which are shortened by the formation of hydrogen bonds. (Adapted, with permission, from Pauling L. 1960. *The nature of the chemical bond and the structure of molecules and crystals: An introduction to modern structural chemistry*, 3rd ed., p. 262. © Cornell University.)



**FIGURE 3-7** Antibody–antigen interaction. The structures, depicted as space filling (a) and as ribbons (b), show the complex between Fab D 1.3 and lysozyme (in purple). (Fischmann T.O. et al. 1991. *J. Biol. Chem.* **266**: 12915.) Images prepared with MolScript, BobScript, and Raster3D.



**FIGURE 3-8** Examples of hydrogen bonds in biological molecules.

bonded atoms belong to groups with a unit of charge (such as  $\text{NH}_3^+$  or  $\text{COO}^-$ ). In other cases, both the donor hydrogen atoms and the negative acceptor atoms have less than a unit of charge.

The biologically most important hydrogen bonds involve hydrogen atoms covalently bound to oxygen atoms ( $\text{O}-\text{H}$ ) or nitrogen atoms ( $\text{N}-\text{H}$ ). Likewise, the negative acceptor atoms are usually nitrogen or oxygen. Table 3-3 lists some of the most important hydrogen bonds. In the absence of surrounding water molecules, bond energies range between 3 and 7 kcal/mol, the stronger bonds involving the greater charge differences between donor and acceptor atoms. Hydrogen bonds are thus weaker than covalent bonds, yet considerably stronger than van der Waals bonds. A hydrogen bond, therefore, will hold two atoms closer together than the sum of their van der Waals radii, but not so close together as a covalent bond would hold them.

Hydrogen bonds, unlike van der Waals bonds, are highly directional. In the strongest hydrogen bonds, the hydrogen atom points directly at the acceptor atom (Fig. 3-9). If it points more than  $30^\circ$  away, the bond energy is much less. Hydrogen bonds are also much more specific than van der Waals bonds because they demand the existence of molecules with complementary donor hydrogen and acceptor groups.

### Some Ionic Bonds Are Hydrogen Bonds

Many organic molecules possess ionic groups that contain one or more units of net positive or negative charge. The negatively charged mononucleotides, for example, contain phosphate groups, which are negatively charged, whereas each amino acid (except proline) has a negative carboxyl group ( $\text{COO}^-$ ) and a positive amino group ( $\text{NH}_3^+$ ), both of which carry a unit of charge. These charged groups are usually neutralized by nearby, oppositely charged groups. The electrostatic forces acting between the oppositely charged groups are called **ionic bonds**. Their average bond energy in an aqueous solution is  $\sim 5$  kcal/mol.

In many cases, either an inorganic cation like  $\text{Na}^+$ ,  $\text{K}^+$ , or  $\text{Mg}^{2+}$  or an inorganic anion like  $\text{Cl}^-$  or  $\text{SO}_4^{2-}$  neutralizes the charge of ionized organic molecules. When this happens in aqueous solution, the neutralizing cations and anions do not carry fixed positions because inorganic ions are usually surrounded by shells of water molecules and thus do not directly bind to oppositely charged groups. Thus, in water solutions, electrostatic bonds to surrounding inorganic cations or anions are usually not of primary importance in determining the molecular shapes of organic molecules.

On the other hand, highly directional bonds result if the oppositely charged groups can form hydrogen bonds to each other. For example,  $\text{COO}^-$  and  $\text{NH}_3^+$  groups are often held together by hydrogen bonds. Because these bonds are stronger than those that involve groups with less than a unit of charge, they are correspondingly shorter. A strong hydrogen bond can also form between a group with a unit charge and a group having less than a unit charge. For example, a hydrogen atom belonging to an amino group ( $\text{NH}_2$ ) bonds strongly to an oxygen atom of a carboxyl group ( $\text{COO}^-$ ).

**TABLE 3-3** Approximate Bond Lengths of Biologically Important Hydrogen Bonds

Bond	Approximate H-Bond Length (Å)
$\text{O}-\text{H}\cdots\text{O}$	$2.70 \pm 0.10$
$\text{O}-\text{H}\cdots\text{O}^-$	$2.63 \pm 0.10$
$\text{O}-\text{H}\cdots\text{N}$	$2.88 \pm 0.13$
$\text{N}-\text{H}\cdots\text{O}$	$3.04 \pm 0.13$
$\text{N}^+-\text{H}\cdots\text{O}$	$2.93 \pm 0.10$
$\text{N}-\text{H}\cdots\text{N}$	$3.10 \pm 0.13$

### Weak Interactions Demand Complementary Molecular Surfaces

Weak binding forces are effective only when the interacting surfaces are close. This proximity is possible only when the molecular surfaces have **complementary structures**, so that a protruding group (or positive charge) on one surface is matched by a cavity (or negative charge) on another. That is, the interacting molecules must have a **lock-and-key relationship**.

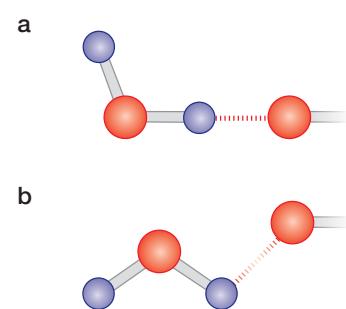
In cells, this requirement often means that some molecules hardly ever bond to other molecules of the same kind because such molecules do not have the properties of symmetry necessary for self-interaction. For example, some polar molecules contain donor hydrogen atoms and no suitable acceptor atoms, whereas other molecules can accept hydrogen bonds but have no hydrogen atoms to donate. On the other hand, there are many molecules with the necessary symmetry to permit strong self-interaction in cells. Water is the most important example of this.

### Water Molecules Form Hydrogen Bonds

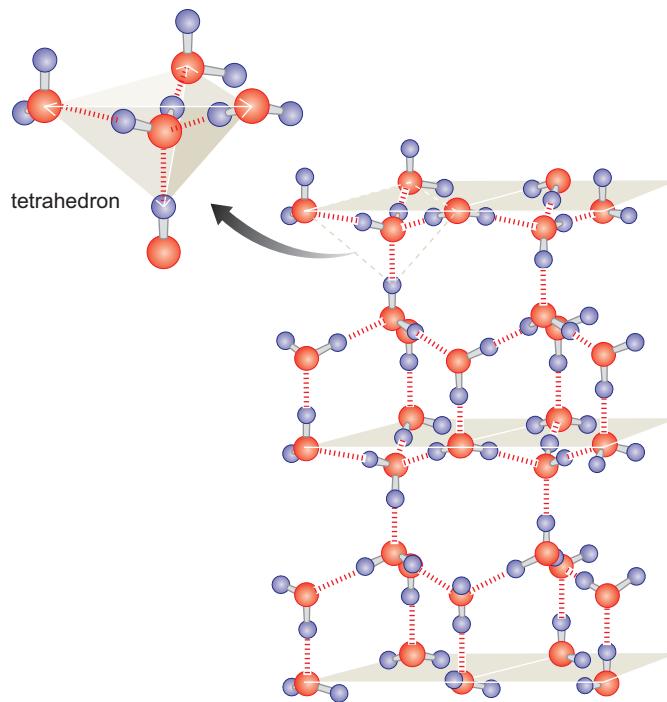
Under physiological conditions, water molecules rarely ionize to form  $\text{H}^+$  and  $\text{OH}^-$  ions. Instead, they exist as polar  $\text{H}-\text{O}-\text{H}$  molecules with both the hydrogen and oxygen atoms forming strong hydrogen bonds. In each water molecule, the oxygen atom can bind to two external hydrogen atoms, whereas each hydrogen atom can bind to one adjacent oxygen atom. These bonds are directed tetrahedrally (Fig. 3-10), and thus, in its solid and liquid forms, each water molecule tends to have four **nearest neighbors**, one in each of the four directions of a tetrahedron. In ice, the bonds to these neighbors are very rigid and the arrangement of molecules fixed. Above the melting temperature ( $0^\circ\text{C}$ ), the energy of thermal motion is sufficient to break the hydrogen bonds and to allow the water molecules to change their nearest neighbors continually. Even in the liquid form, however, at any given instant most water molecules are bound by four strong hydrogen bonds.

### Weak Bonds between Molecules in Aqueous Solutions

The average energy of a secondary, weak bond, although small compared with that of a covalent bond, is nonetheless strong enough compared with heat energy to ensure that most molecules in aqueous solution will form secondary bonds to other molecules. The proportion of bonded to non-



**FIGURE 3-9** Directional properties of hydrogen bonds. (a) The vector along the covalent  $\text{O}-\text{H}$  bond points directly at the acceptor oxygen, thereby forming a strong bond. (b) The vector points away from the oxygen atom, resulting in a much weaker bond.



**FIGURE 3-10** Diagram of a lattice formed by water molecules. The energy gained by forming specific hydrogen bonds between water molecules favors the arrangement of the molecules in adjacent tetrahedrons. (Red spheres) Oxygen atoms; (purple spheres) hydrogen atoms. Although the rigidity of the arrangement depends on the temperature of the molecules, the pictured structure is nevertheless predominant in water as well as in ice. (Adapted, with permission, from Pauling L. 1960. *The nature of the chemical bond and the structure of molecules and crystals: An introduction to modern structural chemistry*, 3rd ed., p. 262. © Cornell University.)

bonded arrangements is given by Equation 3-4, corrected to take into account the high concentration of molecules in a liquid. It tells us that interaction energies as low as 2–3 kcal/mol are sufficient at physiological temperatures to force most molecules to form the maximum number of strong secondary bonds.

The specific structure of a solution at a given instant is markedly influenced by which solute molecules are present, not only because molecules have specific shapes, but also because molecules differ in which types of secondary bonds they can form. Thus, a molecule will tend to move until it is next to a molecule with which it can form the strongest possible bond.

Solutions, of course, are not static. Because of the disruptive influence of heat, the specific configuration of a solution is constantly changing from one arrangement to another of approximately the same energy content. Equally important in biological systems is the fact that metabolism is continually transforming one molecule into another and thus automatically changing the nature of the secondary bonds that can be formed. The solution structure of cells is thus constantly disrupted not only by heat motion, but also by the metabolic transformations of the cell's solute molecules.

### Organic Molecules That Tend to Form Hydrogen Bonds Are Water Soluble

The energy of hydrogen bonds per atomic group is much greater than that of van der Waals contacts; thus, molecules will form hydrogen bonds in preference to van der Waals contacts. For example, if we try to mix water with a compound that cannot form hydrogen bonds, such as benzene, the water and benzene molecules rapidly separate from each other, the water molecules forming hydrogen bonds among themselves while the benzene molecules attach to one another by van der Waals bonds. It is therefore impossible to insert a non-hydrogen-bonding organic molecule into water.

On the other hand, polar molecules such as glucose and pyruvate, which contain a large number of groups that form excellent hydrogen bonds (such as  $=O$  or  $OH$ ), are soluble in water (i.e., they are **hydrophilic** as opposed to **hydrophobic**). Although the insertion of such groups into a water lattice breaks water–water hydrogen bonds, it results simultaneously in the formation of hydrogen bonds between the polar organic molecule and water. These alternative arrangements, however, are not usually as energetically satisfactory as the water–water arrangements, thus even the most polar molecules ordinarily have only limited solubility (see Box 3-1).

Therefore, almost all of the molecules that cells acquire, either through food intake or through biosynthesis, are somewhat insoluble in water. These molecules, by their thermal movements, randomly collide with other molecules until they find complementary molecular surfaces on which to attach and thereby release water molecules for water–water interactions.

### Hydrophobic “Bonds” Stabilize Macromolecules

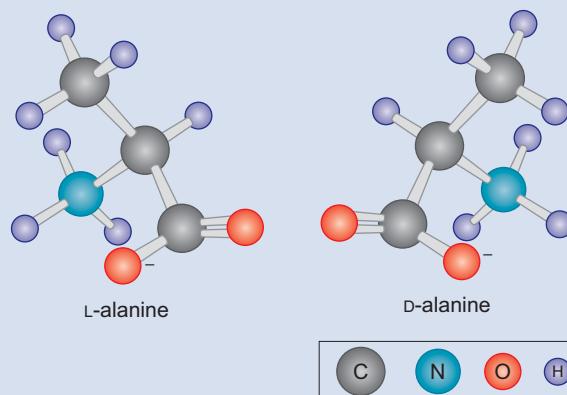
The strong tendency of water to exclude nonpolar groups is frequently referred to as **hydrophobic bonding**. Some chemists like to call all of the bonds between nonpolar groups *in a water solution hydrophobic bonds* (Fig. 3-11). But, in a sense, this term is a misnomer, for the phenomenon that it seeks to emphasize is the *absence*, not the *presence*, of bonds. (The bonds that tend to form between the nonpolar groups are due to van der Waals attractive forces.) On the other hand, the term *hydrophobic bond* is often useful because it emphasizes the fact that nonpolar groups will try

## ► ADVANCED CONCEPTS

**Box 3-1** The Uniqueness of Molecular Shapes and the Concept of Selective Stickiness

Even though most cellular molecules are built up from only a small number of chemical groups, such as OH, NH<sub>2</sub>, and CH<sub>3</sub>, there is great specificity as to which molecules tend to lie next to each other. This is because each molecule has unique bonding properties. One very clear demonstration comes from the specificity of stereoisomers. For example, proteins are always constructed from L-amino acids, never from their mirror images, the D-amino acids (Box 3-1 Fig. 1). Although the D- and L-amino acids have identical covalent bonds, their binding properties to asymmetric molecules are often very different. Thus, most enzymes are specific for L-amino acids. If an L-amino acid is able to attach to a specific enzyme, the D-amino acid is unable to bind.

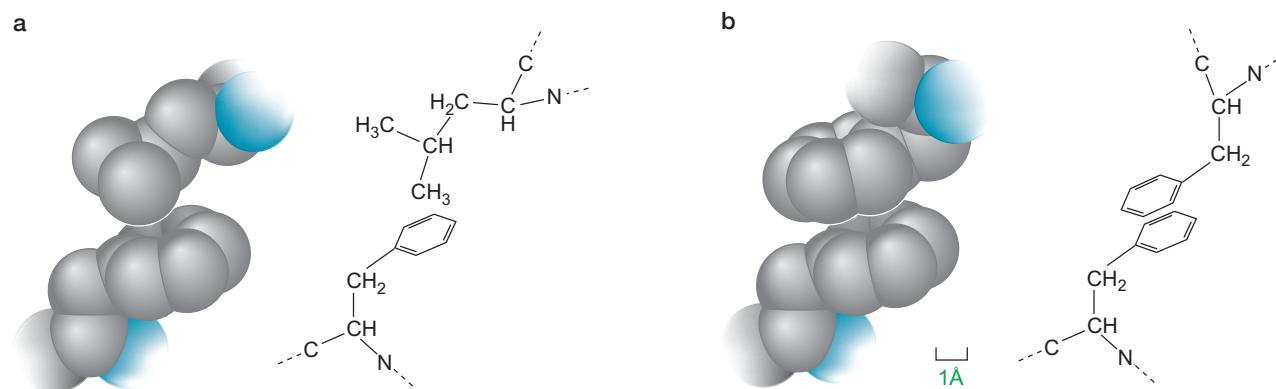
Most molecules in cells can make good “weak” bonds with only a small number of other molecules, partly because most molecules in biological systems exist in an aqueous environment. The formation of a bond in a cell therefore depends not only on whether two molecules bind well to each other, but also on whether bond formation is overall more favorable than the alternative bonds that can form with solvent water molecules.



**BOX 3-1 FIGURE 1** The two stereoisomers of the amino acid alanine. (Adapted, with permission, from Pauling L. 1960. *The nature of the chemical bond and the structure of molecules and crystals: An introduction to modern structural chemistry*, 3rd ed., p. 465. © Cornell University; Pauling L. 1953. *General chemistry*, 2nd ed., p. 498. Courtesy Ava Helen and Linus Pauling Papers, Oregon State University Libraries.)

to arrange themselves so that they are not in contact with water molecules. Hydrophobic bonds are important both in the stabilization of proteins and complexes of proteins with other molecules and in the partitioning of proteins into membranes. They may account for as much as one-half of the total free energy of protein folding.

Consider, for example, the different amounts of energy generated when the amino acids alanine and glycine are bound, in water, to a third molecule that has a surface complementary to alanine. A methyl group is present in alanine but not in glycine. When alanine is bound to the third



**FIGURE 3-11** Examples of van der Waals (hydrophobic) bonds between the nonpolar side groups of amino acids. The hydrogens are not indicated individually. For the sake of clarity, the van der Waals radii are reduced by 20%. The structural formulas adjacent to each space-filling drawing indicate the arrangement of the atoms. (a) Phenylalanine–leucine bond. (b) Phenylalanine–phenylalanine bond. (Adapted, with permission, from Scheraga H.A. 1963. *The proteins*, 2nd ed. [ed. H. Neurath], p. 527. Academic Press, New York. © Harold Scheraga.)

molecule, the van der Waals contacts around the methyl group yield 1 kcal/mol of energy, which is not released when glycine is bound instead. From Equation 3-4, we know that this small energy difference alone would give only a factor of 6 between the binding of alanine and glycine. However, this calculation does not take into consideration the fact that water is trying to exclude alanine much more than glycine. The presence of alanine's  $\text{CH}_3$  group upsets the water lattice much more seriously than does the hydrogen atom side group of glycine. At present, it is still difficult to predict how large a correction factor must be introduced for this disruption of the water lattice by the hydrophobic side groups. It is likely that the water tends to exclude alanine, thrusting it toward a third molecule, with a hydrophobic force of  $\sim 2\text{--}3$  kcal/mol larger than the forces excluding glycine.

We thus arrive at the important conclusion that the energy difference between the binding of even the most similar molecules to a third molecule (when the difference between the similar molecules involves a non-polar group) is at least 2–3 kcal/mol greater in the aqueous interior of cells than under non-aqueous conditions. Frequently, the energy difference is 3–4 kcal/mol, because the molecules involved often contain polar groups that can form hydrogen bonds.

### The Advantage of $\Delta G$ between 2 and 5 kcal/mol

We have seen that the energy of just one secondary bond (2–5 kcal/mol) is often sufficient to ensure that a molecule preferentially binds to a selected group of molecules. Moreover, these energy differences are not so large that rigid lattice arrangements develop within a cell; that is, the interior of a cell never crystallizes, as it would if the energy of secondary bonds were several times greater. Larger energy differences would mean that the secondary bonds would seldom break, resulting in low diffusion rates incompatible with cellular existence.

### Weak Bonds Attach Enzymes to Substrates

Weak bonds are necessarily the basis by which enzymes and their substrates initially combine with each other. Enzymes do not indiscriminately bind all molecules, having noticeable affinity only for their own substrates.

Because enzymes catalyze both directions of a chemical reaction, they must have specific affinities for both sets of reacting molecules. In some cases, it is possible to measure an equilibrium constant for the binding of an enzyme to one of its substrates (Equation 3-4), which consequently enables us to calculate the  $\Delta G$  upon binding. This calculation, in turn, hints at which types of bonds may be involved. For  $\Delta G$  values between 5 and 10 kcal/mol, several strong secondary bonds are the basis of specific enzyme–substrate interactions. Also worth noting is that the  $\Delta G$  of binding is never exceptionally high; thus, enzyme–substrate complexes can be both made and broken apart rapidly as a result of random thermal movement. This explains why enzymes can function quickly, sometimes as often as  $10^6$  times per second. If enzymes were bound to their substrates, or more importantly to their products, by more powerful bonds, they would act much more slowly.

### Weak Bonds Mediate Most Protein–DNA and Protein–Protein Interactions

As we shall see throughout the book, interactions between proteins and DNA, and between proteins and other proteins, lie at the heart of how cells

detect and respond to signals; express genes; replicate, repair, and recombine their DNA; and so on. These interactions—which clearly play an important role in how those cellular processes are regulated—are mediated by weak chemical bonds of the sort we have described in this chapter. Despite the low energy of each individual bond, affinity in these interactions, and specificity as well, results from the combined effects of many such bonds between any two interacting molecules.

In Chapter 6, we return to these matters with a detailed look at how proteins are built, how they adopt particular structures, and how they bind DNA, RNA, and each other.

## HIGH-ENERGY BONDS

---

We now turn to high-energy covalent bonds in biological systems. So far we have considered the formation of weak bonds from the thermodynamic viewpoint. Each time a potential weak bond was considered, the question was posed: Does its formation involve a gain or a loss of free energy? Only when  $\Delta G$  is negative does the thermodynamic equilibrium favor a reaction. This same approach is equally valid for covalent bonds. The fact that enzymes are usually involved in the making or breaking of a covalent bond does not in any sense alter the requirement of a negative  $\Delta G$ .

Upon superficial examination, however, many of the important covalent bonds in cells appear to be formed in violation of the laws of thermodynamics, particularly those bonds joining small molecules together to form large polymeric molecules. The formation of such bonds involves an increase in free energy. Originally, this fact suggested to some people that cells had the unique ability to work in violation of thermodynamics and that this property was, in fact, the real “secret of life.”

Now, however, it is clear that these biosynthetic processes do not violate thermodynamics but, rather, are based on different reactions from those originally postulated. Nucleic acids, for example, do not form by the condensation of nucleoside phosphates; glycogen is not formed directly from glucose residues; proteins are not formed by the union of amino acids. Instead, the monomeric precursors, using energy present in ATP, are first converted to high-energy “activated” precursors, which then spontaneously (with the help of specific enzymes) unite to form larger molecules. Below, we illustrate these ideas by concentrating on the thermodynamics of peptide (protein) and phosphodiester (nucleic acid) bonds. First, however, we must briefly look at some general thermodynamic properties of covalent bonds.

## MOLECULES THAT DONATE ENERGY ARE THERMODYNAMICALLY UNSTABLE

---

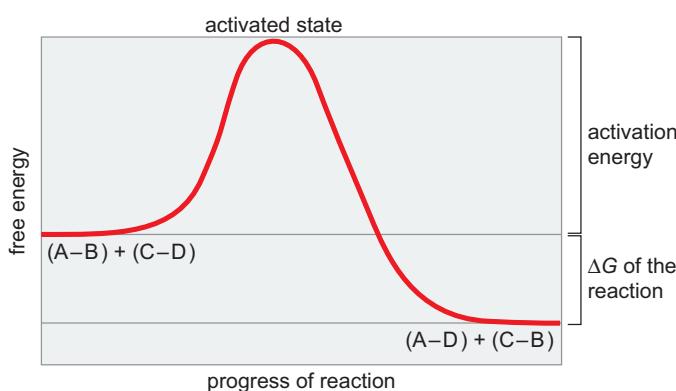
There is great variation in the amount of free energy possessed by specific molecules. This is because covalent bonds do not all have the same bond energy. As an example, the covalent bond between oxygen and hydrogen is considerably stronger than the bond between hydrogen and hydrogen, or oxygen and oxygen. The formation of an O—H bond at the expense of O—O or H—H will thus release energy. Energy considerations, therefore, tell us that a sufficiently concentrated mixture of oxygen and hydrogen will be transformed into water.

A molecule thus possesses a larger amount of free energy if linked together by weak covalent bonds than if it is linked together by strong bonds. This idea seems almost paradoxical at first glance because it means that the stronger the bond, the less energy it can give off. But the notion automatically makes sense when we realize that an atom that has formed a very strong bond has already lost a large amount of free energy in this process. Therefore, the best food molecules (molecules that donate energy) are those molecules that contain weak covalent bonds and are therefore thermodynamically unstable.

For example, glucose is an excellent food molecule because there is a great decrease in free energy when it is oxidized by oxygen to yield carbon dioxide and water. On the other hand, carbon dioxide, composed of strong covalent double bonds between carbon and oxygen, known as **carbonyl bonds**, is not a food molecule in animals. In the absence of the energy donor ATP, carbon dioxide cannot be transformed spontaneously into more complex organic molecules, even with the help of specific enzymes. Carbon dioxide can be used as a primary source of carbon in plants only because the energy supplied by light quanta during photosynthesis results in the formation of ATP.

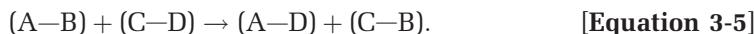
The chemical reactions by which molecules are transformed into other molecules containing less free energy do not occur at significant rates at physiological temperatures in the absence of a catalyst. This is because even a weak covalent bond is, in reality, very strong and is only rarely broken by thermal motion within a cell. For a covalent bond to be broken in the absence of a catalyst, energy must be supplied to push apart the bonded atoms. When the atoms are partially apart, they can recombine with new partners to form stronger bonds. In the process of recombination, the energy released is the sum of the free energy supplied to break the old bond plus the difference in free energy between the old and the new bond (Fig. 3-12).

The energy that must be supplied to break the old covalent bond in a molecular transformation is called the **activation energy**. The activation energy is usually less than the energy of the original bond because molecular rearrangements generally do not involve the production of completely free atoms. Instead, a collision between the two reacting molecules is required, followed by the temporary formation of a molecular complex called the **activated state**. In the activated state, the close proximity of the two molecules makes each other's bonds more labile, so that less energy is needed to break a bond than when the bond is present in a free molecule.



**FIGURE 3-12** The energy of activation of a chemical reaction:  $(A-B)+(C-D) \rightarrow (A-D)+(C-B)$ . This reaction is accompanied by a decrease in free energy.

Most reactions of covalent bonds in cells are therefore described by



The mass action expression for such a reaction is

$$K_{\text{eq}} = \frac{\text{conc}^{A-D} \times \text{conc}^{C-B}}{\text{conc}^{A-B} \times \text{conc}^{C-D}}, \quad [\text{Equation 3-6}]$$

where  $\text{conc}^{A-B}$ ,  $\text{conc}^{C-D}$ , and so on are the concentrations of the several reactants in moles per liter. Here, also, the value of  $K_{\text{eq}}$  is related to  $\Delta G$  by (see also Table 3-4)

$$\Delta G = -RT \ln K_{\text{eq}} \quad \text{or} \quad K_{\text{eq}} = e^{-\Delta G/RT}. \quad [\text{Equation 3-7}]$$

Because energies of activation are generally between 20 and 30 kcal/mol, activated states practically never occur at physiological temperatures. High activation energies are thus barriers preventing spontaneous rearrangements of cellular-covalent bonds.

These barriers are enormously important. Life would be impossible if they did not exist, because all atoms would be in the state of least possible energy. There would be no way to store energy temporarily for future work. On the other hand, life would also be impossible if means were not found to lower the activation energies of certain reactions selectively. This also must happen if cell growth is to occur at a rate sufficiently fast so as not to be seriously impeded by random destructive forces, such as ionization or ultraviolet radiation.

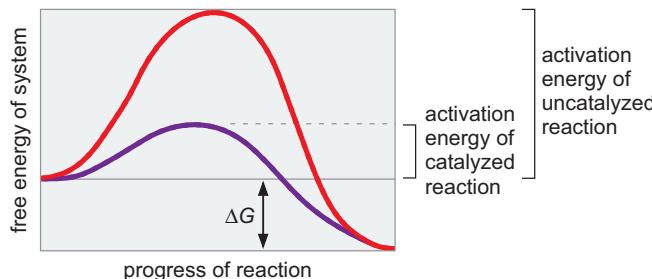
**TABLE 3-4** The Relationship between  $K_{\text{eq}}$  and  $\Delta G$  [ $\Delta G = -RT(\ln K_{\text{eq}})$ ]

$K_{\text{eq}}$	$\Delta G$ (kcal/mol)
$10^{-6}$	8.2
$10^{-5}$	6.8
$10^{-4}$	5.1
$10^{-3}$	4.1
$10^{-2}$	2.7
$10^{-1}$	1.4
$10^0$	0.0
$10^1$	-1.4
$10^2$	-2.7
$10^3$	-4.1

## ENZYMES LOWER ACTIVATION ENERGIES IN BIOCHEMICAL REACTIONS

Enzymes are absolutely necessary for life. The **function of enzymes** is to speed up the **rate** of the chemical reactions requisite to cellular existence by lowering the activation energies of molecular rearrangements to values that can be supplied by the heat of motion (Fig. 3-13). When a specific enzyme is present, there is no longer an effective barrier preventing the rapid formation of the reactants possessing the lowest amounts of free energy. Enzymes never affect the nature of an equilibrium: They merely speed up the rate at which it is reached. Thus, if the thermodynamic equilibrium is unfavorable for the formation of a molecule, the presence of an enzyme can in no way effect the molecule's accumulation.

Because enzymes must catalyze essentially every cellular molecular rearrangement, knowing the free energy of various molecules cannot by itself tell us whether an energetically feasible rearrangement will, in fact, occur. The rate of the reactions must always be considered. Only if a cell possesses a suitable enzyme will the reaction be important.



**FIGURE 3-13** Enzymes lower activation energies and thus speed up the rate of the reaction. The enzyme-catalyzed reaction is shown by the purple curve. Note that  $\Delta G$  remains the same because the equilibrium position remains unaltered.

## FREE ENERGY IN BIOMOLECULES

---

Thermodynamics tells us that all biochemical pathways must be characterized by a decrease in free energy. This is clearly the case for degradative pathways, in which thermodynamically unstable food molecules are converted to more stable compounds, such as carbon dioxide and water, with the evolution of heat. All degradative pathways have two primary purposes: (1) to produce the small organic fragments necessary as building blocks for larger organic molecules and (2) to conserve a significant fraction of the free energy of the original food molecule in a form that can do work. This latter purpose is accomplished by coupling some of the steps in degradative pathways with the simultaneous formation of high-energy molecules such as ATP, which can store free energy.

Not all of the free energy of a food molecule is converted into the free energy of high-energy molecules. If this were the case, a degradative pathway would not be characterized by a decrease in free energy, and there would be no driving force to favor the breakdown of food molecules. Instead, we find that all degradative pathways are characterized by a conversion of at least one-half of the free energy of the food molecule into heat and/or entropy. For example, it is estimated that in cells,  $\sim 40\%$  of the free energy of glucose is used to make new high-energy compounds, the remainder being dissipated into heat energy and entropy.

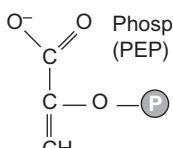
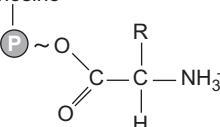
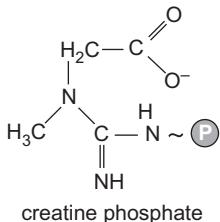
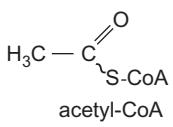
### High-Energy Bonds Hydrolyze with Large Negative $\Delta G$

A high-energy molecule contains one or more bonds whose breakdown by water, called **hydrolysis**, is accompanied by a large decrease in free energy. The specific bonds whose hydrolysis yields these large negative  $\Delta G$  values are called **high-energy bonds**—a somewhat misleading term because it is not the bond energy but the free energy of hydrolysis that is high. Nonetheless, the term *high-energy bond* is generally used, and for convenience, we shall continue this usage by designating high-energy bonds with the symbol  $\sim$ .

The energy of hydrolysis of the average high-energy bond ( $\approx 7$  kcal/mol) is very much smaller than the amount of energy that would be released if a glucose molecule were to be completely degraded in one step (688 kcal/mol). A one-step breakdown of glucose would be inefficient in making high-energy bonds. This is undoubtedly the reason why biological glucose degradation requires so many steps. In this way, the amount of energy released per degradative step is of the same order of magnitude as the free energy of hydrolysis of a high-energy bond.

The most important high-energy compound is ATP. It is formed from inorganic phosphate ( $\textcircled{P}$ ) and ADP, using energy obtained either from degradative reactions or from the Sun, a process known as **photosynthesis**. There are, however, many other important high-energy compounds. Some are directly formed during degradative reactions; others are formed using some of the free energy of ATP. Table 3-5 lists the most important types of high-energy bonds. All involve either phosphate or sulfur atoms. The high-energy pyrophosphate bonds of ATP arise from the union of phosphate groups. The pyrophosphate linkage ( $\textcircled{P} \sim \textcircled{P}$ ) is not, however, the only kind of high-energy phosphate bond: The attachment of a phosphate group to the oxygen atom of a carboxyl group creates a high-energy acyl bond. It is now clear that high-energy bonds involving sulfur atoms play almost as important a role in energy metabolism as those involving phosphorus. The most important molecule containing a high-energy sulfur bond is acetyl-CoA. This bond is the main source of energy for fatty acid biosynthesis.

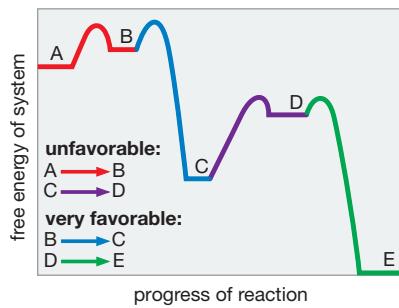
**TABLE 3-5** Important Classes of High-Energy Bonds

Class	Molecular Example	Reaction	$\Delta G$ of Reaction (kcal/mol)
Pyrophosphate	$\text{P} \sim \text{P}$ pyrophosphate	$\text{P} \sim \text{P} \rightleftharpoons \text{P} + \text{P}$	$\Delta G = -6$
Nucleoside diphosphates	Adenosine— $\text{P} \sim \text{P}$ (ADP)	$\text{ADP} \rightleftharpoons \text{AMP} + \text{P}$	$\Delta G = -6$
Nucleoside triphosphates	Adenosine— $\text{P} \sim \text{P} \sim \text{P}$ (ATP)	$\text{ATP} \rightleftharpoons \text{ADP} + \text{P}$	$\Delta G = -7$
		$\text{ATP} \rightleftharpoons \text{AMP} + \text{P} \sim \text{P}$	
Enol phosphates	 Phosphoenolpyruvate (PEP)	$\text{PEP} \rightleftharpoons \text{Pyruvate} + \text{P}$	$\Delta G = -12$
Aminoacyl adenylates	Adenosine 	$\text{AM} \text{P} \sim \text{AA} \rightleftharpoons \text{AMP} + \text{AA}$	$\Delta G = -7$
Guanidinium phosphates	 creatine phosphate	$\text{Creatine} \sim \text{P} \sim \text{P} \rightleftharpoons \text{Creatine} + \text{P}$	$\Delta G = -8$
Thioesters	 acetyl-CoA	$\text{Acetyl CoA} \rightleftharpoons \text{CoA-SH} + \text{Acetate}$	$\Delta G = -8$

The wide range of  $\Delta G$  values of high-energy bonds (see Table 3-5) means that calling a bond “high-energy” is sometimes arbitrary. The usual criterion is whether its hydrolysis can be coupled with another reaction to effect an important biosynthesis. For example, the negative  $\Delta G$  accompanying the hydrolysis of glucose-6-phosphate is 3–4 kcal/mol. But this  $\Delta G$  is not sufficient for efficient synthesis of peptide bonds, and thus this phosphate ester bond is not included among high-energy bonds.

## HIGH-ENERGY BONDS IN BIOSYNTHETIC REACTIONS

The construction of a large molecule from smaller building blocks often requires the input of free energy. Yet a biosynthetic pathway, like a degradative pathway, would not exist if it were not characterized by a net decrease in free energy. This means that many biosynthetic pathways demand an external source of free energy. These free-energy sources are the high-energy compounds. The making of many biosynthetic bonds is coupled with the breakdown of a high-energy bond, so that the net change of free energy is always negative. Thus, high-energy bonds in cells generally have a very



**FIGURE 3-14** Free-energy changes in a multistep metabolic pathway,  $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E$ . Two steps ( $A \rightarrow B$  and  $C \rightarrow D$ ) do not favor the  $A \rightarrow E$  direction of the reaction, because they have small positive  $\Delta G$  values. However, they are insignificant owing to the very large negative  $\Delta G$  values provided in steps  $B \rightarrow C$  and  $D \rightarrow E$ . Therefore, the overall reaction favors the  $A \rightarrow E$  conversion.

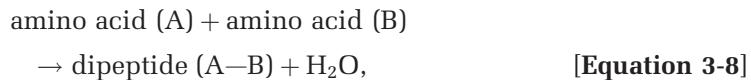
short life. Almost as soon as they are formed during a degradative reaction, they are enzymatically broken down to yield the energy needed to drive another reaction to completion.

Not all of the steps in a biosynthetic pathway require the breakdown of a high-energy bond. Often, only one or two steps involve such a bond. Sometimes this is because the  $\Delta G$ , even in the absence of an externally added high-energy bond, favors the biosynthetic direction. In other cases,  $\Delta G$  is effectively zero or may even be slightly positive. These small positive  $\Delta G$  values, however, are not significant so long as they are followed by a reaction characterized by the hydrolysis of a high-energy bond. Rather, it is the *sum* of all of the free-energy changes in a pathway that is significant, as shown in Figure 3-14. It does not really matter that the  $K_{eq}$  of a specific biosynthetic step is slightly (80:20) in favor of degradation if the  $K_{eq}$  of the succeeding step is 100:1 in favor of the forward biosynthetic direction.

Likewise, not all of the steps in a degradative pathway generate high-energy bonds. For example, only two steps in the lengthy glycolytic (Embden–Meyerhof) breakdown of glucose generate ATP. Moreover, there are many degradative pathways that have one or more steps requiring the breakdown of a high-energy bond. The glycolytic breakdown of glucose is again an example. It uses up two molecules of ATP for every four that it generates. Here, of course, as in every energy-yielding degradative process, more high-energy bonds must be made than consumed.

### Peptide Bonds Hydrolyze Spontaneously

The formation of a dipeptide and a water molecule from two amino acids requires a  $\Delta G$  of 1–4 kcal/mol, depending on which amino acids are being joined. These positive  $\Delta G$  values by themselves tell us that polypeptide chains cannot form from free amino acids. In addition, we must take into account the fact that water molecules have a much, much higher concentration than any other cellular molecules (generally more than 100 times higher). All equilibrium reactions in which water participates are thus strongly pushed in the direction that consumes water molecules. This is easily seen in the definition of equilibrium constants. For example, the reaction forming a dipeptide,



has the equilibrium constant

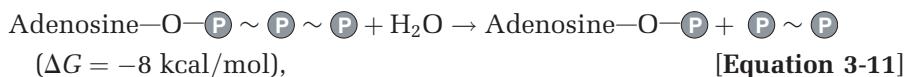
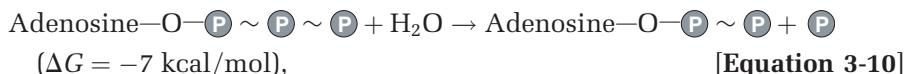
$$K_{eq} = \frac{\text{conc}^{A-B} \times \text{conc}^{\text{H}_2\text{O}}}{\text{conc}^A \times \text{conc}^B}, \quad [\text{Equation 3-9}]$$

where concentrations are given in moles per liter. Thus, for a given  $K_{eq}$  value [related to  $\Delta G$  by the formula  $\Delta G = -RT(\ln K_{eq})$ ], a much greater concentration of water means a correspondingly smaller concentration of the dipeptide. The relative concentrations are, therefore, very important. In fact, a simple calculation shows that hydrolysis may often proceed spontaneously even when the  $\Delta G$  for the nonhydrolytic reaction is –3 kcal/mol.

Thus, in theory, proteins are unstable and, given sufficient time, will spontaneously degrade to free amino acids. On the other hand, in the absence of specific enzymes, these spontaneous rates are too slow to have a significant effect on cellular metabolism. That is, once a protein is made, it remains stable unless its degradation is catalyzed by a specific enzyme.

### Coupling of Negative with Positive $\Delta G$

Free energy must be added to amino acids before they can be united to form proteins. How this happens became clear with the discovery of the fundamental role of ATP as an energy donor. ATP contains three phosphate groups attached to an adenosine molecule (adenosine—O—P~P~P). When one or two of the terminal ~P groups are broken off by hydrolysis, there is a significant decrease of free energy:

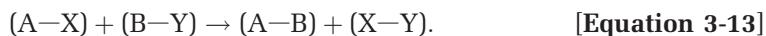


All of these breakdown reactions have negative  $\Delta G$  values considerably greater in absolute value (numerical value without regard to sign) than the positive  $\Delta G$  values accompanying the formation of polymeric molecules from their monomeric building blocks. The essential trick underlying these biosynthetic reactions, which by themselves have a positive  $\Delta G$ , is that they are coupled with the breakage of high-energy bonds, characterized by a negative  $\Delta G$  of greater absolute value. Thus, during protein synthesis, the formation of each peptide bond ( $\Delta G = +0.5 \text{ kcal/mol}$ ) is coupled with the breakdown of ATP to AMP and pyrophosphate, which has a  $\Delta G$  of  $-8 \text{ kcal/mol}$  (see Equation 3-11). This results in a net  $\Delta G$  of  $-7.5 \text{ kcal/mol}$ , more than sufficient to ensure that the equilibrium favors protein synthesis rather than breakdown.

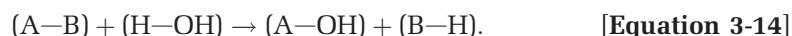
### ACTIVATION OF PRECURSORS IN GROUP TRANSFER REACTIONS

---

When ATP is hydrolyzed to ADP and phosphate, most of the free energy is liberated as heat. Because heat energy cannot be used to make covalent bonds, a coupled reaction cannot be the result of two completely separate reactions, one with a positive  $\Delta G$ , the other with a negative  $\Delta G$ . Instead, a coupled reaction is achieved by two or more successive reactions. These are always **group-transfer** reactions: reactions, not involving oxidations or reductions, in which molecules exchange functional groups. The enzymes that catalyze these reactions are called **transferases**. Consider the reaction



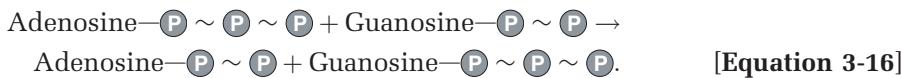
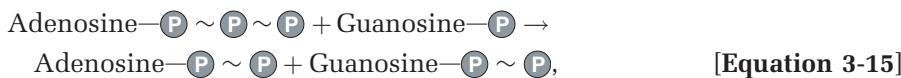
In this example, group X is exchanged with component B. Group-transfer reactions are arbitrarily defined to exclude water as a participant. When water is involved,



This reaction is called a hydrolysis, and the enzymes involved are called **hydrolases**.

The group-transfer reactions that interest us here are those involving groups attached by high-energy bonds. When such a high-energy group is transferred to an appropriate acceptor molecule, it becomes attached to

the acceptor by a high-energy bond. Group transfer thus allows the transfer of high-energy bonds from one molecule to another. For example, Equations 3-15 and 3-16 show how energy present in ATP is transferred to form GTP, one of the precursors used in RNA synthesis:



The high-energy  $\text{P} \sim \text{P}$  group on GTP allows it to unite spontaneously with another molecule. GTP is thus an example of what is called an **activated molecule**; correspondingly, the process of transferring a high-energy group is called **group activation**.

### ATP Versatility in Group Transfer

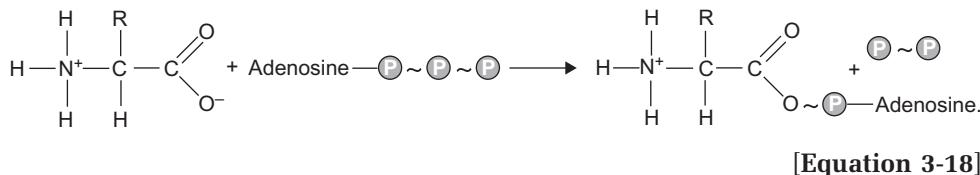
ATP synthesis has a key role in the controlled trapping of the energy of molecules that serve as energy donors. In both oxidative and photosynthetic phosphorylations, energy is used to synthesize ATP from ADP and phosphate:



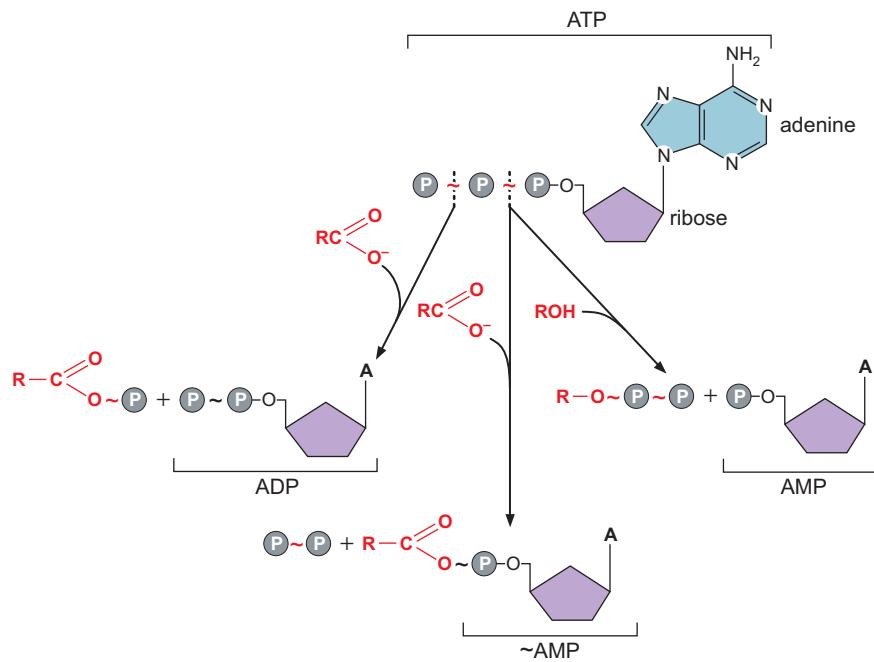
Because ATP is the original biological recipient of high-energy groups, it must be the starting point of a variety of reactions in which high-energy groups are transferred to low-energy molecules to give them the potential to react spontaneously. This central role of ATP relies on the fact that it contains two high-energy bonds whose splitting releases specific groups. This function is seen in Figure 3-15, which shows three important groups arising from ATP:  $\text{P} \sim \text{P}$ , a pyrophosphate group;  $\sim \text{AMP}$ , an adenosyl monophosphate group; and  $\sim \text{P}$ , a phosphate group. It is important to notice that these high-energy groups retain their high-energy quality only when transferred to an appropriate acceptor molecule. For example, although the transfer of a  $\sim \text{P}$  group to a  $\text{COO}^-$  group yields a high-energy  $\text{COO} \sim \text{P}$  acylphosphate group, the transfer of the same group to a sugar hydroxyl group ( $-\text{C}-\text{OH}$ ), as in the formation of glucose-6-phosphate, gives rise to a low-energy bond (<5 kcal/mol decrease in  $\Delta G$  upon hydrolysis).

### Activation of Amino Acids by Attachment of AMP

The activation of an amino acid is achieved by transfer of an AMP group from ATP to the  $\text{COO}^-$  group of the amino acid, as shown by



(In the equation, R represents the specific side group of the amino acid.) The enzymes that catalyze this type of reaction are called **aminoacyl synthetases**. Upon activation, an amino acid (AA) is thermodynamically capable

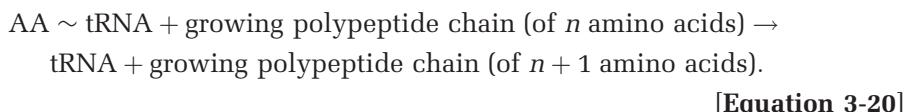


**FIGURE 3-15** Important group transfers involving ATP.

of being efficiently used for protein synthesis. Nonetheless, the AA~AMP complexes are not the direct precursors of proteins. Instead, for a reason we shall explain in Chapter 15, a second group transfer must occur to transfer the amino acid, still activated at its carboxyl group, to the end of a tRNA molecule:



A peptide bond then forms by the condensation of the AA~tRNA molecule onto the end of a growing polypeptide chain:



Thus, the final step of this “coupled reaction,” like that of all other coupled reactions, necessarily involves the removal of the activating group and the conversion of a high-energy bond into one with a lower free energy of hydrolysis. This is the source of the negative  $\Delta G$  that drives the reaction in the direction of protein synthesis.

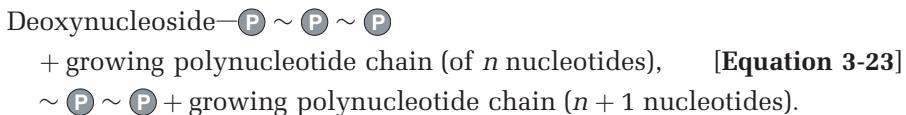
### Nucleic Acid Precursors Are Activated by the Presence of $\text{P} \sim \text{P}$

Both types of nucleic acid, DNA and RNA, are built up from mononucleotide monomers, also called **nucleoside phosphates**. Mononucleotides, however, are thermodynamically even less likely to combine than amino acids. This is because the phosphodiester bonds that link the former together release considerable free energy upon hydrolysis ( $-6 \text{ kcal/mol}$ ). This means that nucleic acids will spontaneously hydrolyze, at a slow rate, to mononucleotides. Thus, it is even more important that activated precursors be used in the synthesis of nucleic acids than in the synthesis of proteins.

The immediate precursors for both DNA and RNA are the nucleoside-5'-triphosphates. For DNA, these precursors are dATP, dGTP, dCTP, and dTTP (d stands for deoxy); for RNA, the precursors are ATP, GTP, CTP, and UTP. ATP, thus, not only serves as the main source of high-energy groups in group-transfer reactions, but is itself a direct precursor for RNA. The other three RNA precursors all arise by group-transfer reactions like those described in Equations 3-15 and 3-16. The deoxytriphosphates are formed in basically the same way: After the deoxymononucleotides have been synthesized, they are transformed to the triphosphate form by group transfer from ATP:



These triphosphates can then unite to form polynucleotides held together by phosphodiester bonds. In this group-transfer reaction, a pyrophosphate bond is broken and a pyrophosphate group is released:



This reaction, unlike that which forms peptide bonds, does not have a negative  $\Delta G$ . In fact, the  $\Delta G$  is slightly positive ( $\sim 0.5$  kcal/mol). This situation immediately poses the question, as polynucleotides obviously form: What is the source of the necessary free energy?

### The Value of $\text{P} \sim \text{P}$ Release in Nucleic Acid Synthesis

The needed free energy comes from the splitting of the high-energy pyrophosphate group that is formed simultaneously with the high-energy phosphodiester bond. All cells contain a powerful enzyme, pyrophosphatase, which breaks down pyrophosphate molecules almost as soon as they are formed:



The large negative  $\Delta G$  means that the reaction is effectively irreversible. This means that once  $\text{P} \sim \text{P}$  is broken down, it never re-forms.

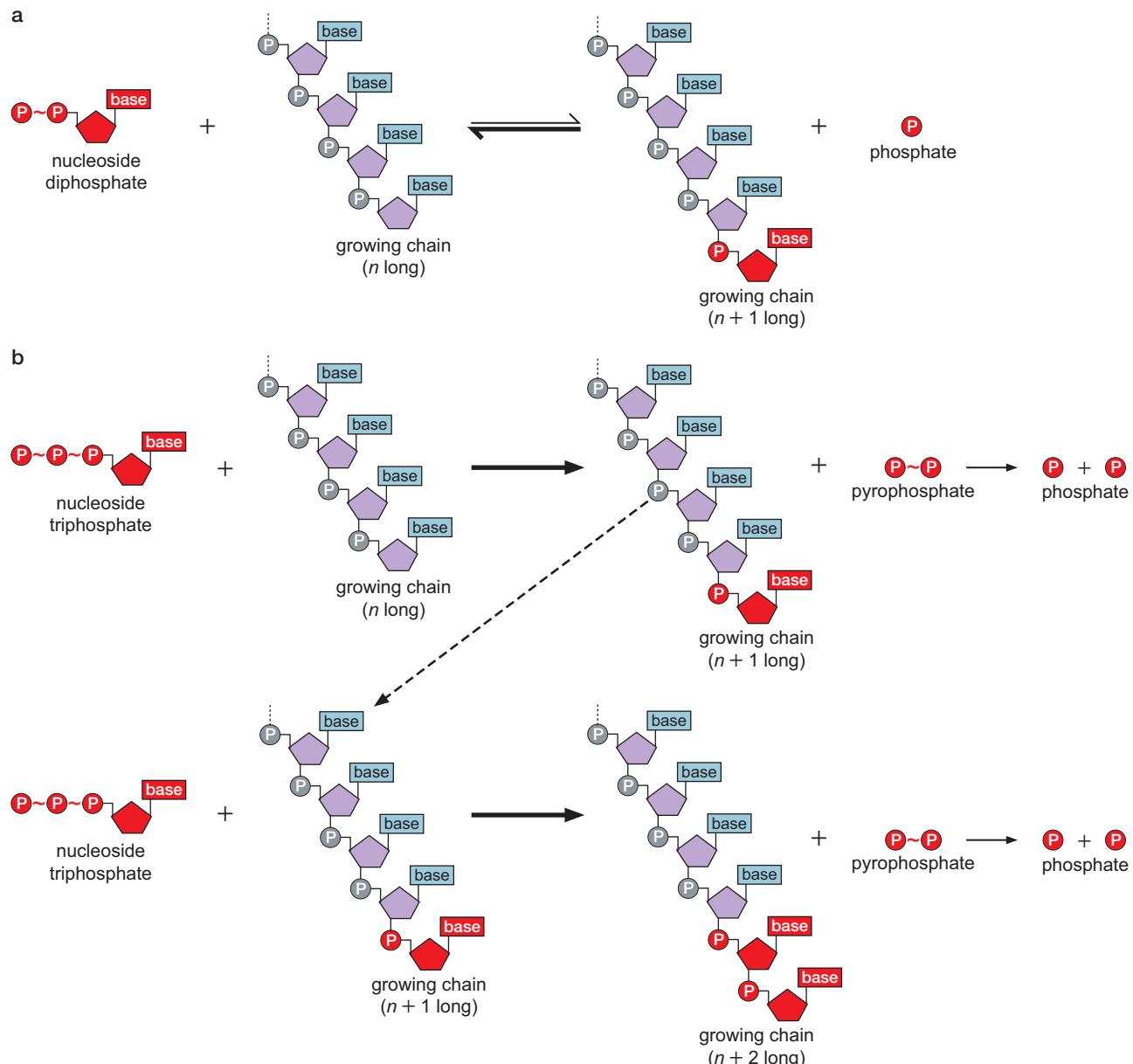
The union of the nucleoside monophosphate group (Equation 3-21), coupled with the splitting of the pyrophosphate groups (Equation 3-24), has an equilibrium constant determined by the combined  $\Delta G$  values of the two reactions: (0.5 kcal/mol) + (-7 kcal/mol). The resulting value ( $\Delta G = -6.5$  kcal/mol) tells us that nucleic acids almost never break down to re-form their nucleoside triphosphate precursors.

Here we see a powerful example of the fact that often it is the free-energy change accompanying a *group of reactions* that determines whether a reaction in the group will take place. Reactions with small, positive  $\Delta G$  values, which by themselves would never take place, are often part of important metabolic pathways in which they are followed by reactions with large negative  $\Delta G$  values. At all times we must remember that a single reaction (or even a single pathway) never occurs in isolation; rather, the nature of the equilibrium is constantly being changed through the addition and removal of metabolites.

### $\text{P} \sim \text{P}$ Splits Characterize Most Biosynthetic Reactions

The synthesis of nucleic acids is not the only reaction in which direction is determined by the release and splitting of  $\text{P} \sim \text{P}$ . In fact, essentially all biosynthetic reactions are characterized by one or more steps that release pyrophosphate groups. Consider, for example, the activation of an amino acid by the attachment of AMP. By itself, the transfer of a high-energy bond from ATP to the AA~AMP complex has a slightly positive  $\Delta G$ . Therefore, it is the release and splitting of ATP's terminal pyrophosphate group that provides the negative  $\Delta G$  that is necessary to drive the reaction.

The great utility of the pyrophosphate split is neatly shown when we consider the problems that would arise if a cell attempted to synthesize nucleic acid from nucleoside diphosphates rather than triphosphates (Fig. 3-16).



**FIGURE 3-16** Two scenarios for nucleic acid biosynthesis. (a) Synthesis of nucleic acids using nucleoside diphosphates. (b) Synthesis of nucleic acids using nucleoside triphosphates.

Phosphate, rather than pyrophosphate, would be liberated as the backbone phosphodiester linkages were made. The phosphodiester linkages, however, are not stable in the presence of significant quantities of phosphate, because they are formed without a significant release of free energy. Thus, the biosynthetic reaction would be easily reversible; if phosphate were to accumulate, the reaction would begin to move in the direction of nucleic acid breakdown according to the law of mass action. Moreover, it is not feasible for a cell to remove the phosphate groups as soon as they are generated (thereby preventing this reverse reaction), because all cells require a significant internal level of phosphate to grow. In contrast, a sequence of reactions that liberate pyrophosphate and then rapidly break it down into two phosphates disconnects the liberation of phosphate from the nucleic acid biosynthesis reaction and thereby prevents the possibility of reversing the biosynthetic reaction (see Fig. 3-16). In consequence, it would be very difficult to accumulate enough phosphate in the cell to drive both reactions in the reverse, or breakdown, direction. It is clear that the use of nucleoside triphosphates as precursors of nucleic acids is not a matter of chance.

This same type of argument tells us why ATP, and not ADP, is the key donor of high-energy groups in all cells. At first, this preference seemed arbitrary to biochemists. Now, however, we see that many reactions using ADP as an energy donor would occur equally well in both directions.

## SUMMARY

---

Many important chemical events in cells do not involve the making or breaking of covalent bonds. The cellular location of most molecules depends on weak, or secondary, attractive or repulsive forces. In addition, weak bonds are important in determining the shape of many molecules, especially very large ones. The most important of these weak forces are hydrogen bonds, van der Waals interactions, hydrophobic bonds, and ionic bonds. Even though these forces are relatively weak, they are still large enough to ensure that the right molecules (or atomic groups) interact with each other. For example, the surface of an enzyme is uniquely shaped to allow the specific attraction of its substrates.

The formation of all chemical bonds—weak interactions as well as strong covalent bonds—proceeds according to the laws of thermodynamics. A bond tends to form when the result would be a release of free energy (negative  $\Delta G$ ). For the bond to be broken, this same amount of free energy must be supplied. Because the formation of covalent bonds between atoms usually involves a very large negative  $\Delta G$ , covalently bound atoms almost never separate spontaneously. In contrast, the  $\Delta G$  values accompanying the formation of weak bonds are only several times larger than the average thermal energy of molecules at physiological temperatures. Single weak bonds are thus frequently being made and broken in living cells.

Molecules having polar (charged) groups interact quite differently from nonpolar molecules (in which the charge is symmetrically distributed). Polar molecules can form good hydrogen bonds, whereas nonpolar molecules can form only van der Waals bonds. The most important polar molecule is water. Each water molecule can form four hydrogen bonds to other water molecules. Although polar molecules

tend to be soluble in water (to various degrees), nonpolar molecules are insoluble because they cannot form hydrogen bonds with water molecules.

Every distinct molecule has a unique molecular shape that restricts the number of molecules with which it can form strong secondary bonds. Strong secondary interactions demand both a complementary (lock-and-key) relationship between the two bonding surfaces and the involvement of many atoms. Although molecules bound together by only one or two secondary bonds frequently fall apart, a collection of these weak bonds can result in a stable aggregate. The fact that double-helical DNA never falls apart spontaneously shows the extreme stability possible in such an aggregate.

The biosynthesis of many molecules appears, at a superficial glance, to violate the thermodynamic law that spontaneous reactions always involve a decrease in free energy ( $\Delta G$  is negative). For example, the formation of proteins from amino acids has a positive  $\Delta G$ . This paradox is removed when we realize that the biosynthetic reactions do not proceed as initially postulated. Proteins, for example, are not formed from free amino acids. Instead, the precursors are first enzymatically converted to high-energy activated molecules, which, in the presence of a specific enzyme, spontaneously unite to form the desired biosynthetic product.

Many biosynthetic processes are thus the result of “coupled” reactions, the first of which supplies the energy that allows the spontaneous occurrence of the second reaction. The primary energy source in cells is ATP. It is formed from ADP and inorganic phosphate, either during degradative reactions (such as fermentation or respiration) or during photosynthesis. ATP contains several high-energy bonds whose hydrolysis has a large negative  $\Delta G$ . Groups linked by

high-energy bonds are called *high-energy groups*. High-energy groups can be transferred to other molecules by group-transfer reactions, thereby creating new high-energy compounds. These derivative high-energy molecules are then the immediate precursors for many biosynthetic steps.

Amino acids are activated by the addition of an AMP group, originating from ATP, to form an AA~AMP molecule. The energy of the high-energy bond in the AA~AMP molecule is similar to that of a high-energy bond of ATP. Nonetheless, the group-transfer reaction proceeds to completion because the high-energy  $\text{P} \sim \text{P}$  molecule, created when the AA~

AMP molecule is formed, is broken down by the enzyme pyrophosphatase to low-energy groups. Thus, the reverse reaction,  $\text{P} \sim \text{P} + \text{AA} \sim \text{AMP} \rightarrow \text{ATP} + \text{AA}$ , cannot occur.

Almost all biosynthetic reactions result in the release of  $\text{P} \sim \text{P}$ . Almost as soon as it is made, it is enzymatically broken down to two phosphate molecules, thereby making a reversal of the biosynthetic reaction impossible. The great utility of the  $\text{P} \sim \text{P}$  split provides an explanation for why ATP, not ADP, is the primary energy donor. ADP cannot transfer a high-energy group and at the same time produce  $\text{P} \sim \text{P}$  groups as a by-product.

## BIBLIOGRAPHY

### Weak Chemical Interactions

- Branden C. and Tooze J. 1999. *Introduction to protein structure*, 2nd ed. Garland Publishing, New York.
- Creighton T.E. 1992. *Proteins: Structure and molecular properties*, 2nd ed. W.H. Freeman, New York.
- . 1983. *Proteins*. W.H. Freeman, San Francisco.
- Donohue J. 1968. Selected topics in hydrogen bonding. In *Structural chemistry and molecular biology* (ed. A. Rich and N. Davidson), pp. 443–465. W.H. Freeman, San Francisco.
- Fersht A. 1999. *Structure and mechanism in protein science: A guide to enzyme catalysis and protein folding*. W.H. Freeman, New York.
- Gray H.B. 1964. *Electrons and chemical bonding*. Benjamin Cummings, Menlo Park, California.
- Klotz I.M. 1967. *Energy changes in biochemical reactions*. Academic Press, New York.
- Kyte J. 1995. *Mechanism in protein chemistry*. Garland Publishing, New York.
- . 1995. *Structure in protein chemistry*. Garland Publishing, New York.
- Lehninger A.L. 1971. *Bioenergetics*, 3rd ed. Benjamin Cummings, Menlo Park, California.
- Lesk A. 2000. *Introduction to protein architecture: The structural biology of proteins*. Oxford University Press, New York.
- Marsh R.E. 1968. Some comments on hydrogen bonding in purine and pyrimidine bases. In *Structural chemistry and molecular biology* (eds. A. Rich and N. Davidson), pp. 485–489. W.H. Freeman, San Francisco.
- Morowitz H.J. 1970. *Entropy for biologists*. Academic Press, New York.
- Pauling L. 1960. *The nature of the chemical bond*, 3rd ed. Cornell University Press, Ithaca, New York.

- Tinoco I., Sauer K., Wang J.C., Puglisi J.D., and Wang J.Z. 2001. *Physical chemistry: Principles and applications in life sciences*, 4th ed. Prentice Hall College Division, Upper Saddle River, New Jersey.

### Strong Chemical Bonds

- Berg J., Tymoczko J.L., and Stryer L. 2006. *Biochemistry*, 6th ed. W.H. Freeman, New York.
- Kornberg A. 1962. On the metabolic significance of phosphorolytic and pyrophosphorolytic reactions. In *Horizons in biochemistry* (eds. M. Kasha and B. Pullman), pp. 251–264. Academic Press, New York.
- Krebs H.A. and Kornberg H.L. 1957. A survey of the energy transformation in living material. *Ergeb. Physiol. Biol. Exp. Pharmakol.* **49**: 212.
- Nelson D.L. and Cox M.M. 2000. *Lehninger principles of biochemistry*, 3rd ed. Worth Publishing, New York.
- Nicholls D.G. and Ferguson S.J. 2002. *Bioenergetics 3*. Academic Press, San Diego.
- Purich D.L., ed. 2002. *Enzyme kinetics and mechanism, Part F: Detection and characterization of enzyme reaction intermediates*. Methods in Enzymology, Vol. 354. Academic Press, San Diego.
- Silverman R.B. 2002. *The organic chemistry of enzyme-catalyzed reactions*. Academic Press, San Diego.
- Tinoco I., Sauer K., Wang J.C., Puglisi J.D., and Wang J.Z. 2001. *Physical chemistry: Principles and applications in life sciences*, 4th ed. Prentice Hall College Division, Upper Saddle River, New Jersey.
- Voet D., Voet J.G., and Pratt C. 2002. *Fundamentals of biochemistry*. John Wiley & Sons, New York.

## QUESTIONS

### MasteringBiology®

*For instructor-assigned tutorials and problems, go to MasteringBiology.*

For answers to even-numbered questions, see Appendix 2: Answers.

**Question 1.** What are the types of bonds that are possible between two macromolecules?

**Question 2.** (True or False. If false, rewrite the statement to be true.) Enzymes lower the  $\Delta G$  of a reaction.

**Question 3.** (True or False. If false, rewrite the statement to be true.) Ionic bonds and hydrogen bonds are stronger than van der Waals bonds.

**Question 4.** (True or False. If false, rewrite the statement to be true.) At 25°C, a 10-fold change in  $K_{\text{eq}}$  corresponds to a 10-fold change in  $\Delta G$ .

**Question 5.** Review Table 3-5. Which major cellular processes involve the reactions of a nucleoside triphosphate breaking down into a nucleoside monophosphate and pyrophosphate as well as pyrophosphate breaking down into two phosphates? Why is the  $\Delta G$  of these reactions significant for these processes?

**Question 6.** What is the primary type of bond responsible for each of the following interactions:

- A. One DNA strand interacting with another strand of DNA in double-stranded DNA.
- B. A dipeptide of two amino acids.

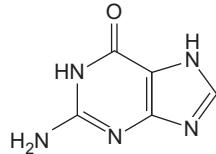
**Question 7.** Describe the general structure of water molecules at a temperature below freezing versus at  $25^{\circ}\text{C}$ , and name the primary type of bond between water molecules.

**Question 8.** Define polar and nonpolar molecules in terms of dipole moments. Do van der Waals interactions occur between polar or nonpolar molecules?

**Question 9.** Calculate the value of  $K_{\text{eq}}$  at  $25^{\circ}\text{C}$ , given the  $\Delta G$  of  $-12 \text{ kcal/mol}$  as for the hydrolysis of PEP into pyruvate and a phosphate.

**Question 10.** Given the equation  $\text{AB} + \text{energy} \rightleftharpoons \text{A} + \text{B}$ , calculate the concentration of A at equilibrium if  $K_{\text{eq}}=8.0 \times 10^5 \text{ mM}$ ,  $[\text{B}]=2 \text{ mM}$ , and  $[\text{AB}]=0.5 \text{ mM}$  (where [x] means “concentration of x”).

**Question 11.** The structure of a nitrogenous base is shown below. Considering this structure alone (not in the context of DNA or RNA), how many possible hydrogen bond acceptors are present? How many possible hydrogen bond donors are present?



**Question 12.**  $\text{Glutamate} + \text{NH}_3 \rightleftharpoons \text{glutamine} + \text{H}_2\text{O} \quad \Delta G = +3.4 \text{ kcal/mol}$

Would coupling this reaction to ATP hydrolysis allow glutamine formation to be favored? Explain why or why not. Write the overall reaction.

**Question 13.** Explain why nucleoside triphosphates rather than nucleoside diphosphates are used in DNA synthesis.

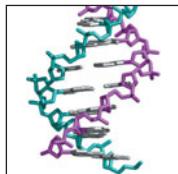
**Question 14.** Researchers studied the interactions between proteins and DNA in more than 100 protein–DNA complexes. The table below provides a subset of the data: the distribution of single hydrogen bonds between a specific base and amino acid.

Amino Acids	DNA Bases			
	Thymine	Cytosine	Adenine	Guanine
Arginine	5	4	7	26
Glutamate	—	11	—	1
Tryptophan	—	—	—	—

- A. Explain why the researchers found no single hydrogen bonds between tryptophan and one of the DNA bases.
- B. Explain why the researchers found single hydrogen bonds between glutamate and each of the DNA bases.
- C. Is there a preference of arginine for one of the amino acids?

Data adapted from Luscombe et al. (2001. *Nucleic Acids Res.* **29**: 2860–2874).

CHAPTER 4



# The Structure of DNA

THE DISCOVERY THAT DNA IS THE PRIME GENETIC molecule, carrying all of the hereditary information within chromosomes, immediately focused attention on its structure. It was hoped that knowledge of the structure would reveal how DNA carries the genetic messages that are replicated when chromosomes divide to produce two identical copies of themselves. During the late 1940s and early 1950s, several research groups in the United States and in Europe engaged in serious efforts—both cooperative and rival—to understand how the atoms of DNA are linked together by covalent bonds and how the resulting molecules are arranged in three-dimensional space. Not surprisingly, there initially were fears that DNA might have very complicated and perhaps bizarre structures that differed radically from one gene to another. Great relief, if not general elation, was thus expressed when the fundamental DNA structure was found to be the double helix. This told us that all genes have roughly the same three-dimensional form and that the differences between two genes reside in the order and number of their four nucleotide building blocks along the complementary strands.

Now, some 50 years after the discovery of the double helix, this simple description of the genetic material remains true and has not had to be appreciably altered to accommodate new findings. Nevertheless, we have come to realize that the structure of DNA is not quite as uniform as was first thought. For example, the chromosomes of some small viruses have single-stranded, not double-stranded, molecules. Moreover, the precise orientation of the base pairs varies slightly from base pair to base pair in a manner that is influenced by the local DNA sequence. Some DNA sequences even permit the double helix to twist in the left-handed sense, as opposed to the right-handed sense originally formulated for DNA's general structure. And some DNA molecules are linear, whereas others are circular. Still additional complexity comes from the supercoiling (further twisting) of the double helix, often around cores of DNA-binding proteins. Clearly, the structure of DNA is richer and more intricate than was at first appreciated. Indeed, there is no one generic structure for DNA. As we see in this chapter, there are, in fact, variations on common themes of structure that arise from the unique physical, chemical, and topological properties of the polynucleotide chain.

## O U T L I N E

DNA Structure, 78



DNA Topology, 93



Visit Web Content for Structural  
Tutorials and Interactive Animations

## DNA STRUCTURE

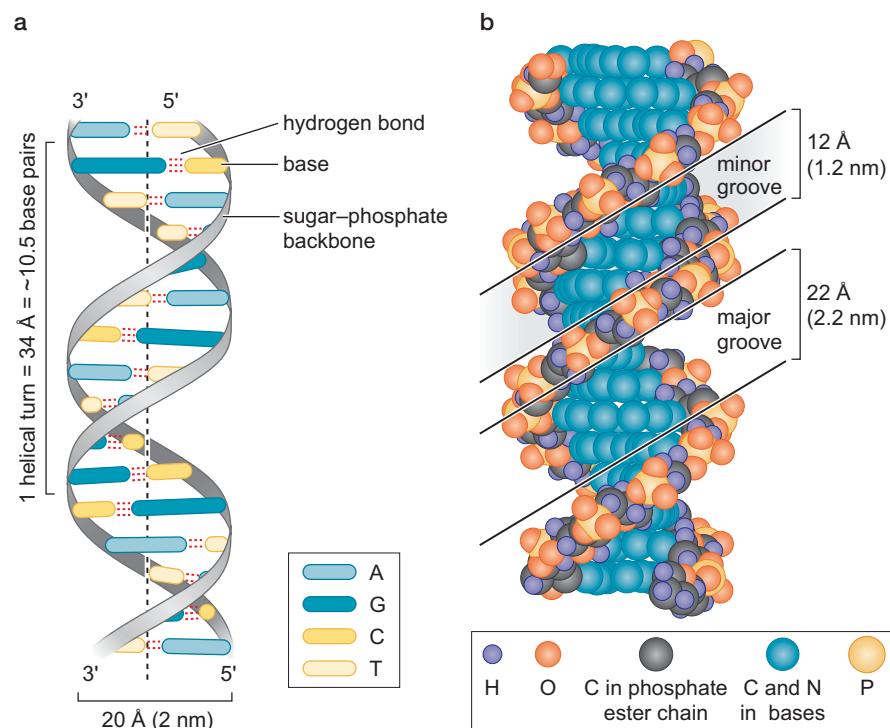
### DNA Is Composed of Polynucleotide Chains



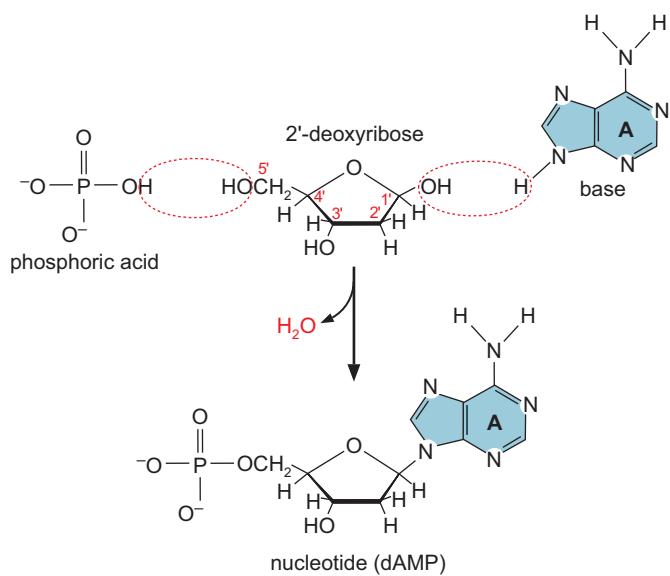
The most important feature of DNA is that it is usually composed of two **polynucleotide chains** twisted around each other in the form of a double helix (Fig. 4-1; see Structural Tutorial 4-1). Figure 4-1a presents the structure of the double helix in a schematic form. Note that if inverted 180° (e.g., by turning this book upside down), the double helix looks superficially the same, because of the complementary nature of the two DNA strands. The space-filling model of the double helix in Figure 4-1b shows the components of the DNA molecule and their relative positions in the helical structure. The backbone of each strand of the helix is composed of alternating sugar and phosphate residues; the bases project inward but are accessible through the major and minor grooves.

Let us begin by considering the nature of the nucleotide, the fundamental building block of DNA. The **nucleotide** consists of a phosphate joined to a sugar, known as **2'-deoxyribose**, to which a base is attached. The phosphate and the sugar have the structures shown in Figure 4-2. The sugar is called 2'-deoxyribose because there is no hydroxyl at position 2' (just two hydrogens). Note that the positions on the sugar are designated with primes to distinguish them from positions on the bases (see the discussion below).

We can think of how the base is joined to 2'-deoxyribose by imagining the removal of a molecule of water between the hydroxyl on the 1' carbon of the sugar and the base to form a glycosidic bond (Fig. 4-2). The sugar and base alone are called a **nucleoside**. Likewise, we can imagine linking the phosphate to 2'-deoxyribose by removing a water molecule from between the phosphate and the hydroxyl on the 5' carbon to make a 5' phosphomonoester. Adding a phosphate (or more than one phosphate) to a **nucleoside** creates a **nucleotide**. Thus, by making a glycosidic bond between the base



**FIGURE 4-1** The helical structure of DNA. (a) Schematic model of the double helix. One turn of the helix (34 Å or 3.4 nm) spans ~10.5 bp. (b) Space-filling model of the double helix. The sugar and phosphate residues in each strand form the backbone, which is traced by the yellow, gray, and red circles, showing the helical twist of the overall molecule. The bases project inward but are accessible through major and minor grooves.



**FIGURE 4-2** Formation of nucleotide by removal of water. The numbers of the carbon atoms in 2'-deoxyribose are labeled in red.

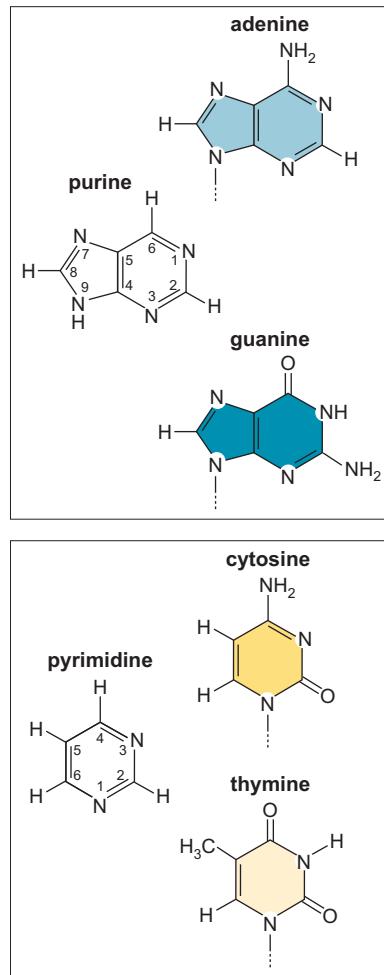
and the sugar, and by making a phosphoester bond between the sugar and the phosphoric acid, we have created a nucleotide (Table 4-1).

Nucleotides are, in turn, joined to each other in polynucleotide chains through the 3'-hydroxyl of 2'-deoxyribose of one nucleotide and the phosphate attached to the 5'-hydroxyl of another nucleotide (Fig. 4-3). This is a **phosphodiester linkage** in which the phosphoryl group between the two nucleotides has one sugar esterified to it through a 3'-hydroxyl and a second sugar esterified to it through a 5'-hydroxyl. Phosphodiester linkages create the repeating, sugar–phosphate backbone of the polynucleotide chain, which is a regular feature of DNA. In contrast, the order of the bases along the polynucleotide chain is irregular. This irregularity as well as the long length is, as we shall see, the basis for the enormous information content of DNA.

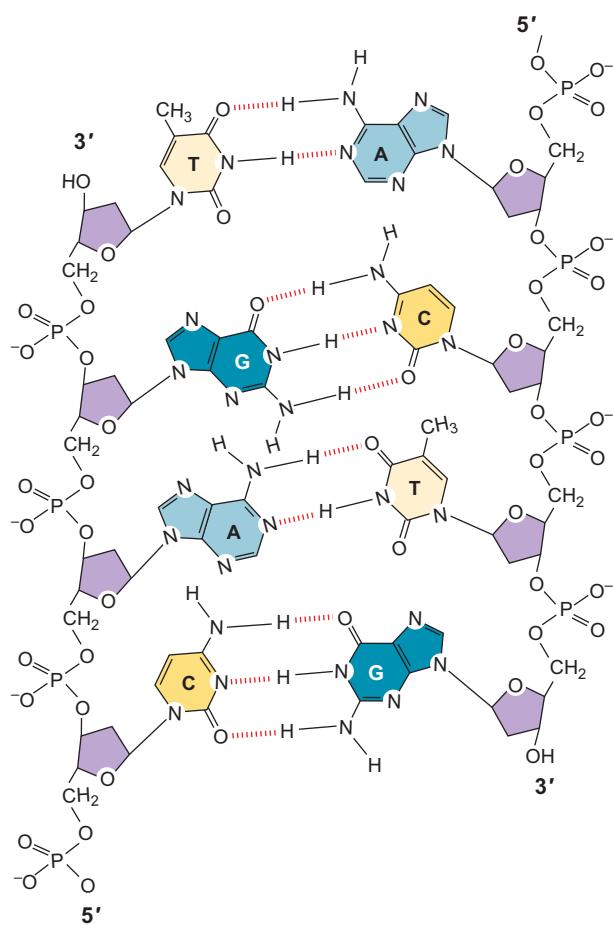
The phosphodiester linkages impart an inherent polarity to the DNA chain. This polarity is defined by the asymmetry of the nucleotides and the way they are joined. DNA chains have a free 5'-phosphate or 5'-hydroxyl at one end and a free 3'-phosphate or 3'-hydroxyl at the other end. The convention is to write DNA sequences from the 5' end (on the left) to the 3' end, generally with a 5'-phosphate and a 3'-hydroxyl.

**TABLE 4-1** Adenine and Related Compounds

	Base Adenine	Nucleotide 2'-Deoxyadenosine	Nucleoside 2'-Deoxyadenosine 5'-Phosphate
Structure			
Molecular weight	135.1	251.2	331.2



**FIGURE 4-4** Purines and pyrimidines. The dotted lines indicate the sites of attachment of the bases to the sugars. For simplicity, hydrogens are omitted from the sugars and bases in subsequent figures, except where pertinent to the illustration.

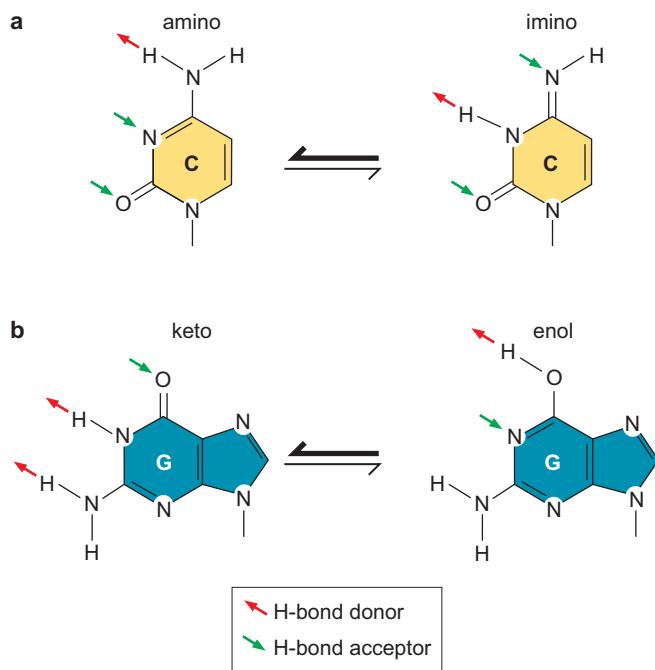


**FIGURE 4-3** Detailed structure of polynucleotide polymer. The structure shows base pairing between purines (blue) and pyrimidines (yellow), and the phosphodiester linkages of the backbone. (Adapted from Dickerson R.E. 1983. *Sci. Am.* **249**: 94. Illustration, Irving Geis. Image from Irving Geis Collection/Howard Hughes Medical Institution. Not to be reproduced without permission.)

### Each Base Has Its Preferred Tautomeric Form

The bases in DNA are flat, heterocyclic rings, consisting of carbon and nitrogen atoms. The bases fall into two classes, **purines** and **pyrimidines**. The purines are **adenine** and **guanine**, and the pyrimidines are **cytosine** and **thymine**. The purines are derived from the double-ringed structure shown in Figure 4-4. Adenine and guanine share this essential structure but with different groups attached. Likewise, cytosine and thymine are variations on the single-ringed structure shown in Figure 4-4. The figure also shows the numbering of the positions in the purine and pyrimidine rings. The bases are attached to the deoxyribose by glycosidic linkages at N1 of the pyrimidines or at N9 of the purines.

Each of the bases exists in two alternative **tautomeric states**, which are in equilibrium with each other. The equilibrium lies far to the side of the conventional structures shown in Figure 4-4, which are the predominant states and the ones important for base pairing. The nitrogen atoms attached to the purine and pyrimidine rings are in the amino form in the predominant state and only rarely assume the imino configuration. Likewise, the oxygen atoms attached to the guanine and thymine normally have the keto form and only rarely take on the enol configuration. As examples, Figure 4-5 shows tautomerization of cytosine into the imino form (Fig. 4-5a) and guanine into the



**FIGURE 4-5 Base tautomers.** Amino  $\rightleftharpoons$  imino and keto  $\rightleftharpoons$  enol tautomerism. (a) Cytosine is usually in the amino form but rarely forms the imino configuration. (b) Guanine is usually in the keto form but is rarely found in the enol configuration.

enol form (Fig. 4-5b). As we shall see, the capacity to form an alternative tautomer is a frequent source of errors during DNA synthesis.

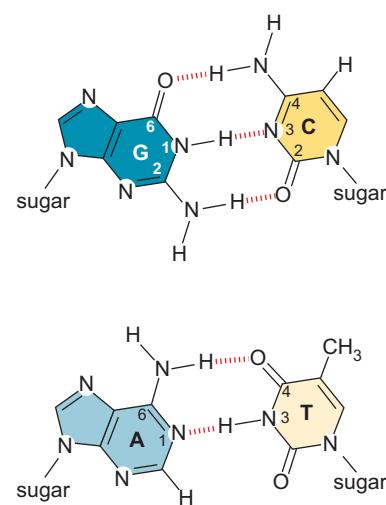
### The Two Strands of the Double Helix Are Wound around Each Other in an Antiparallel Orientation

The double helix consists of two polynucleotide chains that are aligned in opposite orientation. The two chains have the same helical geometry but have opposite 5' to 3' orientations. That is, the 5' to 3' orientation of one chain is antiparallel to the 5' to 3' orientation of the other strand, as shown in Figures 4-1 and 4-3. The two chains interact with each other by pairing between the bases, with adenine on one chain pairing with thymine on the other chain and, likewise, guanine pairing with cytosine. Thus, the base at the 5' end of one strand is paired with the base at the 3' end of the other strand. The antiparallel orientation of the double helix is a stereochemical consequence of the way that adenine and thymine, and guanine and cytosine, pair with each together.

### The Two Chains of the Double Helix Have Complementary Sequences

The pairing between adenine and thymine, and between guanine and cytosine, results in a complementary relationship between the sequence of bases on the two intertwined chains and gives DNA its self-encoding character. For example, if we have the sequence 5'-ATGTC-3' on one chain, the opposite chain must have the complementary sequence 3'-TACAG-5'.

The strictness of the rules for this “Watson–Crick” pairing derives from the complementarity both of shape and of hydrogen-bonding properties between adenine and thymine and between guanine and cytosine (Fig. 4-6). Adenine and thymine match up so that a hydrogen bond can form between the exocyclic amino group at C6 on adenine and the carbonyl at C4 in thymine; and likewise, a hydrogen bond can form between N1 of



**FIGURE 4-6 A:T and G:C base pairs.** The figure shows hydrogen bonding between the bases.

adenine and N3 of thymine. A corresponding arrangement can be drawn between a guanine and a cytosine, so that there is both hydrogen bonding and shape complementarity in this base pair as well. A G:C base pair has three hydrogen bonds, because the exocyclic NH<sub>2</sub> at C2 on guanine lies opposite to, and can hydrogen-bond with, a carbonyl at C2 on cytosine. Likewise, a hydrogen bond can form between N1 of guanine and N3 of cytosine and between the carbonyl at C6 of guanine and the exocyclic NH<sub>2</sub> at C4 of cytosine. Watson–Crick base pairing requires that the bases be in their preferred tautomeric states.

An important feature of the double helix is that the two base pairs have exactly the same geometry; having an A:T base pair or a G:C base pair between the two sugars does not perturb the arrangement of the sugars because the distance between the sugar attachment points is the same for both base pairs. Neither does T:A or C:G. In other words, there is an approximately twofold axis of symmetry that relates the two sugars, and all four base pairs can be accommodated within the same arrangement without any distortion of the overall structure of the DNA. In addition, the base pairs can stack neatly on top of each other between the two helical sugar–phosphate backbones. Thus, the irregularity in the order of base pairs in DNA is embedded in an overall architecture that is relatively regular. This is in contrast to proteins (see Chapter 6) in which the irregular order of amino acids results in enormous diversity in protein structures.

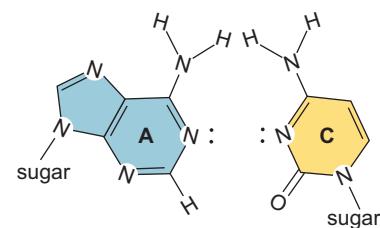
### The Double Helix Is Stabilized by Base Pairing and Base Stacking

The hydrogen bonds between complementary bases are a fundamental feature of the double helix, contributing to the thermodynamic stability of the helix and the specificity of base pairing. Hydrogen bonding might not, at first glance, appear to contribute importantly to the stability of DNA for the following reason: An organic molecule in aqueous solution has all of its hydrogen-bonding properties satisfied by water molecules that come on and off very rapidly. As a result, for every hydrogen bond that is made when a base pair forms, a hydrogen bond with water is broken that was there before the base pair formed. Thus, the net energetic contribution of hydrogen bonds to the stability of the double helix would appear to be modest. However, when polynucleotide strands are separate, water molecules are lined up on the bases. When strands come together in the double helix, the water molecules are displaced from the bases. This creates disorder and increases entropy, thereby stabilizing the double helix. Hydrogen bonds are not the only force that stabilizes the double helix.

A second important contribution comes from stacking interactions between the bases. The bases are flat, relatively water-insoluble molecules, and they tend to stack above each other roughly perpendicular to the direction of the helical axis. Electron cloud interactions ( $\pi-\pi$ ) between bases in the helical stacks contribute significantly to the stability of the double helix. The stacked bases are attracted to each other by transient, induced dipoles between the electron clouds, a phenomenon known as van der Waals interactions. Base stacking also contributes to the stability of the double helix, a hydrophobic effect. Briefly put, water molecules interact more favorably with each other than with the “greasy” or hydrophobic surfaces of the bases. These hydrophobic surfaces are buried by base stacking in the double helix (as compared with the relative lack of stacking in single-stranded DNA), minimizing the exposure of base surfaces to water molecules and hence lowering the free energy of the double helix.

## Hydrogen Bonding Is Important for the Specificity of Base Pairing

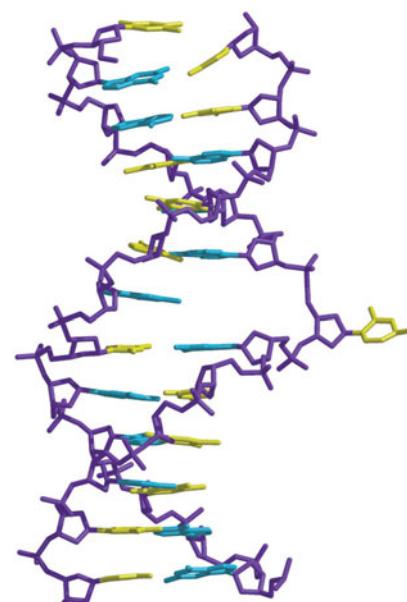
As we have seen, hydrogen bonding per se does not contribute importantly to the stability of DNA. It is, however, particularly important for the specificity of base pairing. Suppose we tried to pair an adenine with a cytosine. If so, we would have a hydrogen-bond acceptor (N1 of adenine) lying opposite a hydrogen-bond acceptor (N3 of cytosine) with no room to put a water molecule in between to satisfy the two acceptors (Fig. 4-7). Likewise, two hydrogen-bond donors, the NH<sub>2</sub> groups at C6 of adenine and C4 of cytosine, would lie opposite each other. Thus, an A:C base pair would be unstable because water would have to be stripped off the donor and acceptor groups without restoring the hydrogen bond formed within the base pair.



**FIGURE 4-7** A:C incompatibility. The structure shows the inability of adenine to form the proper hydrogen bonds with cytosine. The base pair is therefore unstable.

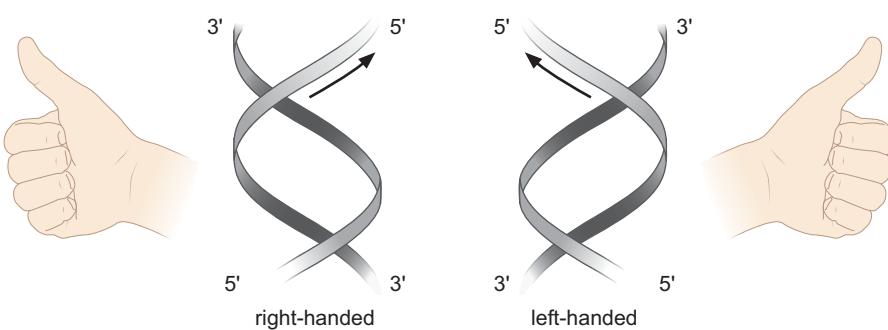
## Bases Can Flip Out from the Double Helix

As we have seen, the energetics of the double helix favor the pairing of each base on one polynucleotide strand with the complementary base on the other strand. Sometimes, however, individual bases can protrude from the double helix in a remarkable phenomenon known as **base flipping** (Fig. 4-8). As we shall see in Chapter 10, certain enzymes that methylate bases or remove damaged bases do so with the base in an extrahelical configuration in which it is flipped out from the double helix, enabling the base to sit in the catalytic cavity of the enzyme. Furthermore, enzymes involved in homologous recombination and DNA repair are believed to scan DNA for homology or lesions by flipping out one base after another. This is not energetically expensive because only one base is flipped out at a time. Clearly, DNA is more flexible than might be assumed at first glance.



## DNA Is Usually a Right-Handed Double Helix

Applying the handedness rule from physics, we can see that each of the polynucleotide chains in the double helix is right-handed. In your mind's eye, hold your right hand up to the DNA molecule in Figure 4-9 with your thumb pointing up and along the long axis of the helix and your fingers following the grooves in the helix. Trace along one strand of the helix in the direction in which your thumb is pointing. Notice that you go around the helix in the same direction as your fingers are pointing. This does not work if you use your left hand. Try it!



**FIGURE 4-9** Left- and right-handed helices. The two polynucleotide chains in the double helix wrap around one another in a right-handed manner.

**FIGURE 4-8** Base flipping. Structure of isolated DNA from the methylase structure, showing the flipped cytosine residue and the small distortions to the adjacent base pairs. (Klimasauskas S. et al. 1994. *Cell* 76: 357.) Image prepared with BobScript, MolScript, and Raster3D.

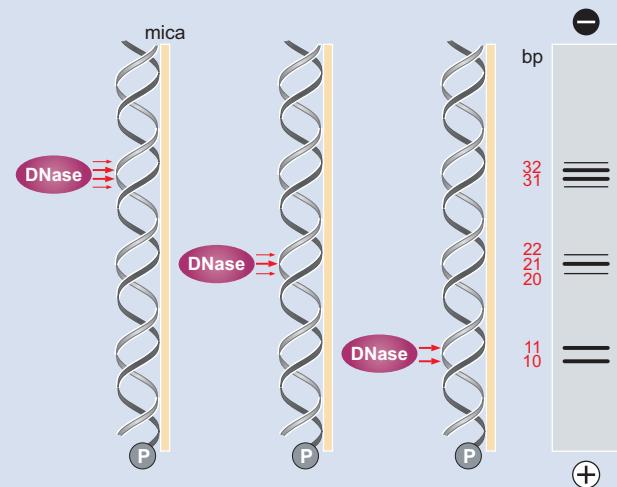
## KEY EXPERIMENTS

### Box 4-1 DNA Has 10.5 bp per Turn of the Helix in Solution: The Mica Experiment

The value of 10 bp per turn varies somewhat under different conditions. A classic experiment that was performed in the 1970s showed that DNA absorbed on a surface has somewhat greater than 10 bp per turn. Short segments of DNA were allowed to bind to a mica surface. The presence of 5'-terminal phosphates on the DNAs held them in a fixed orientation on the mica. The mica-bound DNAs were then exposed to DNase I, an enzyme (a deoxyribonuclease) that cleaves the phosphodiester bonds in the DNA backbone. Because the enzyme is bulky, it is able to cleave phosphodiester bonds only on the DNA surface furthest from the mica (think of the DNA as a cylinder lying down on a flat surface) because of the steric difficulty of reaching the sides or bottom surface of the DNA. As a result, the length of the resulting fragments should reflect the periodicity of the DNA, the number of base pairs per turn.

After the mica-bound DNA was exposed to DNase, the resulting fragments were separated by electrophoresis in a polyacrylamide gel, a jelly-like matrix (Box 4-1 Fig. 1; see also Chapter 7 for an explanation of gel electrophoresis). Because DNA is negatively charged, it migrates through the gel toward the positive pole of the electric field. The gel matrix impedes movement of the fragments in a manner that is proportional to their length such that larger fragments migrate more slowly than smaller fragments. When the experiment is performed, we see clusters of DNA fragments of average sizes 10 and 11, 21, 31, and 32 bp and so forth, that is, in multiples of 10.5,

which is the number of base pairs per turn. This value of 10.5 bp per turn is close to that of DNA in solution as inferred by other methods (see the section titled *The Double Helix Exists in Multiple Conformations*). The strategy of using DNase to probe the structure of DNA is now used to analyze the interaction of DNA with proteins (see Chapter 7).



**BOX 4-1 FIGURE 1** The mica experiment.

A consequence of the helical nature of DNA is its periodicity. Each base pair is displaced (twisted) from the previous one by  $\sim 36^\circ$ . Thus, in the X-ray crystal structure of DNA, it takes a stack of  $\sim 10$  base pairs to go completely around the helix ( $360^\circ$ ) (Fig. 4-1a). That is, the helical periodicity is generally 10 base pairs per turn of the helix. (For further discussion, see Box 4-1, DNA Has 10.5 bp per Turn of the Helix in Solution: The Mica Experiment.)

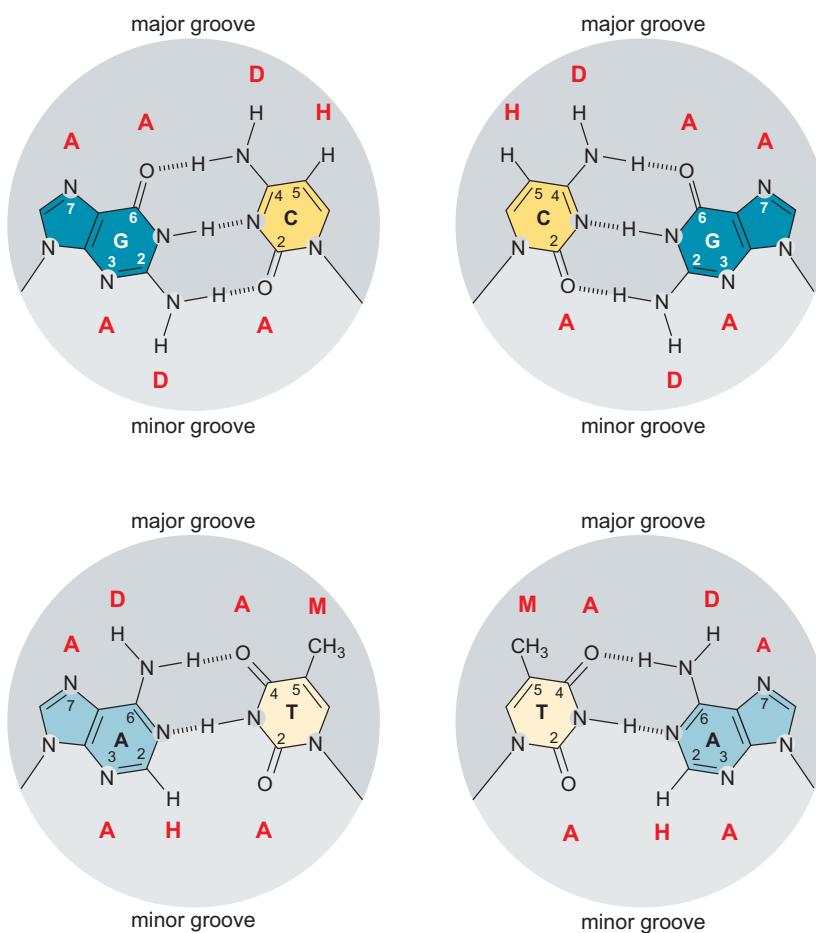
### The Double Helix Has Minor and Major Grooves

As a result of the double-helical structure of the two chains, the DNA molecule is a long, extended polymer with two grooves that are not equal in size to each other. Why are there a minor groove and a major groove? It is a simple consequence of the geometry of the base pair. The angle at which the two sugars protrude from the base pairs (i.e., the angle between the glycosidic bonds) is  $\sim 120^\circ$  (for the narrow angle) or  $240^\circ$  (for the wide angle) (see Figs. 4-1b and 4-6). As a result, as more and more base pairs stack on top of each other, the narrow angle between the sugars on one edge of the base pairs generates a **minor groove** and the large angle on the other edge generates a **major groove**. (If the sugars pointed away from each other in a straight line, i.e., at an angle of  $180^\circ$ , then the two grooves would be of equal dimensions and there would be no minor and major grooves.)

### The Major Groove Is Rich in Chemical Information

The edges of each base pair are exposed in the major and minor grooves, creating a pattern of hydrogen-bond donors and acceptors and of hydrophobic groups (allowing for van der Waals interactions) that identifies the base pair (see Fig. 4-10). The edge of an A:T base pair displays the following chemical groups in the following order in the major groove: a hydrogen-bond acceptor (the N7 of adenine), a hydrogen-bond donor (the exocyclic amino group on C6 of adenine), a hydrogen-bond acceptor (the carbonyl group on C4 of thymine), and a bulky hydrophobic surface (the methyl group on C5 of thymine). Similarly, the edge of a G:C base pair displays the following groups in the major groove: a hydrogen-bond acceptor (at N7 of guanine), a hydrogen-bond acceptor (the carbonyl on C6 of guanine), a hydrogen-bond donor (the exocyclic amino group on C4 of cytosine), and a small nonpolar hydrogen (the hydrogen at C5 of cytosine).

Thus, there are characteristic patterns of hydrogen bonding and of overall shape that are exposed in the major groove that distinguish an A:T base pair from a G:C base pair, and, for that matter, A:T from T:A, and G:C from C:G. We can think of these features as a code in which **A** represents a **hydrogen-bond acceptor**, **D** a **hydrogen-bond donor**, **M** a **methyl group**, and **H** a **nonpolar hydrogen**. In such a code, **ADAM** in the major groove signifies an A:T base pair, and **AADH** stands for a G:C base pair. Likewise, **MADA** stands for a T:A base pair, and **HDAA** is characteristic of a C:G base pair. In all cases, this code of chemical groups in the major groove specifies the identity of the base pair. These patterns are important because



**FIGURE 4-10** Chemical groups exposed in the major and minor grooves along the edges of the base pairs. The letters in red identify hydrogen-bond acceptors (A), hydrogen-bond donors (D), nonpolar hydrogens (H), and methyl groups (M).

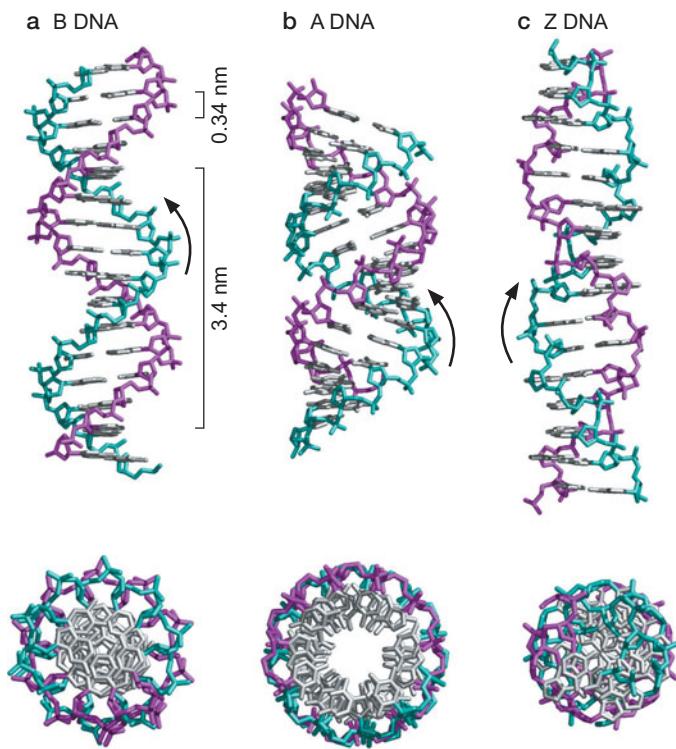
they allow proteins to unambiguously recognize DNA sequences without having to open and thereby disrupt the double helix. Indeed, as we shall see, a principal decoding mechanism relies on the ability of amino acid side chains to protrude into the major groove and to recognize and bind to specific DNA sequences (see Chapter 6).

The minor groove is not as rich in chemical information, and what information is available is less useful for distinguishing between base pairs. The small size of the minor groove is less able to accommodate amino acid side chains. In addition, A:T and T:A base pairs and G:C and C:G base pairs look similar to one another in the minor groove. An A:T base pair has a hydrogen-bond acceptor (at N3 of adenine), a nonpolar hydrogen (at N2 of adenine), and a hydrogen-bond acceptor (the carbonyl on C2 of thymine). Thus, its code is **AHA**. But this code is the same if read in the opposite direction, and hence an A:T base pair does not look very different from a T:A base pair from the point of view of the hydrogen-bonding properties of a protein poking its side chains into the minor groove. Likewise, a G:C base pair exhibits a hydrogen-bond acceptor (at N3 of guanine), a hydrogen-bond donor (the exocyclic amino group on C2 of guanine), and a hydrogen-bond acceptor (the carbonyl on C2 of cytosine), representing the code **ADA**. Thus, from the point of view of hydrogen bonding, C:G and G:C base pairs do not look very different from each other either. The minor groove *does* look different when comparing an A:T base pair with a G:C base pair, but G:C and C:G, or A:T and T:A, cannot be easily distinguished (see Fig. 4-10). Although the minor groove is less useful in distinguishing one base pair from another, the identical pattern of hydrogen-bond acceptors displayed in the minor groove of all Watson–Crick base pairs is frequently exploited by proteins to recognize correctly base-paired, B-form DNA (e.g., DNA polymerases; see Chapter 9).

### The Double Helix Exists in Multiple Conformations

Early X-ray diffraction studies of DNA, which were performed using concentrated solutions of DNA that had been drawn out into thin fibers, revealed two kinds of structures, the B and the A forms of DNA (Fig. 4-11; see Box 4-2, How Spots on an X-Ray Film Reveal the Structure of DNA). The B form, which is observed at high humidity, most closely corresponds to the average structure of DNA under physiological conditions. It has 10 bp per turn and a wide major groove and a narrow minor groove. The A form, which is observed under conditions of low humidity, has 11 bp per turn. Its major groove is narrower and much deeper than that of the B form, and its minor groove is broader and shallower. The vast majority of the DNA in the cell is in the B form, but DNA *does* adopt the A structure in certain DNA–protein complexes. In addition, as we shall see, the A form is similar to the structure that RNA adopts when double-helical.

The B form of DNA represents an ideal structure that deviates in two respects from the DNA in cells. First, DNA in solution, as we have seen, is somewhat more twisted on average than the B form, having on average 10.5 bp per turn of the helix. Second, the B form is an average structure, whereas real DNA is not perfectly regular. Rather, it shows variations in its precise structure from base pair to base pair. This was revealed by comparison of the crystal structures of individual DNAs of different sequences. For example, the two members of each base pair do not always lie exactly in the same plane. Rather, they can display a “propeller twist” arrangement in which the two flat bases counterrotate relative to each other along the long axis of the base pair, giving the base pair a propeller-like character (Fig. 4-12). Moreover, the precise rotation per base pair is not a constant. As a result, the width of the major and minor grooves varies locally. Thus, DNA molecules are

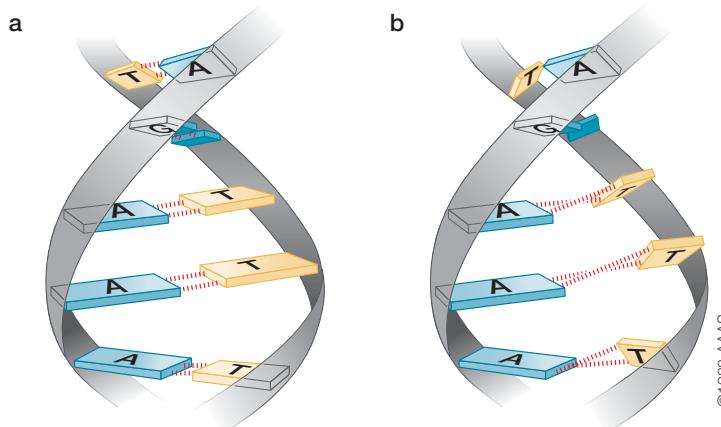


**FIGURE 4-11** Models of the B, A, and Z forms of DNA. The sugar–phosphate backbone of each chain is on the outside in all structures (one purple and one green) with the bases (silver) oriented inward. Side views are shown at the top, and views along the helical axis at the bottom. (a) The B form of DNA, the usual form found in cells, is characterized by a helical turn every 10 base pairs (3.4 nm); adjacent stacked base pairs are 0.34 nm apart. The major and minor grooves are also visible. (b) The more compact A form of DNA has 11 bp per turn and shows a large tilt of the base pairs with respect to the helix axis. In addition, the A form has a central hole (bottom). This helical form is adopted by RNA–DNA and RNA–RNA helices. (c) Z DNA is a left-handed helix and has a zig-zag (hence “Z”) appearance. (Courtesy of C. Kielkopf and P.B. Dervan.)

never perfectly regular double helices. Instead, their exact conformation depends on which base pair (A:T, T:A, G:C, or C:G) is present at each position along the double helix and on the identity of neighboring base pairs. Still, the B form is for many purposes a good first approximation of the structure of DNA in cells.

### DNA Can Sometimes Form a Left-Handed Helix

DNA containing alternative purine and pyrimidine residues can fold into left-handed as well as right-handed helices. To understand how DNA can form a left-handed helix, we need to consider the glycosidic bond that connects the base to the 1' position of 2'-deoxyribose. This bond can be in one of two conformations called *syn* and *anti* (Fig. 4-13). In right-handed DNA, the glycosidic bond is always in the *anti* conformation. In the left-handed



**FIGURE 4-12** The propeller twist between the purine and pyrimidine base pairs of a right-handed helix. (a) The structure shows a sequence of three consecutive A:T base pairs with normal Watson–Crick bonding. (b) A propeller twist causes rotation of the bases about their long axes. (Adapted, with permission, from Aggarwal A.K. et al. 1988. *Science* **242**: 899–907, Fig. 5b. © AAAS.)

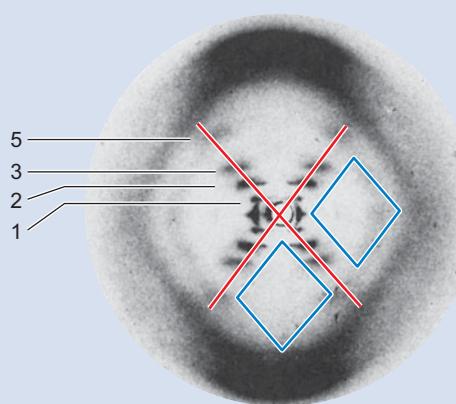
## ► KEY EXPERIMENTS

### Box 4-2 How Spots on an X-Ray Film Reveal the Structure of DNA

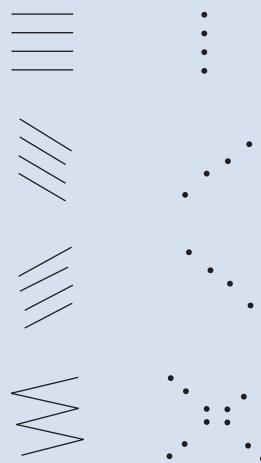
One of the most enduring images in the history of molecular biology is the famous photograph taken by Rosalind Franklin of the X-ray diffraction pattern of an oriented fiber of DNA molecules. Franklin's image is of great historic significance because it provided critical evidence in support of the Watson–Crick model for B-form DNA. In addition, Francis Crick, who had helped develop the theory of the diffraction of helical molecules, was able to infer from the pattern of spots that the strands of DNA are twisted around each other. At first glance, Franklin's image shows no recognizable relationship to a double helix. How then did this mysterious pattern of spots help unravel the atomic structure of the genetic material?

As seen in the figure, Franklin's image consists of a central "Maltese" cross (highlighted in red in Box 4-2 Fig. 1), which is composed of broad spots (the breadth of the spots reflecting disorder in the fiber). The spots are evenly spaced along horizontal "layer" lines (numbered in the figure). Notice that counting up and down from the center of the cross, the spots at the fourth layer line are missing. Notice also that the Maltese cross and the intensely dark regions at the top and bottom of the image create a series of four diamond-shaped areas (two examples of which are highlighted in blue). As we now explain, it can be understood in qualitative terms from a few simple considerations about the nature of wave diffraction that this seemingly arcane pattern of spots corresponds to the structure of the double helix.

The principle underlying X-ray diffraction is that when waves pass through a periodic array, interference occurs between the waves if the wavelength of the waves is similar to the repeat distance of the array. (Hence, X-rays, which have a very short wavelength of 0.15 nm, are used for revealing atomic structure.) If the oscillations of the waves are aligned, the waves reinforce each other ("constructive interference"), but if the troughs of one set of waves are aligned with the peaks of another set of



**BOX 4-2 FIGURE 1** Rosalind Franklin's X-ray diffraction image of DNA revealing the Maltese cross. (Modified, with permission, from Franklin R.E. and Gosling R.G. 1953. *Nature* **171**: 740–741. © Macmillan.)



**BOX 4-2 FIGURE 2** Diffraction pattern of waves passing through parallel lines.

waves, the waves cancel each other out ("destructive interference"). Thus, a beam of waves passing through an array consisting of a horizontal set of lines would generate a row of spots perpendicular (vertical) to the lines (Box 4-2 Fig. 2). Now suppose that the horizontal lines are tilted. This would result in a tilted row of spots (again, perpendicular to the tilt of the lines). Next, suppose that waves are passing through two sets of tilted lines linked to each other in zigzag fashion as in the figure: this results in a cross composed of two tilted rows of spots.

Now let us turn our attention to DNA. Imagine the backbone of one strand of the double helix projected onto a flat surface. Loosely speaking, this would create a linked series of zigs and zags (or, more properly, a sinusoidal curve). If we think of the zigs as generating one set of tilted lines and the zags as generating another set, then waves passing through the zigs and zags will generate two rows of spots that cross each other as in the example above. This is the basis for the Maltese cross in the Franklin photograph, and hence the cross reveals that DNA is helical. Knowledge of the wavelength of X-rays and measurements of the spacing between the layer lines further reveals that the helix has a periodicity of 3.4 nm. Of course, DNA consists of two helical backbones, not one. This, too, is revealed in the Franklin photograph. The helices of DNA are out of register with each other by three-eighths of a helical repeat. It turns out that this offset between the helices creates an additional destructive interference that obliterates the fourth layer line. Thus, the missing fourth layer line shows that DNA is a double helix and tells us how the two helices are aligned relative to each other.

Finally, the DNA backbone is not a smooth line as in our imaginary example. Rather, it is granular at the atomic level, consisting of sugar–phosphate units. This granularity results in additional intensities, particularly north and south of the center of the cross, to create a pattern of four diamonds. In

**Box 4-2** (Continued)

higher-resolution photographs than the one shown here, one can count 10 layer lines from the center of the cross to the north and south poles. This feature of the diffraction pattern reveals that the periodicity of the double helix (3.4 nm) is 10 times the atomic periodicity, corresponding to 10 repeating units at a spacing of 0.34 nm. Because there is one base per

sugar–phosphate unit, the B form of DNA consists of 10 base pairs per helical period (turn of the helix).

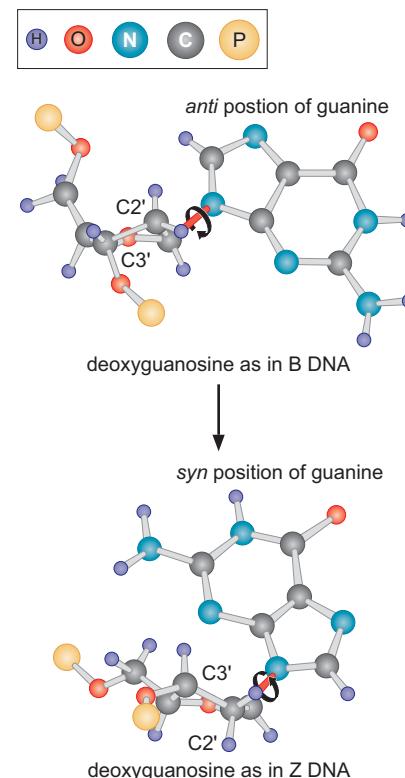
Thus, a rudimentary understanding of wave diffraction makes it possible to coax out of a simple pattern of spots on an X-ray film the principal features of the structure of DNA.

helix, the fundamental repeating unit usually is a purine–pyrimidine dinucleotide, with the glycosidic bond in the *anti* conformation at pyrimidine residues and in the *syn* conformation at purine residues. It is this *syn* conformation at the purine nucleotides that is responsible for the left-handedness of the helix. The change to the *syn* position in the purine residues to alternating *anti*–*syn* conformations gives the backbone of left-handed DNA a zigzag look (hence its designation of **Z DNA**) (see Fig. 4-11), which distinguishes it from right-handed forms. The rotation that effects the change from *anti* to *syn* also causes the sugar group to undergo a change in its pucker. Note, as shown in Figure 4-13, that C3' and C2' can switch locations. In solution, alternating purine–pyrimidine residues assume the left-handed conformation only in the presence of high concentrations of positively charged ions (e.g.,  $\text{Na}^+$ ) that shield the negatively charged phosphate groups. At lower salt concentrations, they form typical right-handed conformations. The physiological significance of Z DNA is uncertain, and left-handed helices probably account at most for only a small proportion of a cell's DNA. Further details of the A, B, and Z forms of DNA are presented in Table 4-2.

### DNA Strands Can Separate (Denature) and Reassociate

Because the two strands of the double helix are held together by relatively weak (noncovalent) forces, you might expect that the two strands could come apart easily. Indeed, the original structure for the double helix suggested that DNA replication would occur in just this manner. The complementary strands of the double helix can also be made to come apart when a solution of DNA is heated above physiological temperatures (to near 100°C) or under conditions of high pH, a process known as **denaturation**. However, this complete separation of DNA strands by denaturation is reversible. When heated solutions of denatured DNA are slowly cooled, single strands often meet their complementary strands and re-form regular double helices (Fig. 4-14). The capacity to renature denatured DNA molecules permits artificial hybrid DNA molecules to be formed by slowly cooling mixtures of denatured DNA from two different sources. Likewise, hybrids can be formed between complementary strands of DNA and RNA. As we shall see in Chapter 7, the ability to form hybrids between two single-stranded nucleic acids, called **hybridization**, is the basis for several indispensable techniques in molecular biology, such as Southern blot hybridization and DNA microarray analysis (see Chapter 7).

Important insights into the properties of the double helix were obtained from classic experiments performed in the 1950s in which the denaturation of DNA was studied under a variety of conditions. In these experiments, DNA denaturation was monitored by measuring the absorbance of ultraviolet light passed through a solution of DNA. DNA maximally absorbs



**FIGURE 4-13** *Syn* and *anti* positions of guanine in B and Z DNA. In right-handed B DNA, the glycosyl bond (red) connecting the base to the deoxyribose group is always in the *anti* position, whereas in left-handed Z DNA, it rotates in the direction of the arrow, forming the *syn* conformation at the purine (here guanine) residues, but remains in the regular *anti* position (no rotation) in the pyrimidine residues. (Adapted, with permission, from Wang A.J.H. et al. 1982. *Cold Spring Harbor Symp. Quant. Biol.* 47: 41. © Cold Spring Harbor Laboratory Press.)

**TABLE 4-2** A Comparison of the Structural Properties of A, B, and Z DNAs as Derived from Single-Crystal X-Ray Analysis

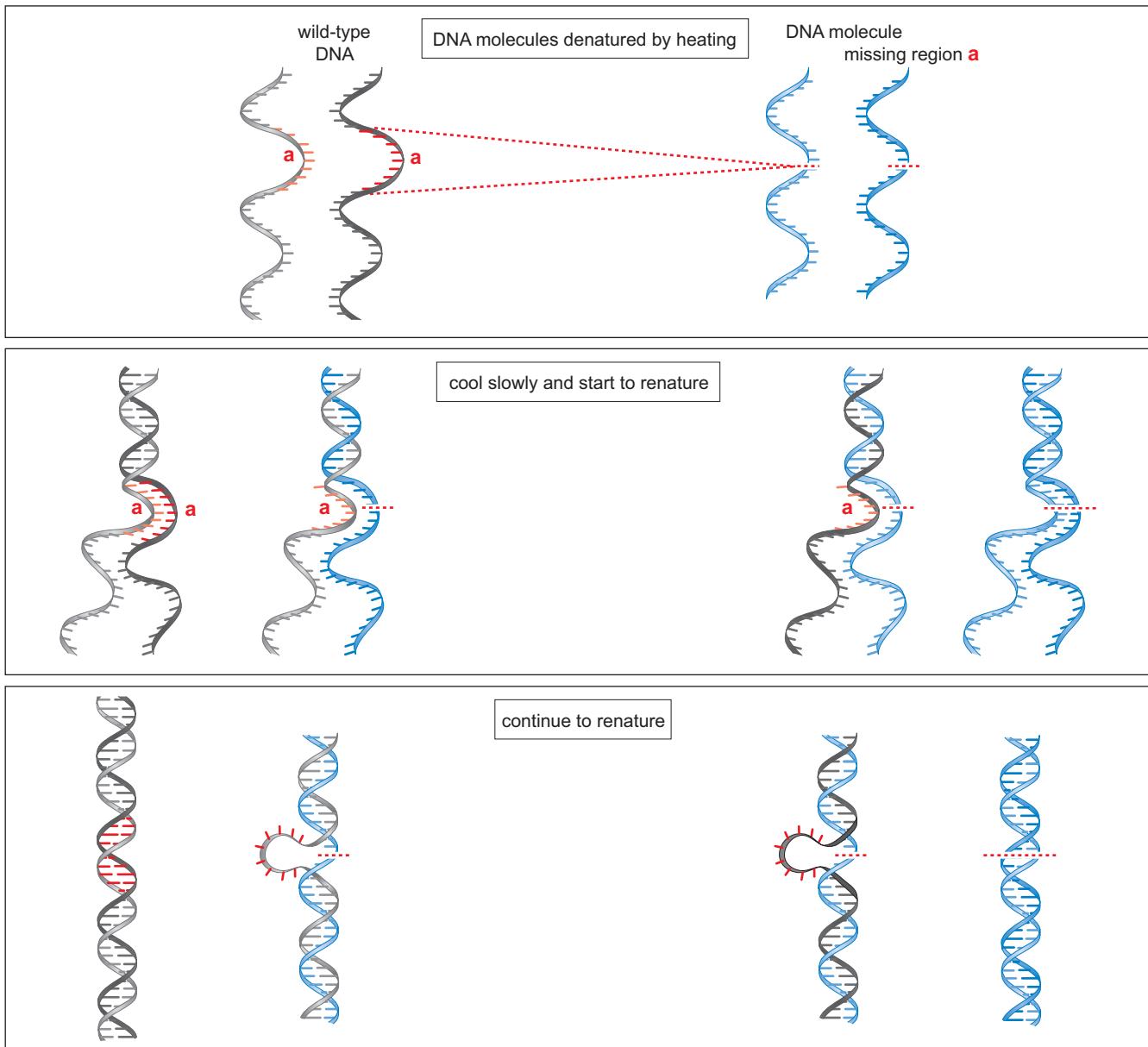
	Helix Type		
	A	B	Z
Overall proportions	Short and broad	Longer and thinner	Elongated and slim
Rise per base pair	2.3 Å	3.32 Å	3.8 Å
Helix-packing diameter	25.5 Å	23.7 Å	18.4 Å
Helix rotation sense	Right-handed	Right-handed	Left-handed
Base pairs per helix repeat	1	1	2
Base pairs per turn of helix	~11	~10	12
Rotation per base pair	33.6°	35.9°	–60° per 2 bp
Pitch per turn of helix	24.6 Å	33.2 Å	45.6 Å
Tilt of base normals to helix axis	+19°	–1.2°	–9°
Base-pair mean propeller twist	+18°	+16°	~0°
Helix axis location	Major groove	Through base pairs	Minor groove
Major-groove proportions	Extremely narrow but very deep	Wide and of intermediate depth	Flattened out on helix surface
Minor-groove proportions	Very broad but shallow	Narrow and of intermediate depth	Extremely narrow but very deep
Glycosyl-bond conformation	anti	anti	anti at C, syn at G

Adapted, with permission, from Dickerson R.E. et al. 1982. *Cold Spring Harbor Symp. Quant. Biol.* 47: 14. © Cold Spring Harbor Laboratory Press.

ultraviolet light at a wavelength of ~260 nm. It is the bases that are principally responsible for this absorption. When the temperature of a solution of DNA is raised to near the boiling point of water, the optical density, (called **absorbance**) at 260 nm markedly increases, a phenomenon known as **hyperchromicity**. The explanation for this increase is that duplex DNA absorbs less ultraviolet light by ~40% than do individual DNA chains. This hypochromicity is due to base stacking, which diminishes the capacity of the bases in duplex DNA to absorb ultraviolet light.

If we plot the optical density of DNA as a function of temperature, we observe that the increase in absorption occurs abruptly over a relatively narrow temperature range. The midpoint of this transition is the **melting point** or  $T_m$  (Fig. 4-15). Like ice, DNA melts: It undergoes a transition from a highly ordered double-helical structure to a much less ordered structure of individual strands. The sharpness of the increase in absorbance at the melting temperature tells us that the denaturation and renaturation of complementary DNA strands is a highly cooperative, zipper-like process. Renaturation, for example, probably occurs by means of a slow nucleation process in which a relatively small stretch of bases on one strand finds and pairs with their complement on the complementary strand (middle panel of Fig. 4-14). The remainder of the two strands then rapidly zipper up from the nucleation site to re-form an extended double helix (lower panel of Fig. 4-14).

The melting temperature of DNA is a characteristic of each DNA that is largely determined by the G:C content of the DNA and the ionic strength of the solution. The higher the percent of G:C base pairs in the DNA (and hence the lower the content of A:T base pairs), the higher is the melting point (Fig. 4-16). Likewise, the higher the salt concentration of the solution, the greater is the temperature at which the DNA denatures. How do we explain this behavior? G:C base pairs contribute more to the stability of DNA than do A:T base pairs because of the greater number of hydrogen bonds for the former (three in a G:C base pair vs. two for A:T), but also, importantly, because the stacking interactions of G:C base pairs with adjacent base pairs are more favorable than the corresponding interactions of A:T base pairs with their neighboring base pairs. The effect of ionic strength reflects another



**FIGURE 4-14** Reannealing and hybridization. A mixture of two otherwise identical double-stranded DNA molecules, one a normal wild-type DNA and the other a mutant missing a short stretch of nucleotides (marked as region *a* in red), is denatured by heating. The denatured DNA molecules are allowed to renature by incubation just below the melting temperature. This treatment results in two types of renatured molecules. One type is composed of completely renatured molecules in which two complementary wild-type strands re-form a helix and two complementary mutant strands re-form a helix. The other type is hybrid molecules, composed of a wild-type and a mutant strand, showing a short unpaired loop of DNA (region *a*).

fundamental feature of the double helix. The backbones of the two DNA strands contain phosphoryl groups that carry a negative charge. These negative charges are close enough across the two strands that, if not shielded, they tend to cause the strands to repel each other, facilitating their separation. At high ionic strength, the negative charges are shielded by cations, thereby stabilizing the helix. Conversely, at low ionic strength, the unshielded negative charges render the helix less stable.

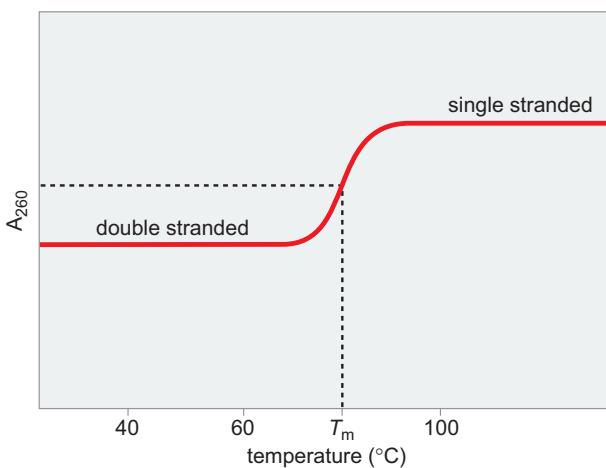
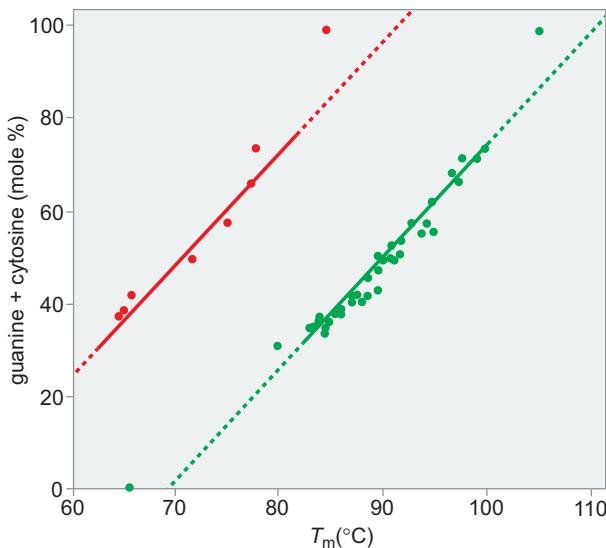


FIGURE 4-15 DNA denaturation curve.

### Some DNA Molecules Are Circles

It was initially believed that all DNA molecules are linear and have two free ends. Indeed, the chromosomes of eukaryotic cells each contain a single (extremely long) DNA molecule. But now we know that some DNAs are circles. For example, the chromosome of the small monkey DNA virus SV40 is a circular, double-helical DNA molecule of  $\sim 5000$  bp. In addition, most (but not all) bacterial chromosomes are circular; *Escherichia coli* has a circular chromosome of  $\sim 5$  million base pairs. Additionally, many bacteria have small autonomously replicating genetic elements known as **plasmids**, which are generally circular DNA molecules.

Interestingly, some DNA molecules are sometimes linear and sometimes circular. The most well-known example is that of the bacteriophage  $\lambda$ , a DNA virus of *E. coli*. The phage  $\lambda$  genome is a linear double-stranded molecule in the virion particle. However, when the  $\lambda$  genome is injected into an *E. coli* cell during infection, the DNA circularizes. This occurs by base pairing between single-stranded regions that protrude from the ends of the DNA and that have complementary sequences, also known as “sticky ends.”



**FIGURE 4-16** Dependence of DNA denaturation on G+C content and on salt concentration. The greater the G+C content, the higher the temperature must be to denature the DNA strand. DNA from different sources was dissolved in solutions of low (red line) and high (green line) concentrations of salt at pH 7.0. The points represent the temperature at which the DNA denatured graphed against the G+C content. (Data from Marmur J. and Doty P. 1962. *J. Mol. Biol.* 5: 120. © Elsevier.)

## DNA TOPOLOGY

Because DNA is a flexible structure, its exact molecular parameters are a function of both the surrounding ionic environment and the nature of the DNA-binding proteins with which it is complexed. Because their ends are free, linear DNA molecules can freely rotate to accommodate changes in the number of times the two chains of the double helix twist about each other. But if the two ends are covalently linked to form a circular DNA molecule and if there are no interruptions in the sugar–phosphate backbones of the two strands, then the absolute number of times the chains can twist about each other cannot change. Such a **covalently closed, circular DNA (cccDNA)** is said to be topologically constrained. Even the linear DNA molecules of eukaryotic chromosomes are subject to topological constraints because of their extreme length, entrainment in chromatin, and interaction with other cellular components (see Chapter 8). Despite these constraints, DNA participates in numerous dynamic processes in the cell. For example, the two strands of the double helix, which are twisted around each other, must rapidly separate in order for DNA to be duplicated and to be transcribed into RNA. Thus, understanding the topology of DNA and how the cell both accommodates and exploits topological constraints during DNA replication, transcription, and other chromosomal transactions is of fundamental importance in molecular biology.

### Linking Number Is an Invariant Topological Property of Covalently Closed, Circular DNA

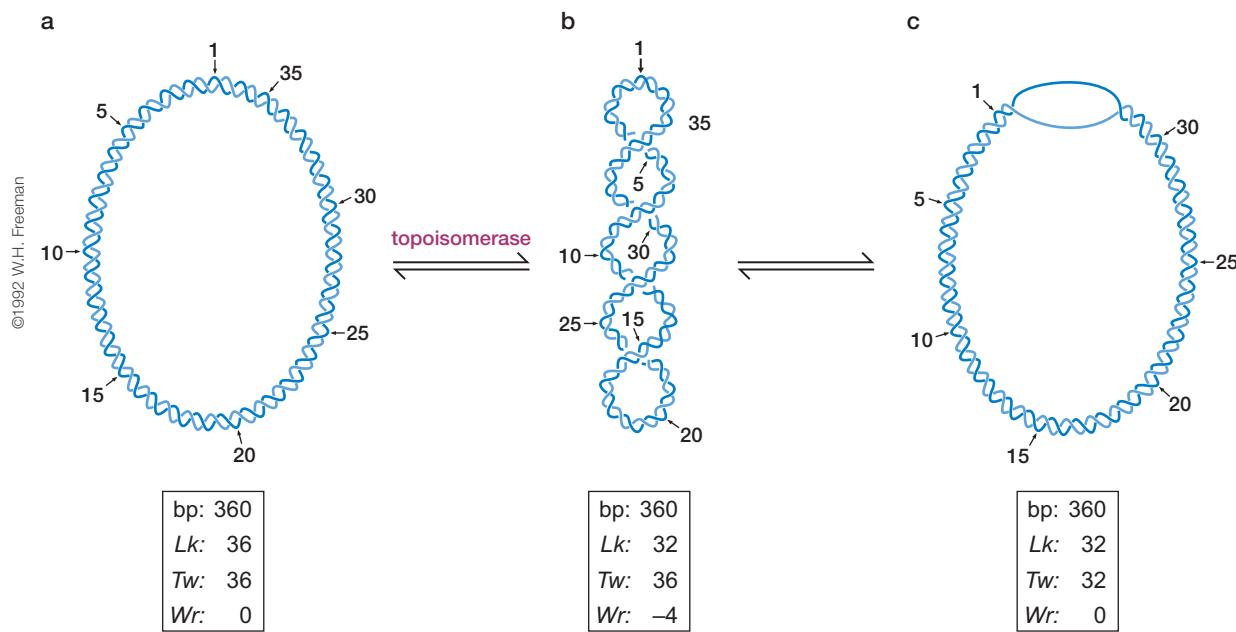
Let us consider the topological properties of covalently closed, circular DNA, which is referred to as cccDNA. Because there are no interruptions in either polynucleotide chain, the two strands of cccDNA cannot be separated from each other without the breaking of a covalent bond. If we wished to separate the two circular strands without permanently breaking any bonds in the sugar–phosphate backbones, we would have to pass one strand through the other strand repeatedly (we will encounter an enzyme that can perform just this feat!). The number of times one strand would have to be passed through the other strand in order for the two strands to be entirely separated from each other is called the **linking number** (Fig. 4-17). The linking number, which is always an integer, is an invariant topological property of cccDNA, no matter how much the shape of the DNA molecule is distorted.

### Linking Number Is Composed of Twist and Writhe

The linking number is the sum of two geometric components called the **twist** and the **writhe** (see Interactive Animation 4-1). Let us consider twist first. Twist is simply the number of helical turns of one strand about the other, that is, the number of times one strand completely wraps around the other strand. Consider a cccDNA that is lying flat on a plane. In this flat conformation, the linking number is fully composed of twist. Indeed, the twist can be easily determined by counting the number of times the two strands cross each other (see Fig. 4-17a). The helical crossovers (twist) in a right-handed helix are defined as positive such that the linking number of DNA will have a positive value.

But cccDNA is generally not lying flat on a plane. Rather, it is usually stressed torsionally such that the long axis of the double helix crosses over itself, often repeatedly, in three-dimensional space (Fig. 4-17b). This is





**FIGURE 4-17** Topological states of covalently closed, circular (ccc) DNA. The figure shows conversion of the relaxed (a) to the negatively supercoiled (b) form of DNA. The strain in the supercoiled form may be taken up by supertwisting (b) or by local disruption of base pairing (c). (Adapted from a diagram provided by Dr. M. Gellert.) (Modified, with permission, from Kornberg A. and Baker T.A. 1992. *DNA replication*, 2nd ed., Fig. 4.21, p. 32. © W.H. Freeman.)

called *writhe*. To visualize the distortions caused by torsional stress, think of the coiling of a telephone cord that has been overtwisted.

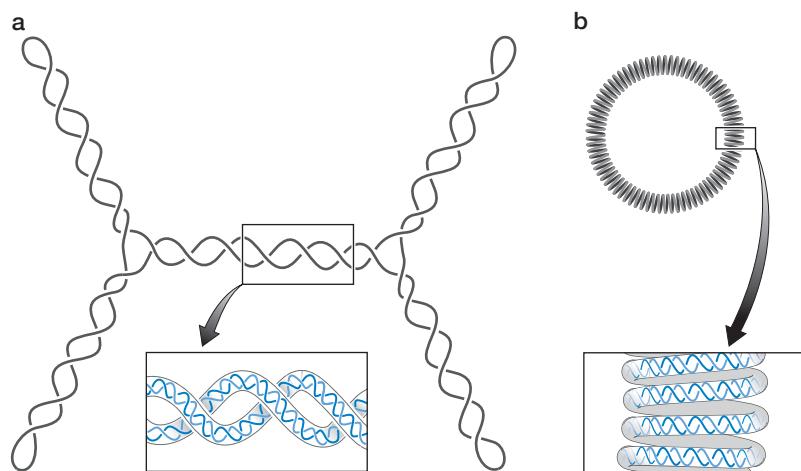
Writhe can take two forms. One form is the **interwound** or **plectonemic** writhe, in which the long axis is twisted around itself, as depicted in Figures 4-17b and 4-18a. The other form of writhe is a **toroid** or **spiral** in which the long axis is wound in a cylindrical manner, as often occurs when DNA wraps around protein (Fig. 4-18b). The **writhing number (Wr)** is the total number of interwound and/or spiral writhes in cccDNA. For example, the molecule shown in Figure 4-17b has a writhing number of 4.

Interwound writhe and spiral writhe are topologically equivalent to each other and are readily interconvertible geometric properties of cccDNA. In addition, twist and writhe are interconvertible. A molecule of cccDNA can readily undergo distortions that convert some of its twist to writhe or some of its writhe to twist without the breakage of any covalent bonds. The only constraint is that the sum of the **twist number (Tw)** and the writhing number (*Wr*) must remain equal to the **linking number (Lk)**. This constraint is described by the equation

$$Lk = Tw + Wr.$$

### Lk° Is the Linking Number of Fully Relaxed cccDNA under Physiological Conditions

Consider cccDNA that is free of **supercoiling** (i.e., it is said to be **relaxed**) and whose twist corresponds to that of the B form of DNA in solution under physiological conditions (~10.5 bp per turn of the helix). The linking number (*Lk*) of such cccDNA under physiological conditions is assigned the symbol **Lk°**. *Lk°* for such a molecule is the number of base pairs divided by 10.5. For a cccDNA of 10,500 base pairs, *Lk* = +1000. (The sign is positive because the



**FIGURE 4-18** Two forms of writhe of supercoiled DNA. The figure shows interwound (a) and toroidal (b) writhe of cccDNA of the same length. (a) The interwound or plectonemic writhe is formed by twisting of the double-helical DNA molecule over itself as depicted in the example of a branched molecule. (b) Toroidal or spiral writhe is depicted in this example by cylindrical coils. (Modified, with permission, from Kornberg A. and Baker T.A. 1992. *DNA replication*, I 1–22, p. 33. © W.H. Freeman. Used by permission of Dr. Nicholas Cozzarelli.)

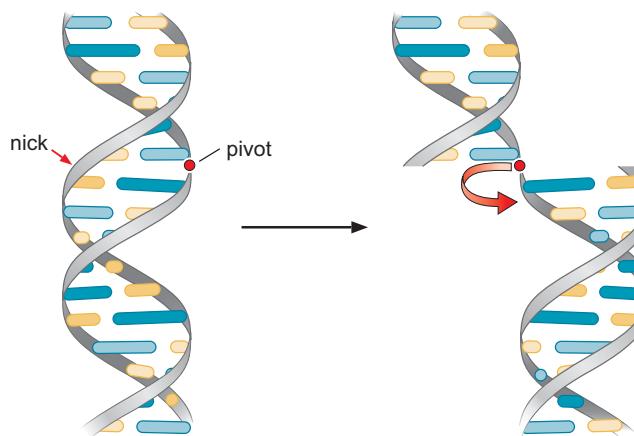
twists of DNA are right-handed.) One way to see this is to imagine pulling one strand of the 10,500-bp cccDNA out into a flat circle. If we did this, then the other strand would cross the flat circular strand 1000 times.

How can we remove supercoils from cccDNA if it is not already relaxed? One procedure is to treat the DNA mildly with the enzyme DNase I, so as to break on average one phosphodiester bond (or a small number of bonds) in each DNA molecule. Once the DNA has been “nicked” in this manner, it is no longer topologically constrained, and the strands can rotate freely, allowing writhe to dissipate (Fig. 4-19). If the nick is then repaired, the resulting cccDNA molecules will be relaxed and will have on average an  $Lk$  that is equal to  $Lk^0$ . (Because of rotational fluctuation at the time the nick is repaired, some of the resulting cccDNAs will have an  $Lk$  that is somewhat higher than  $Lk^0$ , and others will have an  $Lk$  that is somewhat lower. Thus, the relaxation procedure will generate a narrow spectrum of cccDNAs whose average  $Lk$  is equal to  $Lk^0$ .)

### DNA in Cells Is Negatively Supercoiled

The extent of supercoiling is measured by the difference between  $Lk$  and  $Lk^0$ , which is called the **linking difference**:

$$\Delta Lk = Lk - Lk^0.$$



**FIGURE 4-19** Relaxing DNA with DNase I.

If the  $\Delta Lk$  of a cccDNA is significantly different from 0, then the DNA is torsionally strained, and hence it is supercoiled. If  $Lk < Lk^0$  and  $\Delta Lk < 0$ , then the DNA is said to be “negatively supercoiled.” Conversely, if  $Lk > Lk^0$  and  $\Delta Lk > 0$ , then the DNA is “positively supercoiled.” For example, the molecule shown in Figure 4-17b is negatively supercoiled and has a linking difference of  $-4$  because its  $Lk$  (32) is 4 less than that (36) for the relaxed form of the molecule shown in Figure 4-17a.

Because  $\Delta Lk$  and  $Lk^0$  are dependent on the length of the DNA molecule, it is more convenient to refer to a normalized measure of supercoiling. This is the **superhelical density**, which is assigned the symbol  $\sigma$  and is defined as

$$\sigma = \Delta Lk / Lk^0.$$

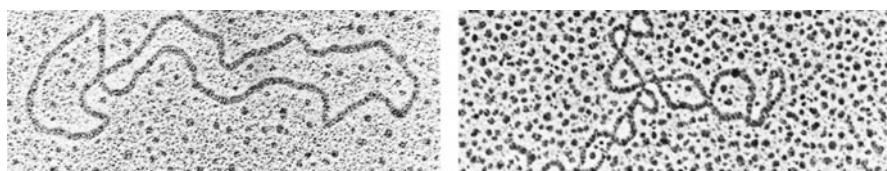
Circular DNA molecules purified from both bacteria and eukaryotes are usually negatively supercoiled, having values of  $\sigma$  of approximately  $-0.06$ . The electron micrograph shown in Figure 4-20 compares the structures of bacteriophage DNA in its relaxed form with its supercoiled form.

What does superhelical density mean biologically? Negative supercoils can be thought of as a store of free energy that aids in processes that require strand separation, such as DNA replication and transcription. Because  $Lk = Tw + Wr$ , negative supercoils can be converted into untwisting of the double helix (cf. Fig. 4-17a with 4-17b). Regions of negatively supercoiled DNA, therefore, have a tendency to unwind partially. Thus, strand separation can be accomplished more easily in negatively supercoiled DNA than in relaxed DNA.

The only organisms that have been found to have positively supercoiled DNA are certain thermophiles, microorganisms that live under conditions of extreme high temperatures, such as in hot springs. In this case, the positive supercoils can be thought of as a store of free energy that helps keep the DNA from denaturing at the elevated temperatures. Insofar as positive supercoils can be converted into more twist (positively supercoiled DNA can be thought of as being overwound), strand separation requires more energy in thermophiles than in organisms whose DNA is negatively supercoiled.

### Nucleosomes Introduce Negative Supercoiling in Eukaryotes

As we shall see in Chapter 8, DNA in the nucleus of eukaryotic cells is packaged in small particles known as **nucleosomes** in which the double helix is wrapped almost two times around the outside circumference of a protein core. You will be able to recognize this wrapping as the toroid or spiral form of writhe. Importantly, it occurs in a left-handed manner.



**FIGURE 4-20** Electron micrograph of supercoiled DNA. The left electron micrograph is a relaxed (nonsupercoiled) DNA molecule of bacteriophage PM2. The right electron micrograph shows the phage in its supertwisted form. (Electron micrographs courtesy of Wang J.C. 1982. *Sci. Am.* **247**: 97.)

(Convince yourself of this by applying the handedness rule in your mind's eye to DNA wrapped around the nucleosome in Chapter 8, Fig. 8-18.) It turns out that writhe in the form of left-handed spirals is equivalent to negative supercoils. Thus, the packaging of DNA into nucleosomes introduces negative superhelical density.

### Topoisomerases Can Relax Supercoiled DNA

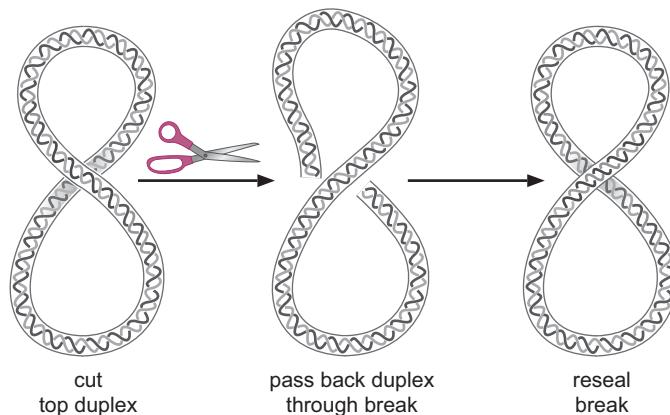
As we have seen, the linking number is an invariant property of DNA that is topologically constrained. It can be changed only by introducing interruptions into the sugar–phosphate backbone. A remarkable class of enzymes known as **topoisomerases** are able to do just this by introducing transient single-strand or double-strand breaks into the DNA (see Interactive Animation 4-2).

Topoisomerases are of two general types. Type II topoisomerases change the linking number in steps of two. They make transient double-strand breaks in the DNA through which they pass a segment of uncut duplex DNA before resealing the break. This type of reaction is shown schematically in Figure 4-21. Type II topoisomerases require the energy of ATP hydrolysis for their action. Type I topoisomerases, in contrast, change the linking number of DNA in steps of one. They make transient single-strand breaks in the DNA, allowing the uncut strand to pass through the break before resealing the nick (Fig. 4-22). In contrast to the type II topoisomerases, type I topoisomerases do not require ATP. How topoisomerases relax DNA and promote other related reactions in a controlled and concerted manner is explained below.

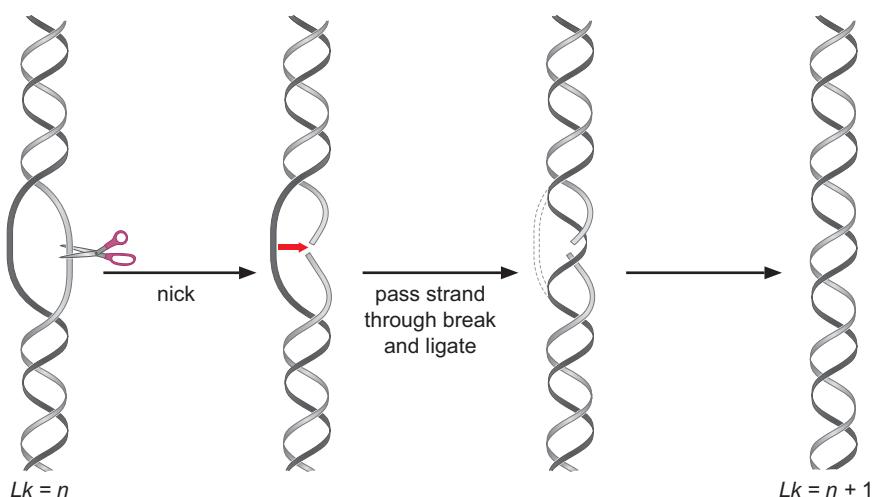


### Prokaryotes Have a Special Topoisomerase That Introduces Supercoils into DNA

Both prokaryotes and eukaryotes have type I and type II topoisomerases that are capable of removing supercoils from DNA. In addition, however, prokaryotes have a special type II topoisomerase known as “DNA gyrase” that introduces, rather than removes, negative supercoils. DNA gyrase is responsible for the negative supercoiling of chromosomes in prokaryotes. This negative supercoiling facilitates the unwinding of the DNA duplex, which stimulates many reactions of DNA including initiation of both transcription and DNA replication.



**FIGURE 4-21** Schematic for changing the linking number in DNA with topoisomerase II. Topoisomerase II binds to DNA, creates a double-strand break, passes uncut DNA through the gap, and then reseals the break.



**FIGURE 4-22** Schematic mechanism of action for topoisomerase I. The enzyme cuts a single strand of the DNA duplex, passes the uncut strand through the break, and then reseals the break. The process increases the linking number by +1.

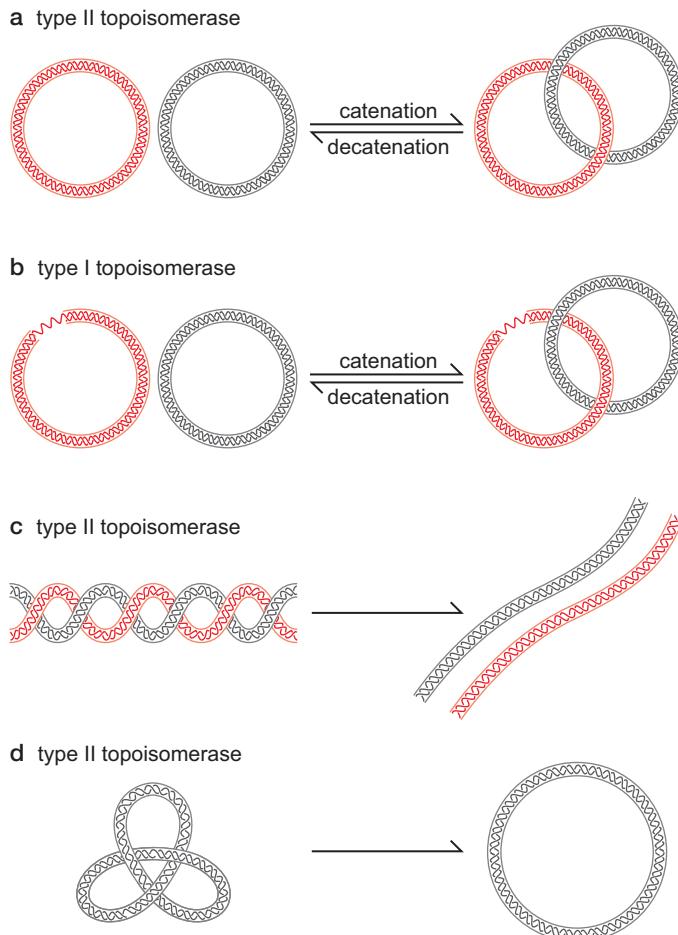
### Topoisomerases Also Unknot and Disentangle DNA Molecules

In addition to relaxing supercoiled DNA, topoisomerases promote several other reactions important to maintaining the proper DNA structure within cells. The enzymes use the same transient DNA break and strand passage reaction that they use to relax DNA to perform these reactions.

Topoisomerases can both **catenate** and **decatenate** circular DNA molecules. Circular DNA molecules are said to be catenated if they are linked together like two rings of a chain (Fig. 4-23a). Of these two activities, the ability of topoisomerases to decatenate DNA is of clear biological importance. As we shall see in Chapter 9, catenated DNA molecules are commonly produced as a round of DNA replication is finished (see Chapter 9, Fig. 9-36). Topoisomerases play the essential role of unlinking these DNA molecules to allow them to separate into the two daughter cells for cell division. Decatenation of two covalently closed, circular DNA molecules requires passage of the two DNA strands of one molecule through a double-strand break in the second DNA molecule. This reaction therefore depends on a type II topoisomerase. The requirement for decatenation explains why type II topoisomerases are essential cellular proteins. However, if at least one of the two catenated DNA molecules carries a nick or a gap, then a type I enzyme may also unlink the two molecules (Fig. 4-23b).

Although we often focus on circular DNA molecules when considering topological issues, the long linear chromosomes of eukaryotic organisms also experience topological problems. For example, during a round of DNA replication, the two double-stranded daughter DNA molecules will often become entangled (Fig. 4-23c). These sites of entanglement, just like the links between catenated DNA molecules, block the separation of the daughter chromosomes during mitosis. Therefore, DNA disentanglement, generally catalyzed by a type II topoisomerase, is also required for a successful round of DNA replication and cell division in eukaryotes.

On occasion, a DNA molecule becomes knotted (Fig. 4-23d). For example, some site-specific recombination reactions (which we shall discuss in detail in Chapter 12) give rise to knotted DNA products. Once



**FIGURE 4-23** Topoisomerases decatenate, disentangle, and unknot DNA.

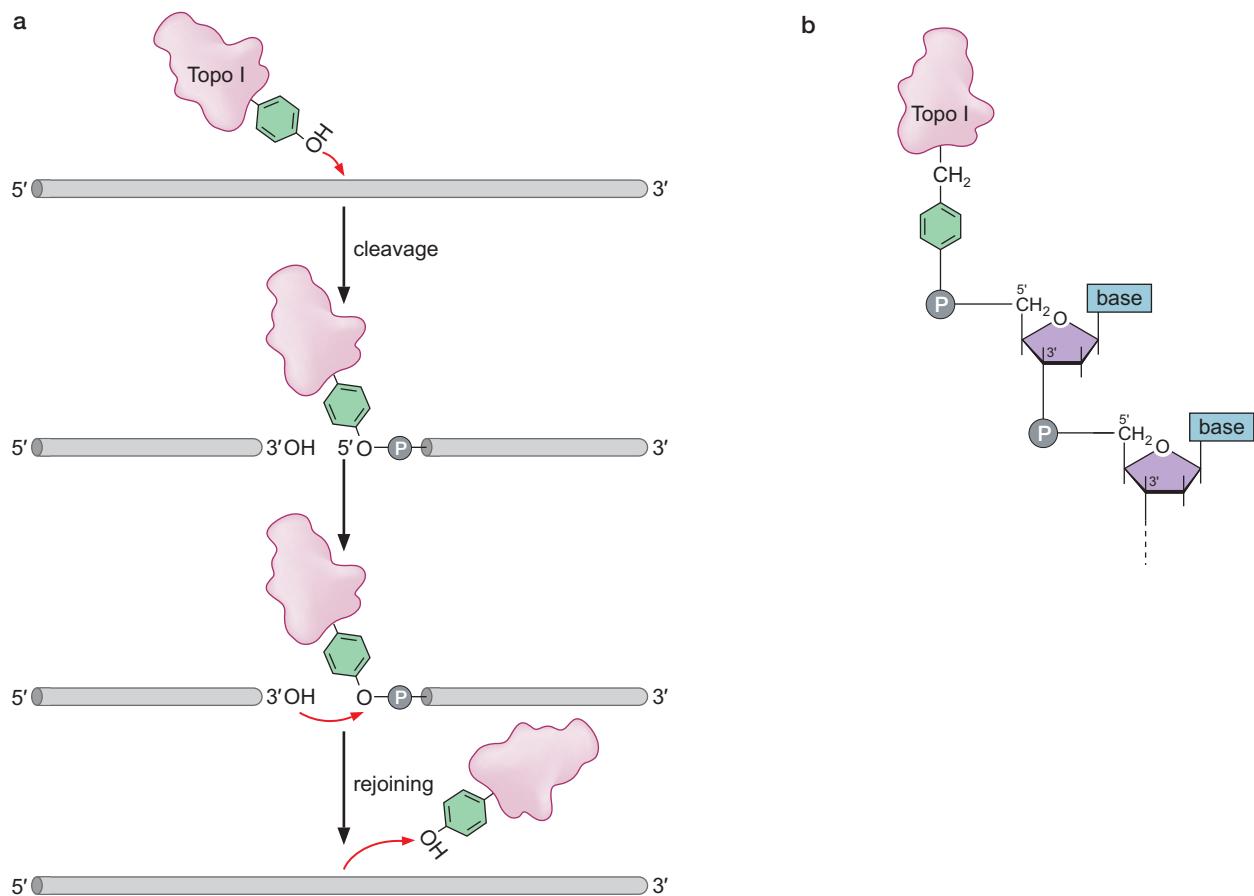
(a) Type II topoisomerases can catenate and decatenate covalently closed, circular DNA molecules by introducing a double-strand break in one DNA and passing the other DNA molecule through the break. (b) Type I topoisomerases can catenate and decatenate molecules only if one DNA strand has a nick or a gap because these enzymes cleave only one DNA strand at a time. (c) Entangled long linear DNA molecules, generated, for example, during the replication of eukaryotic chromosomes, can be disentangled by a topoisomerase. (d) DNA knots can also be unknotted by topoisomerase action.

again, a type II topoisomerase can “untie” a knot in duplex DNA. If the DNA molecule is nicked or gapped, then a type I enzyme also can do this job.

### Topoisomerases Use a Covalent Protein–DNA Linkage to Cleave and Rejoin DNA Strands

To perform their functions, topoisomerases must cleave a DNA strand (or two strands) and then rejoin the cleaved strand (or strands). Topoisomerases are able to promote both DNA cleavage and rejoicing without the assistance of other proteins or high-energy co-factors (e.g., ATP; also see below) because they use a covalent-intermediate mechanism. DNA cleavage occurs when a tyrosine residue in the active site of the topoisomerase attacks a phosphodiester bond in the backbone of the target DNA (Fig. 4-24). This attack generates a break in the DNA, whereby the topoisomerase is covalently joined to one of the broken ends via a phosphotyrosine linkage. The other end of the DNA terminates with a free OH group. This end is also held tightly by the enzyme, as we shall see below.

The phospho-tyrosine linkage conserves the energy of the phosphodiester bond that was cleaved. Therefore, the DNA can be resealed simply by reversing the original reaction: The OH group from one broken DNA end attacks the phospho-tyrosine bond re-forming the DNA phosphodiester bond. This reaction rejoins the DNA strand and releases the



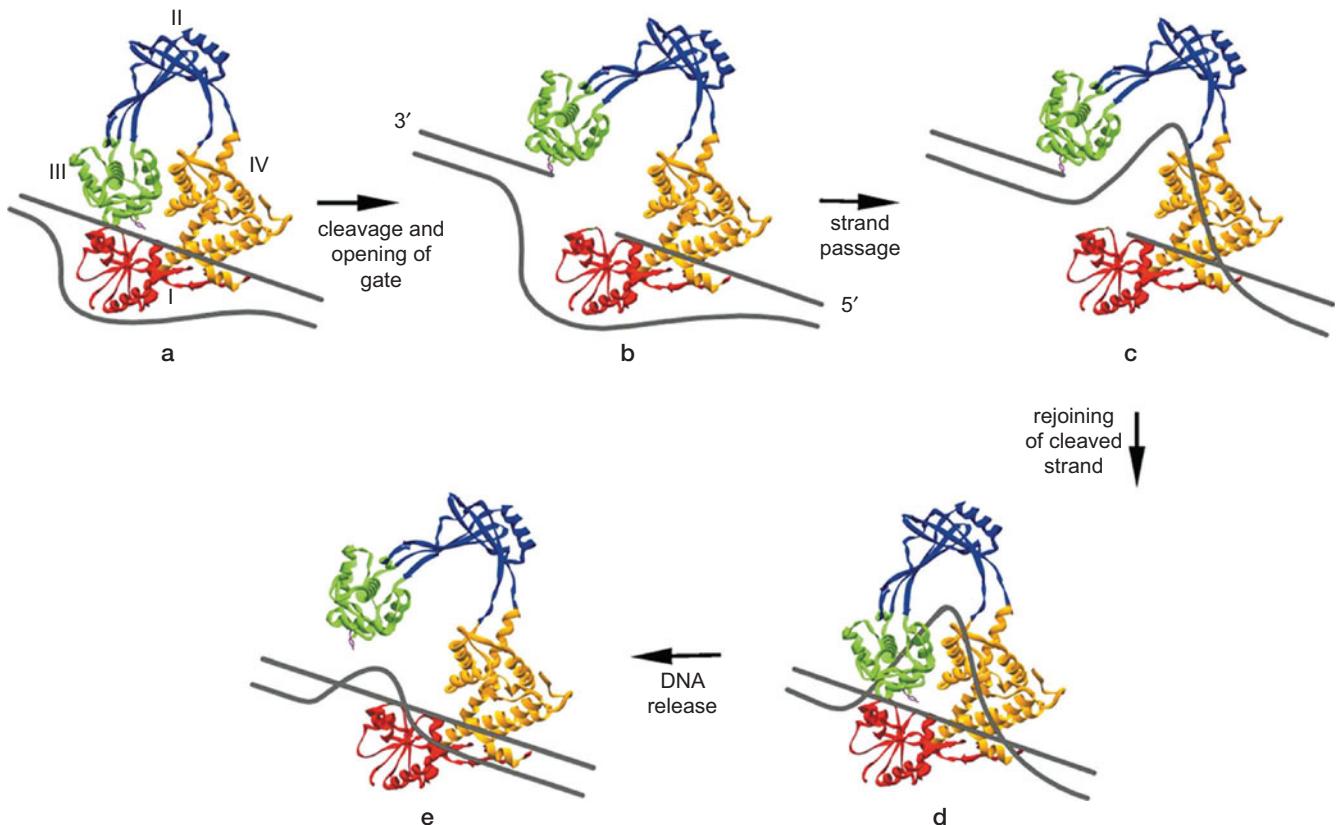
**FIGURE 4-24** Topoisomerases cleave DNA using a covalent tyrosine–DNA intermediate. (a) Schematic of the cleavage and rejoining reaction. For simplicity, only a single strand of DNA is shown. See Figure 4-25 for a more realistic picture. The same mechanism is used by type II topoisomerases, although two enzyme subunits are required, one to cleave each of the two DNA strands. Topoisomerases sometimes cut to the 5' side and sometimes to the 3' side. (b) Close-up view of the phosphotyrosine covalent intermediate.

topoisomerase, which can then go on to catalyze another reaction cycle. Although as noted above, type II topoisomerases require ATP hydrolysis for activity, the energy released by this hydrolysis is used to promote conformational changes in the topoisomerase–DNA complex rather than to cleave or rejoin DNA.

### Topoisomerases Form an Enzyme Bridge and Pass DNA Segments through Each Other

Between the steps of DNA cleavage and DNA rejoining, the topoisomerase promotes passage of a second segment of DNA through the break. Topoisomerase function thus requires that DNA cleavage, strand passage, and DNA rejoining all occur in a highly coordinated manner. Structures of several different topoisomerases have provided insight into how the reaction cycle occurs. Here we explain a model for how a type I topoisomerase relaxes DNA.

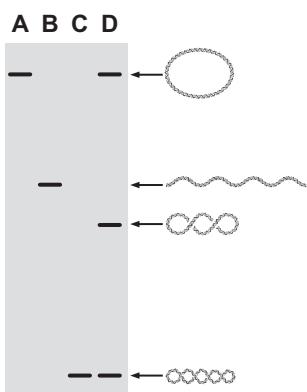
To initiate a relaxation cycle, the topoisomerase binds to a segment of duplex DNA in which the two strands are melted (Fig. 4-25a). Melting of the DNA strands is favored in highly negatively supercoiled DNA (see above),



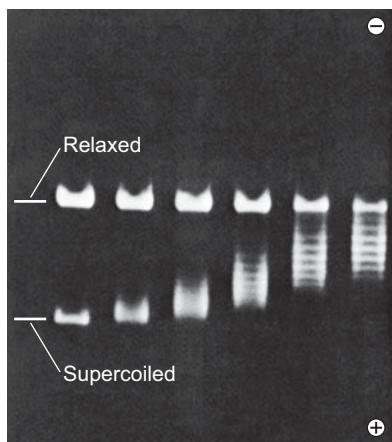
**FIGURE 4-25** Model for the reaction cycle catalyzed by a type I topoisomerase. A series of proposed steps for the relaxation of one turn of a negatively supercoiled plasmid DNA. The two strands of DNA are dark gray (not drawn to scale). The four domains of the protein are labeled in panel a: (red) Domain I; (blue) II; (green) III; (orange) IV. (Adapted, with permission, from Champoux J. 2001. *Annu. Rev. Biochem.* **70**: 369–413. © Annual Reviews.)

making this DNA an excellent substrate for relaxation. One of the DNA strands binds in a cleft in the enzyme that places it near the active-site tyrosine. This strand is cleaved to generate the covalent DNA–tyrosine intermediate (Fig. 4-25b). The success of the reaction requires that the other end of the newly cleaved DNA also be tightly bound by the enzyme. After cleavage, the topoisomerase undergoes a large conformational change to open up a gap in the cleaved strand, with the enzyme bridging the gap. The second (uncleaved) DNA strand then passes through the gap and binds to a DNA-binding site in an internal “donut-shaped” hole in the protein (Fig. 4-25c). After strand passage occurs, a second conformational change in the topoisomerase–DNA complex brings the cleaved DNA ends back together (Fig. 4-25d); rejoining of the DNA strand occurs by attack of the OH end on the phospho-tyrosine bond (see above). After rejoining, the enzyme must open up one final time to release the DNA (Fig. 4-25e). This product DNA is identical to the starting DNA molecule, except that the linking number has been increased by 1.

This general mechanism, in which the enzyme provides a “protein bridge” during the strand passage reaction, can also be applied to the type II topoisomerases. The type II enzymes, however, are dimeric (or in some cases, tetrameric). Two topoisomerase subunits, with their active-site tyrosine residues, are required to cleave the two DNA strands and make the double-stranded DNA break that is an essential feature of the type II topoisomerase mechanism.



**FIGURE 4-26** Schematic of electrophoretic separation of DNA topoisomers. (Lane A) Relaxed or nicked circular DNA; (lane B) linear DNA; (lane C) highly supercoiled cccDNA; (lane D) a ladder of topoisomers.



**FIGURE 4-27** Separation of relaxed and supercoiled DNA by gel electrophoresis. Relaxed and supercoiled DNA topoisomers are resolved by gel electrophoresis. The speed with which the DNA molecules migrate increases as the number of superhelical turns increases. (Courtesy of J.C. Wang.)

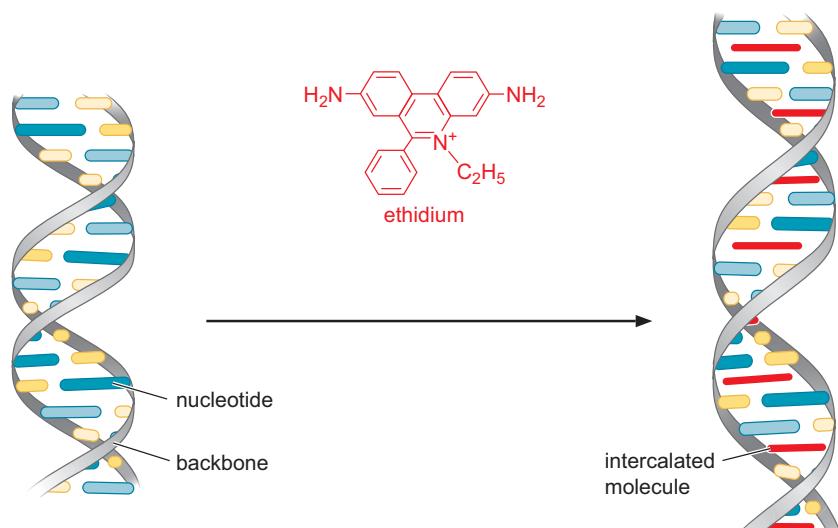
## DNA Topoisomers Can Be Separated by Electrophoresis

Covalently closed, circular DNA molecules of the same length but of different linking numbers are called **DNA topoisomers**. Even though topoisomers have the same molecular weight, they can be separated from each other by electrophoresis through a gel of agarose (see Chapter 7 for an explanation of **gel electrophoresis**). The basis for this separation is that the greater the writhe, the more compact the shape of a cccDNA. Once again, think of how supercoiling a telephone cord causes it to become more compact. The more compact the DNA, the more easily (up to a point) it is able to migrate through the gel matrix (Fig. 4-26). Thus, a fully relaxed cccDNA migrates more slowly than a highly supercoiled topoisomer of the same circular DNA. Figure 4-27 shows a ladder of DNA topoisomers resolved by gel electrophoresis. Molecules in adjacent rungs of the ladder differ from each other by a linking number difference of just 1. Obviously, electrophoretic mobility is highly sensitive to the topological state of DNA (see Box 4-3, Proving that DNA Has a Helical Periodicity of  $\sim 10.5$  bp per Turn from the Topological Properties of DNA Rings).

## Ethidium Ions Cause DNA to Unwind

**Ethidium** is a large, flat, multiringed cation. Its planar shape enables ethidium to slip, or intercalate, between the stacked base pairs of DNA (Fig. 4-28). Because it fluoresces when exposed to ultraviolet light and because its fluorescence increases dramatically after intercalation, ethidium is used as a stain to visualize DNA.

When an ethidium ion intercalates between two base pairs, it causes the DNA to unwind by  $26^\circ$ , reducing the normal rotation per base pair from  $\sim 36^\circ$  to  $\sim 10^\circ$ . In other words, ethidium decreases the twist of DNA. Imagine the extreme case of a DNA molecule that has an ethidium ion between every base pair. Instead of 10 bp per turn, it would have 36! When ethidium binds to linear DNA or to a nicked circle, it simply causes the helical pitch to increase. But consider what happens when ethidium binds to covalently closed, circular DNA. The linking number of the cccDNA does not change (no covalent bonds are broken and resealed), but the twist decreases by  $26^\circ$  for each molecule of ethidium that has bound to the DNA. Because



**FIGURE 4-28** Intercalation of ethidium into DNA. Ethidium increases the spacing of successive base pairs, distorts the regular sugar–phosphate backbone, and decreases the twist of the helix.

## ► KEY EXPERIMENTS

### Box 4-3 Proving that DNA Has a Helical Periodicity of $\sim 10.5$ bp per Turn from the Topological Properties of DNA Rings

The observation that DNA topoisomers can be separated from each other electrophoretically is the basis for a simple experiment that proves that DNA has a helical periodicity of  $\sim 10.5$  bp per turn in solution. Consider three cccDNAs of sizes 3990, 3995, and 4011 bp that were relaxed to completion by treatment with type I topoisomerase. When subjected to electrophoresis through agarose, the 3990- and 4011-bp DNAs show essentially identical mobilities. Because of thermal fluctuation, topoisomerase treatment actually generates a narrow spectrum of topoisomers, but for simplicity, let us consider the mobility of only the most abundant topoisomer (that corresponding to the cccDNA in its most relaxed state). The mobilities of the most abundant topoisomers for the 3990- and 4011-bp DNAs are indistinguishable because the 21-bp difference between them is negligible compared with the sizes of the rings. The most abundant topoisomer for the 3995-bp ring, however, is found to migrate slightly more rapidly than the other two rings even though it is only 5 bp larger than the 3990-bp ring. How are

we to explain this anomaly? The 3990- and 4011-bp rings in their most relaxed states are expected to have linking numbers equal to  $Lk^0$ , that is, 380 in the case of the 3990-bp ring (dividing the size by 10.5 bp) and 382 in the case of the 4011-bp ring. Because  $Lk$  is equal to  $Lk^0$ , the linking difference ( $\Delta Lk = Lk - Lk^0$ ) in both cases is 0, and there is no writhe. But because the linking number must be an integer, the most relaxed state for the 3995-bp ring would be either of two topoisomers having linking numbers of 380 or 381. However,  $Lk^0$  for the 3995-bp ring is 380.5. Thus, even in its most relaxed state, a covalently closed circle of 3995 bp would necessarily have about half a unit of writhe (its linking difference would be 0.5), and hence it would migrate more rapidly than the 3990- and 4011-bp circles. In other words, to explain how rings that differ in length by 21 bp (two turns of the helix) have the same mobility, whereas a ring that differs in length by only 5 bp (about half a helical turn) shows a different mobility, we must conclude that DNA in solution has a helical periodicity of  $\sim 10.5$  bp per turn.

$Lk = Tw + Wr$ , this decrease in  $Tw$  must be compensated for by a corresponding increase in  $Wr$ . If the circular DNA is initially negatively supercoiled (as is normally the case for circular DNAs isolated from cells), then the addition of ethidium will increase  $Wr$ . In other words, the addition of ethidium will relax the DNA. If enough ethidium is added, the negative supercoiling will be brought to 0, and if even more ethidium is added,  $Wr$  will increase above 0, and the DNA will become positively supercoiled.

Because the binding of ethidium increases  $Wr$ , its presence greatly affects the migration of cccDNA during gel electrophoresis. In the presence of non-saturating amounts of ethidium, negatively supercoiled circular DNAs are more relaxed and migrate more slowly, whereas relaxed cccDNAs become positively supercoiled and migrate more rapidly.

## SUMMARY

DNA is usually in the form of a right-handed double helix. The helix consists of two polydeoxynucleotide chains. Each chain is an alternating polymer of deoxyribose sugars and phosphates that are joined together via phosphodiester linkages. One of four bases protrudes from each sugar: adenine and guanine, which are purines, and thymine and cytosine, which are pyrimidines. Although the sugar–phosphate backbone is regular, the order of bases is irregular, and this is responsible for the information content of DNA. Each chain has a 5' to 3' polarity, and the two chains of the double helix are oriented in an antiparallel manner—that is, they run in opposite directions.

The polynucleotide chains are held together by base pairing and base stacking. Pairing is mediated by hydrogen bonds and results in the release of water molecules, increasing entropy. Base stacking also contributes to the stability of the double helix by favorable electron cloud

interactions between the bases (van der Waals forces) and by burying the hydrophobic surfaces of the bases (the hydrophobic effect).

Hydrogen bonding is specific: Adenine on one chain is paired with thymine on the other chain, whereas guanine is paired with cytosine. This strict base pairing reflects the fixed locations of hydrogen atoms in the purine and pyrimidine bases in the forms of those bases found in DNA. Adenine and cytosine almost always exist in the amino as opposed to the imino tautomeric form, whereas guanine and thymine almost always exist in the keto as opposed to enol form. The complementarity between the bases on the two strands gives DNA its self-coding character.

The two strands of the double helix fall apart (denature) upon exposure to high temperature, extremes of pH, or any agent that causes the breakage of hydrogen bonds. Following slow return to normal cellular conditions, the denatured

single strands can specifically reassociate to biologically active double helices (renature or anneal).

DNA in solution has a helical periodicity of  $\sim 10.5$  bp per turn of the helix. The stacking of base pairs upon each other creates a helix with two grooves. Because the sugars protrude from the bases at an angle of  $\sim 120^\circ$ , the grooves are unequal in size. The edges of each base pair are exposed in the grooves, creating a pattern of hydrogen-bond donors and acceptors and of hydrophobic groups that identifies the base pair. The wider—or *major*—groove is richer in chemical information than the narrow—or *minor*—groove and is more important for recognition by nucleotide sequence-specific binding proteins.

Almost all cellular DNAs are extremely long molecules, with only one DNA molecule within a given chromosome. Eukaryotic cells accommodate this extreme length in part by wrapping the DNA around protein particles known as nucleosomes. Most DNA molecules are linear, but some DNAs are circles, as is often the case for the chromosomes of prokaryotes and for certain viruses.

DNA is flexible. Unless the molecule is topologically constrained, it can freely rotate to accommodate changes in the number of times the two strands twist about each other. DNA is topologically constrained when it is in the form of a covalently closed circle or when it is entrained in chromatin.

The linking number is an invariant topological property of covalently closed, circular DNA. It is the number of times one strand would have to be passed through the other strand in order to separate the two circular strands. The linking number is the sum of two interconvertible geometric properties: twist, which is the number of times the two strands are wrapped around each other; and the writhing number, which is the number of times the long axis of the DNA crosses over itself in space. DNA is relaxed under physiological conditions when it has  $\sim 10.5$  bp per turn and is free of writhe. If the linking number is decreased, then the DNA becomes torsionally stressed, and it is said to be negatively supercoiled. DNA in cells is usually negatively supercoiled by  $\sim 6\%$ .

The left-handed wrapping of DNA around nucleosomes introduces negative supercoiling in eukaryotes. In prokaryotes, which lack histones, the enzyme DNA gyrase is responsible for generating negative supercoils. DNA gyrase is a member of the type II family of topoisomerases. These enzymes change the linking number of DNA in steps of two by making a transient break in the double helix and passing a region of duplex DNA through the break. Some type II topoisomerases relax supercoiled DNA, whereas DNA gyrase generates negative supercoils. Type I topoisomerases also relax supercoiled DNAs but do so in steps of one in which one DNA strand is passed through a transient nick in the other strand.

## BIBLIOGRAPHY

### Books

- Bloomfield V.A., Crothers D.M., Tinoco I. Jr., and Heast J.E. 2000. *Nucleic acids: Structures, properties, and functions*. University Science Books, Sausalito, California.
- Watson J.D., ed. 1982. *Structures of DNA*. Cold Spring Harbor Symposium on Quantitative Biology, Vol. 47. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

### DNA Structure

- Chambers D.A., ed. 1995. DNA: The double-helix—Perspective and prospective at forty years. *Ann. N.Y. Acad. Sci.* **758**: 1–472.
- Dickerson R.E. 1983. The DNA helix and how it is read. *Sci. Am.* **249**: 94–111.
- Franklin R.E. and Gosling R.G. 1953. Molecular configuration in sodium thymonucleate. *Nature* **171**: 740–741.

- Roberts R.J. 1995. On base flipping. *Cell* **82**: 9–12.
- Watson J.D. and Crick F.H.C. 1953a. Molecular structure of nucleic acids: A structure for deoxyribonucleic acids. *Nature* **171**: 737–738.
- . 1953b. Genetical implications of the structure of deoxyribonucleic acids. *Nature* **171**: 964–967.
- Wilkins M.H.F., Stokes A.R., and Wilson H.R. 1953. Molecular structure of deoxypentose nucleic acids. *Nature* **171**: 738–740.

### DNA Topology

- Dröge P. and Cozzarelli N.R. 1992. Topological structure of DNA knots and catenanes. *Methods Enzymol.* **212**: 120–130.
- Wang J.C. 2002. Cellular roles of DNA topoisomerases: A molecular perspective. *Nat. Rev. Mol. Cell Biol.* **3**: 430–440.

## QUESTIONS

### MasteringBiology®

*For instructor-assigned tutorials and problems, go to MasteringBiology.*

For answers to even-numbered questions, see Appendix 2: Answers.

**Question 1.** Explain what is meant by the following adjectives assigned to double-stranded DNA or the two strands that make up DNA.

- A. Polar
- B. Antiparallel
- C. Complementary

### Question 2.

- A. Calculate the approximate number of base pairs in four helical turns of the B-form DNA (based on the values from early X-ray diffraction structures). Calculate the approximate number of base pairs in four helical turns of the B-form DNA (based on the values for DNA in solution). Calculate the vertical length of the four helical turns in the B form of DNA.

- B.** Calculate the approximate number of base pairs in four helical turns of the A-form DNA (based on the values from early X-ray diffraction structures).

**Question 3.** True or False. If false, rewrite the sentence to be true.

- A. A DNA sequence with a higher percentage of G:C base pairs than percentage of A:T base pairs has a higher melting temperature ( $T_m$ ) primarily because of the three hydrogen bonds found in G:C base pairs compared to the two hydrogen bonds found in A:T base pairs.
- B. Entropy and base stacking are the primary factors contributing to the stability of double-stranded DNA.
- C. Base stacking determines the specificity of base pairing in DNA.

**Question 4.** For each base-pair set, state if the two pairs are distinguishable using the minor groove, the major groove, both, or neither.

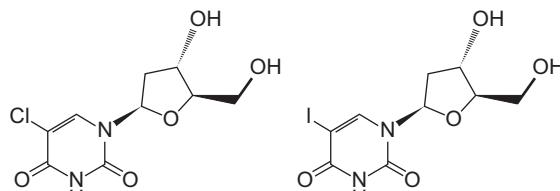
G:C versus C:G \_\_\_\_\_

A:T versus G:C \_\_\_\_\_

A:T versus T:A \_\_\_\_\_

**Question 5.** Structural analogs of nucleosides vary slightly in the chemical structure compared to the primary four nucleosides found in DNA. The structures of two nucleoside analogs used in DNA synthesis experiments are shown below.

- A. On the lines labeled 1 and 2, name the deoxyribonucleoside that these structures mimic.



1. \_\_\_\_\_ 2. \_\_\_\_\_

- B. Circle the chemical group(s) on each analog that differs from the conventional deoxyribonucleoside.

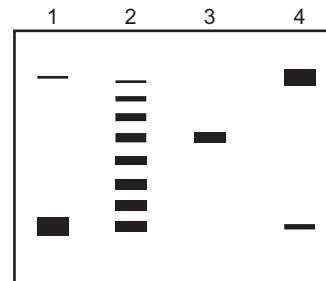
**Question 6.**

- A. You isolate a supercoiled 10,000-bp plasmid from *E. coli* cells. Assume the plasmid is covalently closed circular DNA (cccDNA). You treat the 10,000-bp plasmid with DNase I under conditions such that the DNase I nicks on average once per DNA molecule.

- Name the bond that DNase I breaks.
- How does this treatment change the topological state of the 10,000-bp plasmid?

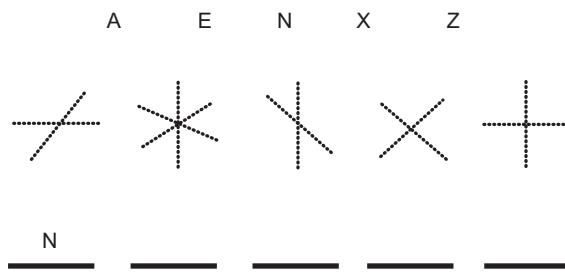
- B.** You then inactivate the DNase I and treat the DNA with DNA ligase plus ATP to create a new population of cccDNA. (DNA ligase is an enzyme that seals the nicks and requires ATP for enzymatic activity.) Briefly describe the topological state of the resulting cccDNAs. Explain your reasoning with respect to  $Lk$  (linking number) and  $Lk^0$ .

**Question 7.** You isolated a 10,500-bp plasmid (supercoiled, cccDNA) from *E. coli*. The plasmid contains one unique recognition site for EcoRI, a restriction enzyme. Restriction enzymes recognize a specific sequence and cut both strands of the DNA at that sequence (Chapter 21). You briefly incubated the cccDNA at 37°C in four separate reactions containing the components listed below. You ran the reaction on an agarose gel and visualized the DNA using ethidium bromide and UV light. The reactions included the appropriate buffer and ATP when required. An agarose gel containing four lanes of possible products is given below. For each reaction, indicate which lane on the gel contains the products that you would expect to see on your agarose gel.



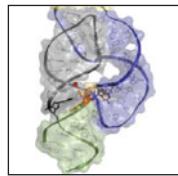
- Buffer alone
- DNase I (brief treatment)
- Topoisomerase I
- EcoRI

**Question 8.** Use your knowledge of X-ray diffraction to match the letter below with its corresponding diffraction pattern. Assume that the diffraction pattern is based on a repeating pattern (or array) of that letter, but for simplicity only a single letter is shown. (Hint: Break each letter down into its component lines and think of the diffraction pattern each line would generate. Then, combine the patterns from the individual lines to get the final diffraction pattern. An example has been completed for you.)



*This page intentionally left blank*

CHAPTER 5



# The Structure and Versatility of RNA

AT FIRST GLANCE, RNA APPEARS VERY SIMILAR TO DNA. Certainly, its chemical composition only differs from that of DNA in seemingly minor respects: a hydroxyl group in place of a hydrogen atom in its backbone and the absence of a methyl group on one of its four bases. Indeed, for many years, RNA was seen as simply playing a supporting role to DNA in information transfer. Certain viruses have RNA genomes, but in the main, DNA is the repository of genetic information in Nature. Instead, RNA was largely seen as the shuttle that transferred genetic information from DNA to the ribosome, the adaptor that decoded that information, and as a structural component of the ribosome. We now recognize, however, that RNA is far richer and more intricate in structure than DNA and far more versatile in function than first appreciated.

RNA is principally found as a single-stranded molecule. Yet by means of intrastrand base pairing, RNA exhibits extensive double-helical character and is capable of folding into a wealth of diverse tertiary structures. These structures are full of surprises, such as nonclassical base pairs, base–backbone interactions, and knot-like configurations. Most remarkable of all, some RNA molecules are enzymes, one of which performs a reaction that is at the core of information transfer from nucleic acid to protein and hence is of profound evolutionary significance.

## RNA CONTAINS RIBOSE AND URACIL AND IS USUALLY SINGLE-STRANDED

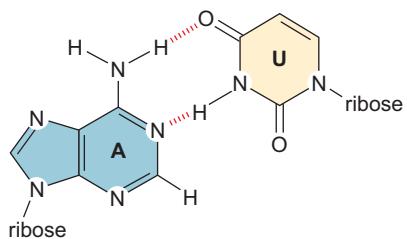
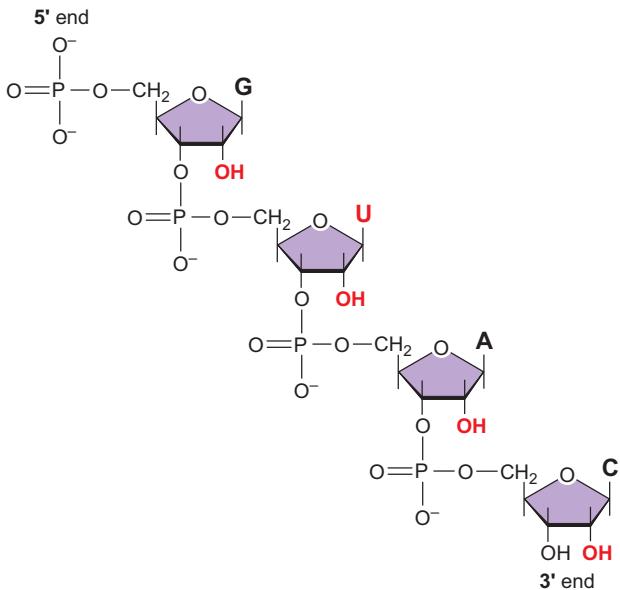
RNA differs from DNA in three respects (Fig. 5-1). First, the backbone of RNA contains ribose rather than 2'-deoxyribose. That is, ribose has a hydroxyl group at the 2' position. Second, RNA contains **uracil** in place of thymine. Uracil has the same single-ringed structure as thymine, except that it lacks the methyl group at position 5 (the **5 methyl** group). Thymine is in effect **5 methyl-uracil**. In addition, like thymine, uracil pairs with adenine (see Fig. 5-2). Given the similarity of the two bases, why has evolution selected for the presence of an extra methyl group in DNA? As we shall see in Chapter 10, the base cytosine undergoes spontaneous deamination to yield uracil, which repair systems can recognize as foreign to DNA and

## O U T L I N E

- RNA Contains Ribose and Uracil and Is Usually Single-Stranded, 107
- RNA Chains Fold Back on Themselves to Form Local Regions of Double Helix Similar to A-Form DNA, 108
- RNA Can Fold Up into Complex Tertiary Structures, 110
- Nucleotide Substitutions in Combination with Chemical Probing Predict RNA Structure, 111
- Directed Evolution Selects RNAs That Bind Small Molecules, 114
- Some RNAs Are Enzymes, 114

Visit Web Content for Structural Tutorials and Interactive Animations

**FIGURE 5-1 Structural features of RNA.** The figure shows the structure of the backbone of RNA, composed of alternating phosphate and ribose moieties. The features of RNA that distinguish it from DNA are highlighted in red.



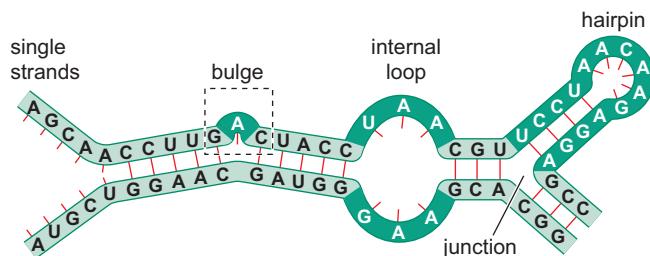
**FIGURE 5-2 Uracil pairs with adenine.** Notice that the 5' position where uracil differs from thymine is not involved in base pairing.

restore to cytosine. If the genetic material contained uracil, then uracil arising from cytosine deamination would go undetected by the surveillance systems that maintain the genome.

Third, RNA is usually found as a single polynucleotide chain. Except for certain viruses, such as those that cause influenza and acquired immune deficiency syndrome, RNA is not the genetic material and does not need to be capable of serving as a template for its own replication. Rather, RNA functions as the intermediate, the messenger RNA (mRNA), between the gene and the protein-synthesizing machinery. Another function of RNA is as an adaptor, the transfer RNA (tRNA), between the codons in the mRNA and amino acids. RNA can also play a structural role, as in the case of the RNA components of the ribosome. Yet another role for RNA is as a regulatory molecule, which through sequence complementarity binds to, and interferes with the translation of, certain mRNAs (see Chapter 20). Finally, some RNAs (including one of the structural RNAs of the ribosome) are enzymes that catalyze essential reactions in the cell. In all of these cases, the RNA is copied as a single strand off only one of the two strands of the DNA template, and its complementary strand does not exist. RNA is capable of forming long double helices, but these are unusual in nature.

### RNA CHAINS FOLD BACK ON THEMSELVES TO FORM LOCAL REGIONS OF DOUBLE HELIX SIMILAR TO A-FORM DNA

Despite being single-stranded, RNA molecules often exhibit a great deal of double-helical character (Fig. 5-3). This is because RNA chains frequently fold back on themselves to form base-paired segments between short stretches of complementary sequences. If the two stretches of complementary sequence are near each other, the RNA may adopt a **stem-loop structure** in which the intervening RNA is looped out from the end of the double-helical segment (Fig. 5-3). Stretches of double-helical RNA may also exhibit **internal loops** (unpaired nucleotides on either side of the stem), **bulges** (an unpaired nucleotide on one side of the bulge), or **junctions** (Fig. 5-3).

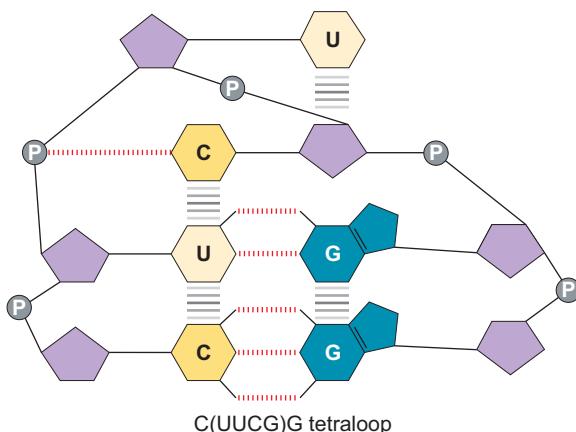


**FIGURE 5-3 Double-helical characteristics of RNA.** In an RNA molecule having regions of complementary sequences, the intervening (noncomplementary) stretches of RNA may become “looped out” to form one of the structures illustrated in the figure: a bulge, an internal loop, or a hairpin loop.

The stability of such stem-loop structures is in some instances enhanced by the special properties of the loop. For example, a stem-loop with the “tetraloop” sequence UUCG is unexpectedly stable because of special base-stacking interactions in the loop (Fig. 5-4). Base pairing can also take place between sequences that are not contiguous to form complex structures aptly named **pseudoknots** (Fig. 5-5). The regions of base pairing in RNA can be a regular double helix or they can contain discontinuities, such as noncomplementary nucleotides that bulge out from the helix.

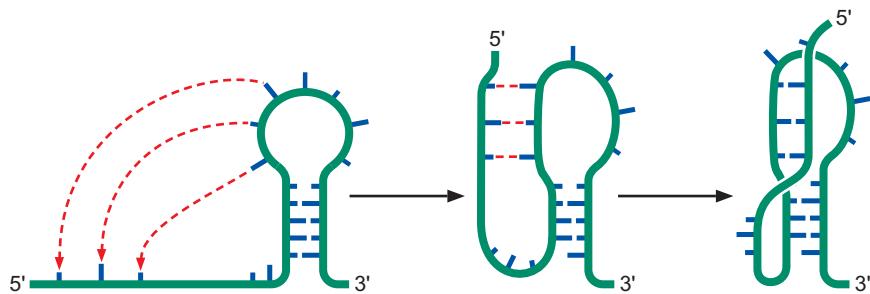
A feature of RNA that adds to its propensity to form double-helical structures is additional non-Watson–Crick base pairs. One such example is the G:U base pair, which has hydrogen bonds between N3 of uracil and the carbonyl on C6 of guanine, and between the carbonyl on C2 of uracil and N1 of guanine (Fig. 5-6). Non-Watson–Crick base pairs can be found in all combinations in RNA (GA and GU are the most abundant in ribosomal RNA). Because such noncanonical base pairs can occur as well as the two conventional Watson–Crick base pairs, RNA chains have an enhanced capacity for self-complementarity. Thus, RNA frequently exhibits local regions of base pairing but not the long-range, regular helicity of DNA.

The presence of 2'-hydroxyls in the RNA backbone prevents RNA from adopting a B-form helix. Rather, double-helical RNA resembles the A-form structure of DNA. As such, the minor groove is wide and shallow, and hence accessible; but recall that the minor groove offers little sequence-specific information. Meanwhile, the major groove is so narrow and deep that it is not very accessible to amino acid side chains from interacting proteins. Thus, the RNA double helix is quite distinct from the DNA double helix in its detailed atomic structure and less well suited for sequence-specific interactions with proteins. However, there are many examples of proteins that do



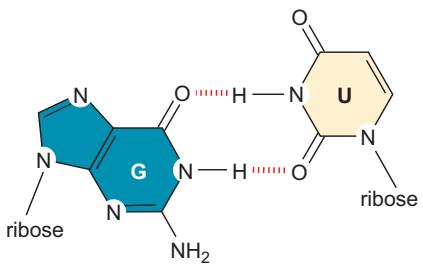
**FIGURE 5-4 Tetraloop.** Base-stacking interactions promote and stabilize the tetraloop structure. The gray circles between the riboses shown in purple represent the phosphate moieties of the RNA backbone. Horizontal lines represent base-stacking interactions.

**FIGURE 5-5 Pseudoknot.** The pseudoknot structure is formed by base pairing between noncontiguous complementary sequences.



bind to RNA in a sequence-specific manner, often relying for recognition on hairpin loops, bulges, and distortions caused by noncanonical base pairs. Examples are tRNA synthetases with their respective tRNAs; the notorious plant protein toxin ricin, which cleaves a critical glycosidic bond in the “sarcin/ricin” loop of the large RNA of the large subunit of the eukaryotic ribosome; and the human U1A protein, which binds to the U-shaped, U1A polyadenylation inhibition element in mRNA, blocking poly(A) polymerase and limiting the length of the poly(A) tail.

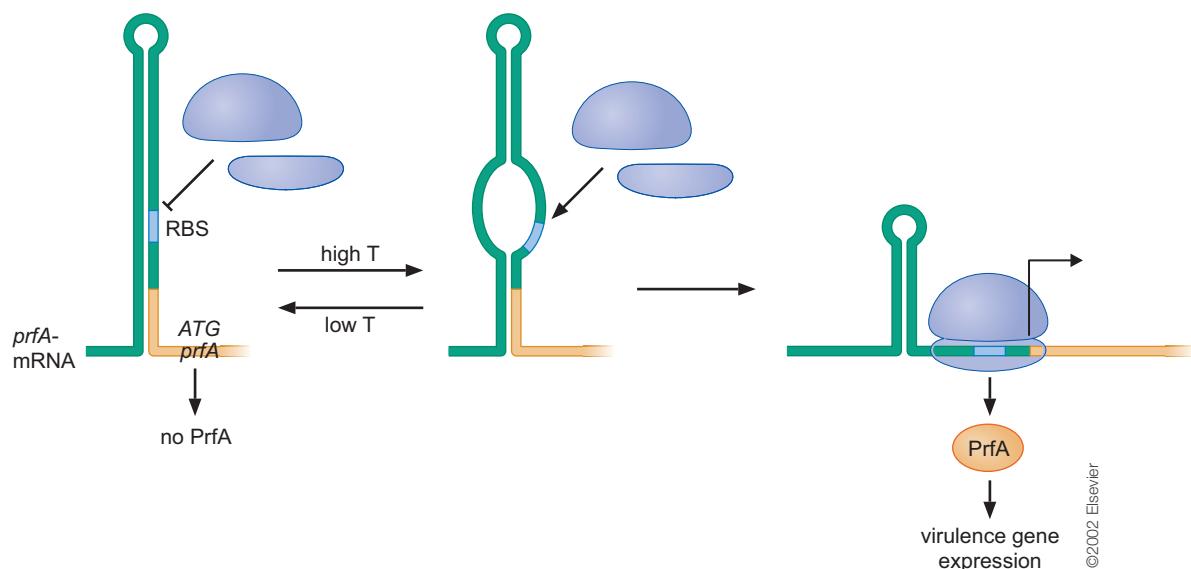
Sometimes RNA secondary structures can have important biological functions without the intervention of proteins. A striking example comes from the field of bacterial pathogenesis. Pathogenic bacteria express virulence genes that are responsible for causing disease in animals. Typically, virulence genes are not expressed outside the host; rather, expression is induced by stimuli in the infected animal. One such stimulus is temperature, which is higher inside than outside the animal. How is this increase in temperature detected by the pathogen and how does the thermosensor activate virulence genes? The answer is known for the food-borne pathogen *Listeria monocytogenes*, which causes severe illness in immune-compromised individuals and pregnant women. Virulence genes in *L. monocytogenes* are turned on by a transcription factor called PrfA, whose synthesis is, in turn, controlled at the level of translation by an upstream region in its mRNA that contains the ribosome-binding site (see Fig. 5-7). As we shall see in Chapter 15, the ribosome-binding site is a sequence that is recognized by the ribosome in the initiation of protein synthesis. The upstream region folds back on itself to form a temperature-sensitive secondary RNA structure that masks the ribosome-binding site, such that it is inaccessible to the ribosome at 30°C. At 37°C, however, the structure melts, allowing the translation machinery to gain access to the ribosome-binding site and produce PrfA. A demonstration that the secondary structure is necessary and sufficient for thermoregulation comes from the use of a fusion of the upstream region to the gene for the green fluorescence protein (see Box 5-2). When transplanted into *E. coli*, the fusion causes fluorescence at 37°C but not 30°C.



**FIGURE 5-6 G:U base pair.** The structure shows hydrogen bonds that allow base pairing to occur between guanine and uracil.

## RNA CAN FOLD UP INTO COMPLEX TERTIARY STRUCTURES

Freed of the constraint of forming long-range regular helices, RNA can adopt a wealth of tertiary structures. This is because RNA has enormous rotational freedom in the backbone of its non-base-paired regions. Thus, RNA can fold up into complex tertiary structures frequently involving unconventional base pairing, such as the base triples and base–backbone interactions seen in tRNAs (see, e.g., the illustration of the U:A:U base triple in Fig. 5-8). Proteins can assist the formation of tertiary structures by large RNA molecules,



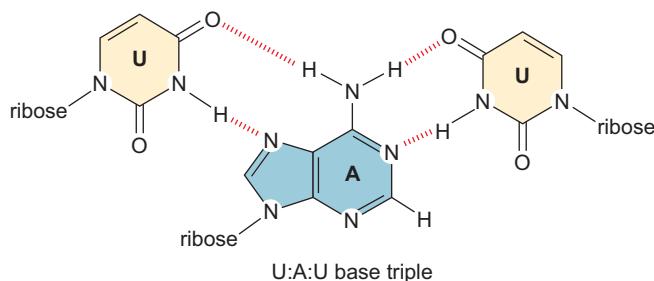
**FIGURE 5-7** A thermosensor for virulence gene expression. The *prfA* regulatory gene in *L. monocytogenes* is controlled at the level of translation by temperature-dependent unmasking of the ribosome-binding site (RBS). The blue ellipses represent ribosome subunits. (Adapted, with permission from Elsevier, from Johannsson J. et al. 2002. *Cell* 110: 551–561, Fig. 7.)

such as those found in the ribosome. Proteins shield the negative charges of backbone phosphates, whose electrostatic repulsive forces would otherwise destabilize the structure.

The tertiary structures formed by RNA are not necessarily static. Rather, the same RNA molecule might exist in one or more alternative conformations. This capacity to switch between alternative structures can sometimes be of important biological significance, as in the case of riboswitches (below and Chapter 20) and the mRNA for the murine leukemia virus (see Box 5-1).

## NUCLEOTIDE SUBSTITUTIONS IN COMBINATION WITH CHEMICAL PROBING PREDICT RNA STRUCTURE

As we have seen, RNA molecules exhibit diverse structures involving stretches of helix, bulges, loops, and long-range, tertiary interactions. How are these structures determined? Traditional approaches include nuclear magnetic resonance (NMR) and X-ray crystallography, but solving structures by these methods is challenging and NMR cannot be used for large RNA molecules. An alternative approach is to probe the secondary structure of an RNA molecule with chemicals that react with unpaired bases in combination with



**FIGURE 5-8** U:A:U base triple. The structure shows one example of hydrogen bonding that allows unusual triple base pairing.

## ► MEDICAL CONNECTIONS

**Box 5-1** An RNA Switch Controls Protein Synthesis by Murine Leukemia Virus

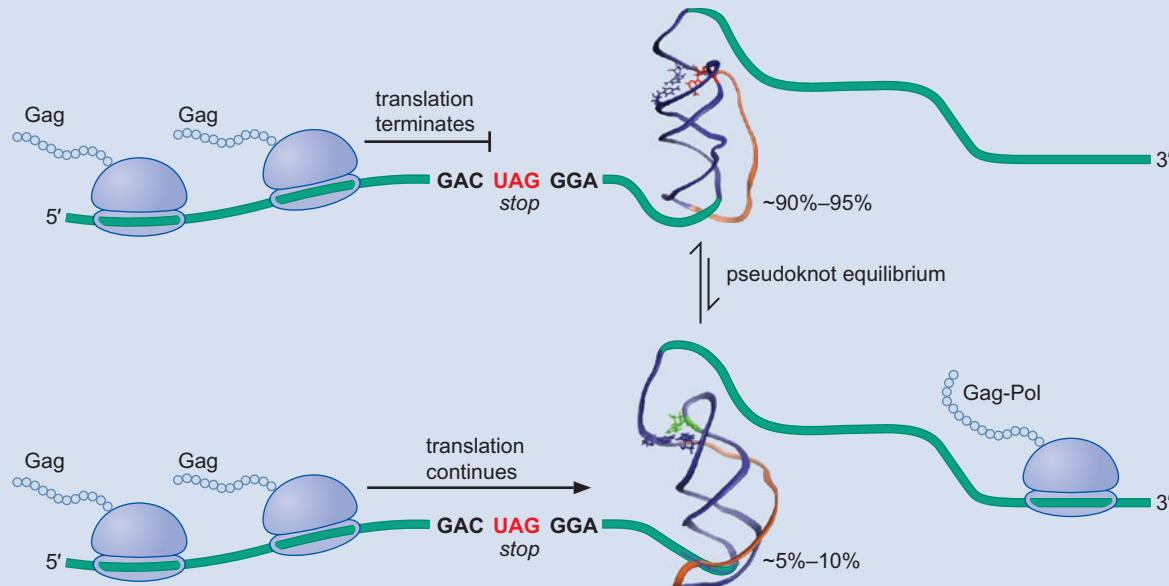
Murine leukemia virus (MLV) is an RNA virus that causes cancer in mice and certain other vertebrates. Like human immunodeficiency virus (HIV), it is a member of a class of RNA viruses known as **retroviruses** that replicate via a DNA intermediate. The RNA genome is copied into DNA by the enzyme reverse transcriptase. The RNA genome, which is also the mRNA, encodes a structural protein of the virus called Gag and a polyprotein composed of Gag and the enzyme reverse transcriptase called Pol. The gene for Gag is immediately upstream of the gene for Pol. Gag is either produced by itself or as a fusion protein, Gag-Pol, by extended translation into the downstream Pol gene. When Gag is produced by itself, the ribosome stops translation at a stop codon (see Chapter 15) located at the end of the Gag coding sequence. When the fusion protein is produced, the ribosome reads through the stop codon, continuing translation through the Pol coding sequence to create Gag-Pol. Because the structural protein is needed in greater abundance than the enzyme, it is important for the virus to maintain a proper ratio of the two proteins. The virus does this by limiting readthrough to 5%–10% of the translating ribosomes.

This design has many advantages. It eliminates the need for additional promoter elements in an already compact genome and links production of the viral enzyme to synthesis of the

structural component, allowing easy incorporation into the virus during viral budding.

How is MLV able to control translational readthrough of its mRNA? Victoria D'Souza and Steven Goff and their colleagues solved this problem by using nuclear magnetic resonance to determine the 3D structure of the MLV mRNA sequence downstream from the stop codon for the Gag coding sequence (Houck-Loomis et al. 2011). What they discovered is that the downstream sequence has a pseudoknot and that the pseudoknot does not have one structure but, rather, is in dynamic equilibrium with two conformations. One conformation limits translation to the synthesis of Gag—the inactive conformation—and the other allows readthrough to create Gag-Pol—the active conformation (see Box 5-1 Fig. 1).

To ensure that only a limited number of mRNAs read through the stop codon, the pseudoknot acts as a proton sensor. At physiological pH, the concentration of protons is such that only 5%–10% of the pseudoknots sense the protons and fold into the active conformation. To achieve this, the molecule uses the N1 nitrogen atom of an adenine, which is not generally protonated, to acquire a proton and to form a triple base in the molecule and change its conformation to the active form.

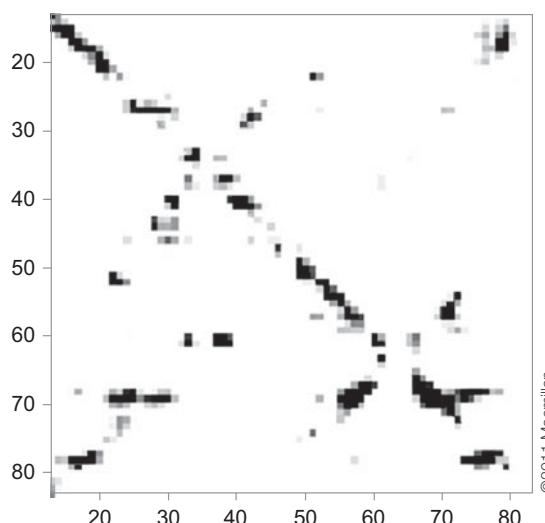


**BOX 5-1 FIGURE 1** An equilibrium between two pseudoknot conformations controls translation through a stop codon. Ribosomes translating MLV RNA encounter a stop codon (UAG) just downstream from the Gag open reading frame. When the pseudoknot is in the inactive conformation (top), translation terminates at the stop codon, resulting in the synthesis of Gag protein. When, however, the pseudoknot is in the active conformation (bottom), the ribosome is able to read through the stop codon, resulting in the synthesis of the Gag-Pol fusion protein. The equilibrium between the two conformations dictates the ratios of Gag and Gag-Pol. The adenine in the unprotonated and protonated forms are shown in red and green, respectively. (Figure kindly provided by V. D'Souza.)

algorithms that predict structure from the known energetics of stacking and hydrogen bond interactions. Such chemical approaches are often unreliable. In a striking innovation, a “mutate-and-map” strategy has been devised that allows predictions of RNA structures to be made with high confidence.

Mutate-and-map is a two-dimensional procedure that combines mutational and chemical modification approaches. First, a library of nucleotide substitutions is made in which each nucleotide is replaced with its complement within a selected RNA sequence. Next, each mutant RNA is chemically modified by a procedure known as SHAPE (for selective 2'-hydroxyl acylation analyzed by primer extension). In SHAPE, RNAs are treated with a chemical reagent (e.g., *N*-methylisatoic anhydride) that preferentially acetylates the 2'-OH of nucleotides that are unpaired. The position of unpaired nucleotides is then determined by a primer extension strategy in which DNA primers are elongated with reverse transcriptase (see Chapter 12). The reverse transcriptase ceases elongation when it encounters a chemical modification, and the positions of chemical modification are then determined from the size of the primer extension products.

Figure 5-9 shows the results of applying mutate-and-map to a type of RNA known as a riboswitch (as we shall see in Chapter 20, riboswitches are RNAs that bind to specific small molecules). The horizontal axis indicates positions along the RNA (from 5' to 3') and the vertical axis indicates the nucleotide substitution mutation for each mutant RNA (with the mutation at the 5' end at the top and the 3' end at the bottom). Boxes identify nucleotides that had reacted with the acylating reagent and hence are inferred to be unpaired. The conspicuous diagonal corresponds to unpaired nucleotides at positions that had been mutated. That is, each position of acylation along the diagonal matches each position of mutation. Boxes off the diagonal are nucleotides that had become unpaired as a direct consequence of a mutation at a different nucleotide. These often represent nucleotides that had become unpaired as a consequence of a nucleotide substitution in the complementary member of a base pair. In some cases, however, mutations are seen to destabilize an entire helix, causing several adjacent nucleotides to become unpaired. In a final step the data are analyzed using an RNA structure modeling algorithm (such as *RNAstructure*). Not only is the algorithm able to predict helices but also long-range interactions involving a very small number of adjacent nucleotides. As a consequence, the mutate-and-map strategy makes it possible to predict secondary as well as tertiary interactions. Applications of



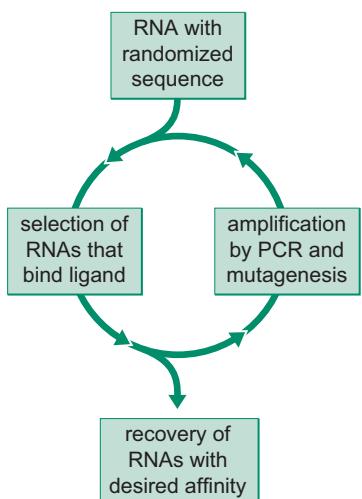
**FIGURE 5-9** A two-dimensional strategy for predicting RNA secondary and tertiary structure. Shown is the mutate-and-map data for a ribozyme in which the horizontal axis shows sites of acetylation along the RNA and the vertical axis shows sites of mutation. Sites of acylation are indicated by the gray and black boxes (with the degree of darkness indicating the extent of modification). Only the most significant data are shown based on the use of a statistical analysis algorithm. (Adapted, with permission, from Kladwang W. et al. 2011. *Nat. Chem.* 3: 954–962, Fig. 2A, p. 956. © MacMillan.)

mutate-and-map to RNAs whose structures are independently known (such as the riboswitch example shown in Fig. 5-9) have established the reliability of the method.

## DIRECTED EVOLUTION SELECTS RNAs THAT BIND SMALL MOLECULES

Researchers have taken advantage of the potential structural complexity of RNA to generate novel RNA species (not found in Nature) that have specific desirable properties by a process of directed evolution known as SELEX (for systematic evolution of ligands by exponential enrichment). By synthesizing RNA molecules with randomized sequences, it is possible to generate mixtures of oligonucleotides representing enormous sequence diversity. For example, a mixture of oligoribonucleotides of length 20 and having four possible nucleotides at each position would have a potential complexity of  $4^{20}$  sequences, or  $10^{12}$  sequences! From mixtures of diverse oligoribonucleotides, RNA molecules can be selected biochemically (e.g., by affinity chromatography) that have particular properties, such as an affinity for a specific small molecule or protein. Such RNAs are known as **aptamers**. Successive rounds of amplification by the polymerase chain reaction (Chapter 7) and sequence diversification achieved by use of a mutagenic polymerase followed by rounds of selection can enrich for aptamers with progressively higher and higher affinities for the small molecule or protein ligand (Fig. 5-10). Examples of ligands recognized by RNA aptamers are ATP, kanamycin, tobramycin, neomycin, cyanocobalamine, the prion protein PrP, and the coagulation factor X11a. A clever strategy, based on rounds of SELEX, has produced yet another example—RNA molecules having a high affinity for a fluorophore in a manner similar to that of the green fluorescent protein (see Box 5-2).

Indeed, Nature has done just this, as we shall see in Chapter 20. Metabolic operons in bacteria are sometimes under the control of regulatory RNA elements known as **riboswitches** that bind and respond to small molecule ligands in controlling gene transcription and translation. Examples of metabolites that are recognized by these riboswitches are the amino acid lysine, the nucleobase guanine, the enzyme co-factor co-enzyme B12, and the metabolite glucosamine-6-phosphate, as we discuss below.



**FIGURE 5-10** Cycle for creating RNAs that bind small molecules by SELEX.

## SOME RNAs ARE ENZYMES

It was widely believed for many years that only proteins could be enzymes. An enzyme must be able to bind a substrate, perform a chemical reaction, release the product, and repeat this sequence of events many times. Proteins are well-suited to this task because they are composed of many different kinds of amino acids (20) and they can fold into complex tertiary structures with binding pockets for the substrate and small molecule co-factors and an active site for catalysis. Now we know that RNAs, which as we have seen can similarly adopt complex tertiary structures, can also be biological catalysts (see Interactive Animation 5-1). Such RNA enzymes are known as **ribozymes**, and they exhibit many of the features of a classical enzyme, such as an active site, a binding site for a substrate, and a binding site for a co-factor, such as a metal ion.

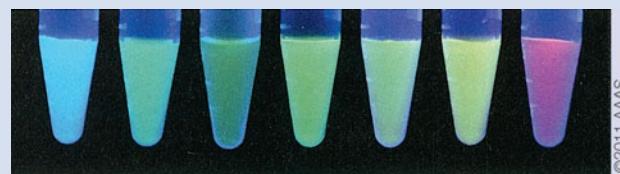
One of the first ribozymes to be discovered was **RNase P**, an endoribonuclease that is involved in generating tRNA molecules from larger, precursor RNAs. Specifically, RNase P cleaves off a leader segment from the 5' end of

## ► TECHNIQUES

### Box 5-2 Creating an RNA Mimetic of the Green Fluorescent Protein by Directed Evolution

One of the most useful proteins in molecular biology is the green fluorescent protein (GFP), which emits green light when excited by ultraviolet light. The green fluorescent protein, which self-generates a covalently bound fluorophore, was discovered in the jellyfish *Aequorea victoria*. The ability to express functional GFP in a variety of organisms, ranging from bacteria to fish and mice, has enabled researchers to use GFP for numerous scientific applications. For example, the protein can be used as a reporter for gene expression in living cells and even for visualizing the locations within cells of proteins to which it is fused. As we have seen, the enormous diversity of RNA, its capacity to fold up into complex tertiary structures, and the invention of SELEX have made it possible to create tailor-made aptamers with many kinds of useful properties. A particularly striking recent application of SELEX is the creation of RNA aptamers that bind to small molecule fluorophores to create mimetics of GFP, developed by Jeremy Paige and colleagues (see Paige et al. 2011. *Science* 333: 642–646). These RNA–fluorophore complexes are similar in color and brightness to GFP but with additional useful features as we explain.

The self-generated fluorophore in GFP is 4-hydroxybenzylidene imidazolinone, whose fluorescence is enabled by specific contacts with the protein moiety of GFP by suppressing intramolecular movement, which prevents fluorescence in the free fluorophore. As a starting point, Paige et al. used a derivative (3,5-dimethoxy-4-hydroxybenzylidene imidazolinone) of the natural fluorophore in GFP that similarly requires suppression of intramolecular movement in order to fluoresce. Ten rounds of SELEX yielded RNA molecules that bound to 3,5-dimethoxy-4-hydroxybenzylidene imidazolinone that had been immobilized on agarose. One of the evolved RNAs, called Spinach, induced bright green fluorescence when bound to the free fluorophore. Using other fluorophores and additional rounds of SELEX, the authors generated RNA–fluorophore complexes exhibiting a palette of colors ranging from blue to red (Box 5-2 Fig. 1). As a demonstration of the utility of these aptamers, Paige et al. fused the Spinach-coding sequence to the 3' end of the gene for 5S RNA, a noncoding component

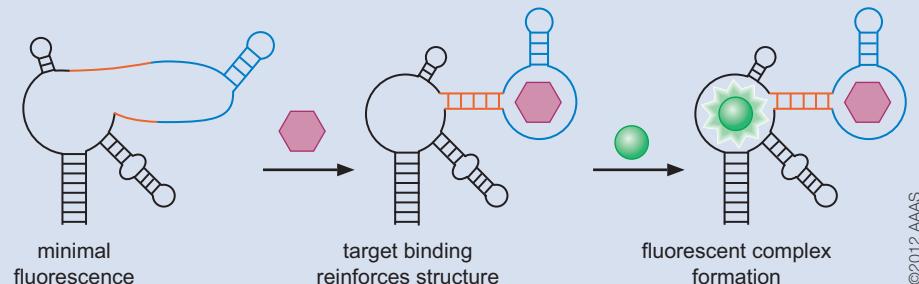


**BOX 5-2 FIGURE 1** RNA–fluorophore complexes exhibiting a range of different colors. (Reproduced, with permission, from Paige J.S. et al. 2011. *Science* 333: 642–646, Fig. 2D. © AAAS.)

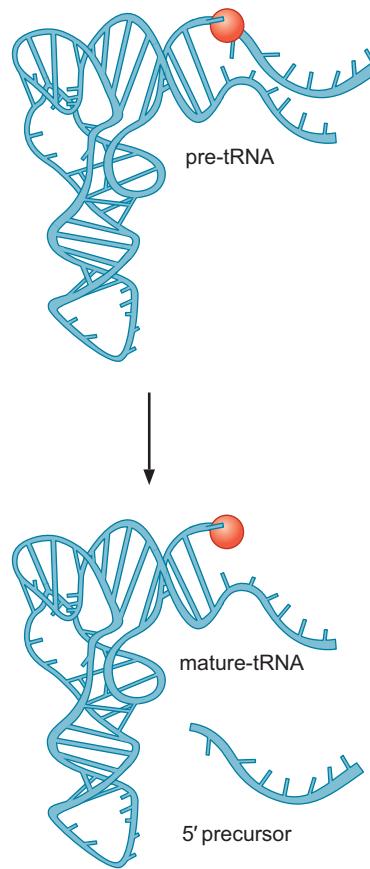
of the large subunit of the ribosome, which is synthesized by RNA polymerase III (see Chapter 15). Using the 5S-Spinach construct, the authors were able to visualize the movement of the ribosomal RNA from the nucleus into the cytosol of the cell.

More recently, Spinach has been further modified to serve as a sensor for cellular metabolites. This was accomplished by joining Spinach to an aptamer that binds a metabolite, such as S-adenosyl methionine (SAM) or ATP (Box 5-2 Fig. 2). The resulting sensor RNA, which is composed of Spinach and a metabolite binding-domain, is designed such that it is unable to bind the fluorophore unless the structure is also stabilized by the binding of the metabolite. One such sensor RNA was used to image SAM in living *E. coli* cells. Sensor-containing cells that had been deprived of methionine (a biosynthetic precursor for SAM) were treated with the fluorophore, followed by addition of methionine to the growth medium, resulting in a marked stimulation in cellular fluorescence. Sensors of this kind may provide a powerful new way to monitor changes in the levels of metabolites in real time in living cells.

These examples underscore the remarkable versatility of RNA. Nature has exploited this versatility in natural selection to create riboswitches, ribozymes, tRNA molecules, and regulatory RNAs, which we shall consider in Chapters 15 and 20. But now molecular biologists are also beginning to exploit this versatility to create a wide variety of RNA molecules that promise to be useful for humanity.



**BOX 5-2 FIGURE 2** Using SELEX to create metabolite sensors. The metabolite sensor contains bonding domains for a fluorophore (shown in green) and a metabolite (shown in purple). A stable, fluorescent complex is only formed when the RNA has bound both small molecules. (Adapted, with permission, from Paige J.S. et al. 2012. *Science* 335: 1194, Fig. 1A. © AAAS.)



**FIGURE 5-11** RNase P cleaves a segment of RNA from the 5' end of a precursor to tRNA molecules.

the precursor RNA in helping to generate the mature and functional tRNA, as depicted in Figure 5-11. RNase P is composed of both RNA and protein; however, the RNA moiety alone is the catalyst. The protein moiety of RNase P facilitates the reaction by shielding the negative charges on the RNA so that it can bind effectively to its negatively charged substrate. The RNA moiety is able to catalyze cleavage of the tRNA precursor in the absence of the protein if a small, positively charged counterion, such as the peptide spermidine, is used to shield the repulsive, negative charges.

As we see below, tRNA molecules fold into an L-shaped tertiary structure with the 5' and 3' termini of the molecule at one end. The crystal structure of RNase P reveals that the site of cleavage of the phosphodiester backbone is located within the catalytic center of the RNaseP RNA moiety, with the protein moiety interacting with the leader (Fig. 5-12).

An example of an RNA that is both a ribozyme and a riboswitch is a structure found in the 5'-untranslated region of the mRNA for the protein enzyme GlmS. GlmS catalyzes the synthesis of the metabolite glucosamine-6-phosphate. The RNA is a ribozyme that degrades the mRNA for GlmS, but the activity of the ribozyme is dependent on glucosamine-6-phosphate; hence, it is also a riboswitch. Thus, when glucosamine-6-phosphate levels are high, the mRNA is degraded, curtailing synthesis of the metabolite.

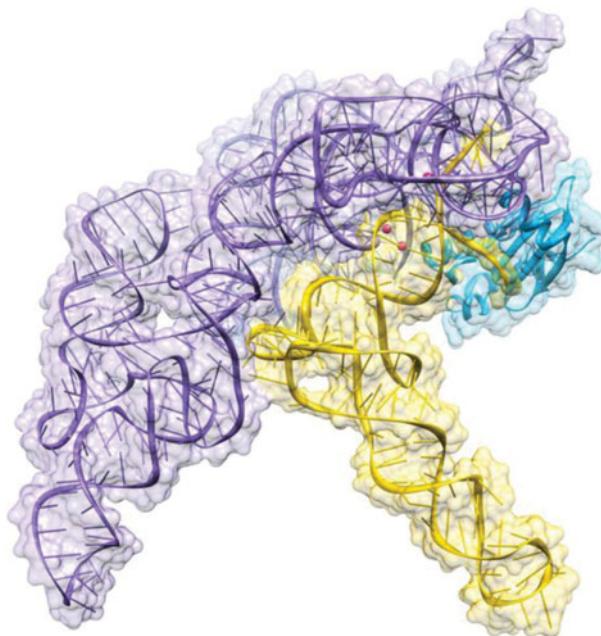
Other ribozymes perform trans-esterification reactions involved in the removal of intervening sequences known as **introns** from precursors to certain mRNAs, tRNAs, and ribosomal RNAs (rRNAs) in a process known as **RNA splicing** (see Chapter 14).

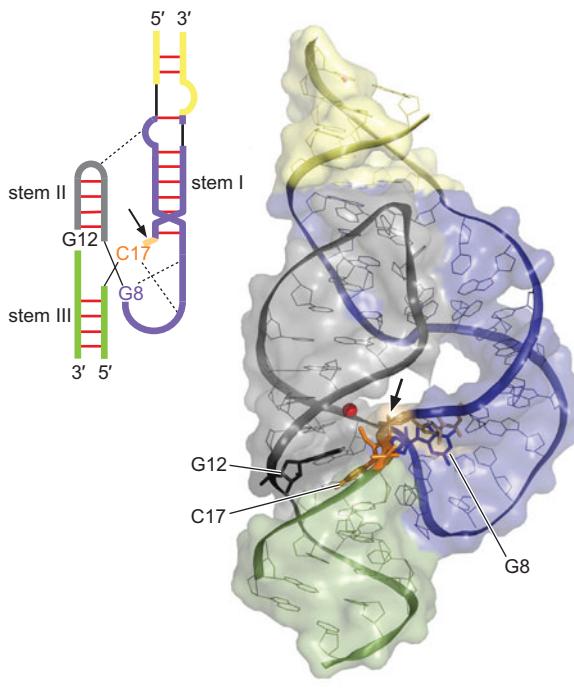
### The Hammerhead Ribozyme Cleaves RNA by the Formation of a 2', 3' Cyclic Phosphate

Let us look in more detail at the structure and function of one particular ribozyme, the **hammerhead** (see Structural Tutorial 5-1). The hammerhead is a sequence-specific ribonuclease that is found in certain infectious RNA agents of plants known as **viroids**, which depend on self-cleavage to propagate. When the viroid replicates, it produces multiple copies of itself in one



**FIGURE 5-12** Structure of RNase P. The crystal structure of a bacterial ribonuclease P holoenzyme, illustrated here, shows the RNA subunit (in purple) and the protein subunit (in blue) in complex with tRNA (in yellow). Metal ions in the catalytic center are shown as small red spheres. (This structure, assembled from coordinates in the Protein Data Base [PDB: 3Q1R], is based on the description by Reiter N.J. et al. 2010. *Nature* **468**: 784–789.)





**FIGURE 5-13 Structure of the hammerhead ribozyme.** (Upper left) A cartoon of the hammerhead secondary structure with its three stems highlighted in color. (Dotted lines) Watson–Crick base-pair interactions; (orange with arrow) the scissile bond at C17. (Diagonal lines) Extrahelical interactions. (Adapted, with permission, from McKay D.B. and Wedekind J.E. 1999. *The RNA world*, 2nd ed. [ed. Gesteland R.F. et al.], p. 267, Fig. 1A. © Cold Spring Harbor Laboratory Press.) (Right) The 3D structure of the hammerhead with a magnesium (red) in the catalytic center. This view of the structure shows stem I (top right), stem II (middle left), and stem loop III (bottom) with the colors corresponding to those in the cartoon. It is uncertain whether the manganese ion participates in catalysis (site shown in orange with arrow). The site of cleavage is at cytosine 17 (C17). (Image kindly prepared by V. D’Souza using PyMOL, from coordinates in the Protein Data Base [PDB: 20EU] based on the description by Martick M. et al. 2008. *Chem Biol.* 15: 332–342.)

continuous RNA chain. Single viroids arise by cleavage, and this cleavage reaction is performed by the RNA sequence around the junction. One such self-cleaving sequence is called the “hammerhead” because of the shape of its secondary structure, which consists of three base-paired stems (I, II, and III) surrounding a core of noncomplementary nucleotides required for cleavage (Fig. 5-13). The cleavage reaction takes place at cytosine 17. The tertiary structure of the hammerhead shows that the catalytic center is located near the junction of the three stems at the core of the ribozyme (Fig. 5-13).

To understand how the hammerhead works, let us first look at how RNA undergoes hydrolysis under alkaline conditions. At high pH, the 2'-hydroxyl of the ribose in the RNA backbone can become deprotonated, and the resulting negatively charged oxygen can attack the scissile phosphate at the 3' position of the same ribose. This reaction breaks the RNA chain, producing a 2',3' cyclic phosphate and a free 5'-hydroxyl. Each ribose in an RNA chain can undergo this reaction, completely cleaving the parent molecule into nucleotides. (Why is DNA not similarly susceptible to alkaline hydrolysis?) Many protein ribonucleases also cleave their RNA substrates via the formation of a 2',3' cyclic phosphate. Working at normal cellular pH, these protein enzymes use a metal ion, bound at their active site, to activate the 2'-hydroxyl of the RNA. The hammerhead also cleaves RNA via the formation of a 2',3' cyclic phosphate. Interestingly, the three-dimensional (3D) structure reveals a magnesium ion near the catalytic center, but the exact mechanism of the cleavage reaction and the significance of the metal ion are not yet understood.

Because the normal reaction of the hammerhead is self-cleavage, it is not really a catalyst; each molecule normally promotes a reaction one time only, thus having a turnover number of 1. But the hammerhead can be engineered to function as a true ribozyme by dividing the molecule into two portions—one, the ribozyme, that contains the catalytic core; and the other, the substrate, that contains the cleavage site. The substrate binds to the ribozyme at stems I and III. After cleavage, the substrate is released and replaced by a fresh uncut substrate, thereby allowing repeated rounds of cleavage.

## A Ribozyme at the Heart of the Ribosome Acts on a Carbon Center

All of the examples considered so far are ribozymes that act on phosphorous centers. But as we see in Chapter 15, one of the RNA components of the ribosome, which was traditionally thought to serve only a structural role, is now known to be the enzyme **peptidyl transferase**, which is responsible for peptide-bond formation during protein synthesis. In this case, the ribozyme acts on a carbon center rather than a phosphorous center in catalyzing the reaction. In addition, and as we consider in Chapter 17, the discovery that one of the most fundamental enzymatic reactions in living cells is catalyzed by an RNA molecule has been taken as support for the hypothesis that contemporary, protein-based life arose from an earlier RNA World.

## SUMMARY

RNA differs from DNA in the following ways: Its backbone contains ribose rather than 2'-deoxyribose; it contains the pyrimidine uracil in place of thymine; and it usually exists as a single polynucleotide chain, without a complementary chain. As a consequence of being a single strand, RNA can fold back on itself to form short stretches of double helix between regions that are complementary to each other. RNA allows a greater range of base pairing than does DNA. Thus, as well as A:U and C:G pairing, non-Watson–Crick pairing is also seen, such as U pairing with G. This capacity to form noncanonical base pairs adds to the propensity of RNA to form double-helical segments. Freed of the constraint of forming long-range regular helices, RNA can form complex tertiary

structures, which are often based on unconventional interactions between bases and the sugar–phosphate backbone.

Some RNAs act as enzymes—they catalyze chemical reactions in the cell and *in vitro*. These RNA enzymes are known as **ribozymes**. Most ribozymes act on phosphorous centers, as in the case of the ribonuclease RNase P. RNase P is composed of protein and RNA, but it is the RNA moiety that is the catalyst. The hammerhead is a self-cleaving RNA, which cuts the RNA backbone via the formation of a 2',3' cyclic phosphate. Peptidyl transferase is an example of a ribozyme that acts on a carbon center. This ribozyme, which is responsible for the formation of the peptide bond, is one of the RNA components of the ribosome.

## BIBLIOGRAPHY

### Books

- Bloomfield V.A., Crothers D.M., Tinoco I., Jr., and Heast J.E. 2000. *Nucleic acids: Structures, properties, and functions*. University Science Books, Sausalito, California.
- Gesteland R.F., Cech T.R., and Atkins J.F., eds. 2006. *The RNA world*, 3rd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

### RNA Structure

- Darnell J.E. Jr. 1985. RNA. *Sci Am.* **253**: 68–78.
- Doudna J.A. and Cech T.R. 2002. The chemical repertoire of natural ribozymes. *Nature* **418**: 222–228.

- Doudna J.A. and Lorsch J.R. 2005. Ribozyme catalysis: Not different, just worse. *Nat. Struct Mol Biol.* **15**: 394–402.
- Houck-Loomis B., Durney M.A., Salguero C., Shankar N., Nagle J.M., Goff S.P., and D'Souza V.M. 2011. An equilibrium-dependent retroviral mRNA switch regulates translational recoding. *Nature*. **480**: 561–564.
- Nelson J.A. and Uhlenbeck O.C. 2006. When to believe what you see. *Mol Cell.* **23**: 447–450.
- Kladwang W., VanLang C.C., Cordero P., and Das R. 2011. A two-dimensional mutate-and-map strategy for non-coding RNA structure. *Nat. Chem.* **3**: 954–962.

## QUESTIONS

**MasteringBiology**®

For instructor-assigned tutorials and problems, go to *MasteringBiology*.

For answers to even-numbered questions, see Appendix 2: Answers.

### Question 1.

- A. Draw the structure of deoxyguanosine. Circle and label the appropriate hydrogen bond donors (D) or acceptors (A) that

participate in hydrogen bonding when a base pair with deoxyctydine forms in DNA.

- B. Draw the structure of guanosine. Circle and label the appropriate hydrogen bond donors (D) or acceptors (A) that participate in hydrogen bonding when a base pair with uridine forms in RNA.

**Question 2.** Researchers discovered a new virus and characterized its genome by determining the base composition and the percentage of each base. You want to categorize the virus by its genetic material. Remember that viruses may contain single-stranded DNA or RNA or double-stranded DNA or RNA as the genetic material. Given just the sequence information, propose a way to distinguish if the genetic material is RNA or DNA. Propose a way to distinguish if the genetic information is single-stranded or double-stranded.

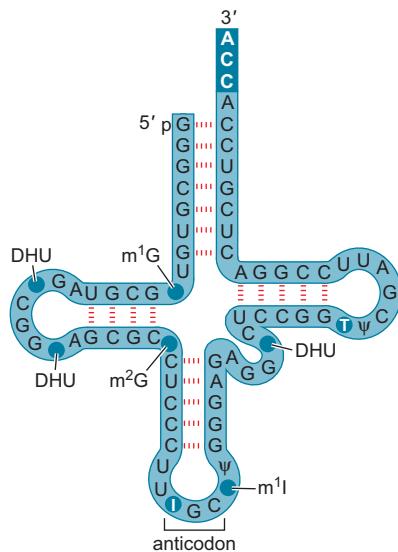
**Question 3.** Explain why the helical structure of DNA differs from the helical structure of double-stranded RNA and how that difference in structure affects the ability of proteins to interact with helical RNA.

**Question 4.** Justify the biological reason for the presence of uracil in RNA but not in DNA.

**Question 5.** Given the following sequence of RNA, propose the potential hairpin structure for this RNA. Indicate base pairing with a dotted line.

5'-AGGACCCUUCGGGGUUCU-3'

**Question 6.** The yeast alanine tRNA is shown below (adapted from Chapter 2, Fig. 2-14). Identify the type(s) of secondary structure present in this tRNA. Identify any noncanonical base pairing.



**Question 7.** Researchers want to find a way to confer antibiotic resistance to selected bacterial cells. To do so, they decide to identify a sequence of RNA (aptamer) by SELEX that specifically binds to the antibiotic of interest and expresses that RNA aptamer in the selected cells to confer antibiotic resistance.

- Given what you know about RNA, hypothesize what general properties of the RNA aptamer allow for specific binding to the antibiotic.
- Instead of using SELEX, the researchers could start with an isolated pool of RNAs encoded by the bacterial genome and select for the RNA(s) that bind the antibiotic to identify their

desired RNA. Give two advantages of using SELEX rather than this method.

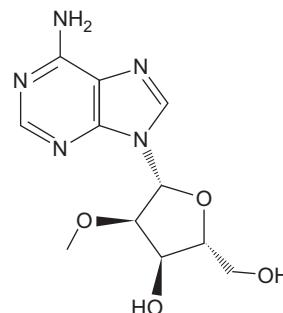
**Question 8.** Describe the properties shared between an enzyme and a ribozyme.

**Question 9.** Some RNAs act as plant pathogens or viroids. Some viroids are capable of carrying out a catalytic reaction. You are given a sequence of RNA isolated as a viroid to a specific plant species. Identify some characteristics that would indicate that this viroid acts as a catalytic hammerhead.

**Question 10.** Scientists design modified versions of the self-cleaving hammerhead as potential therapeutics. To target cleavage of another RNA molecule, describe how the engineered structure of the hammerhead is modified. Why does this change make the hammerhead a true ribozyme?

**Question 11.** Some ribozymes like those found in the hepatitis delta virus and the ribozyme/riboswitch found in the *glms* mRNA are capable of folding into a nested double pseudoknot. Discuss several advantages to pseudoknot formation in ribozymes and riboswitches.

**Question 12.** The structure given below is a nucleoside analog.



- Name the nucleoside that the structure mimics.
- Imagine a strand of RNA containing this nucleoside analog (linked by phosphodiester bonds as in normal RNA). In a high-pH environment, is the RNA strand subject to hydrolysis where the nucleoside analog is present? Explain why or why not.

**Question 13.** DNA serves as the genetic material. Give three cellular roles for RNA.

**Question 14.** You find a complex composed of a protein moiety and an RNA moiety that is capable of cleaving an RNA substrate. You want to know which moiety is responsible for catalysis. You perform an *in vitro* cleavage assay in the proper buffer conditions containing a low concentration of Mg<sup>2+</sup>. The table below summarizes the results from performing seven separate reactions. The + symbol indicates the included reaction components. Spermidine is a positively charged peptide not specific to this reaction.

Protein moiety	-	+	-	+	+	-	-	+
RNA moiety	-	-	+	+	-	+	+	+
Spermidine	-	-	-	-	+	+	+	+
Percent cleavage	0	0	0	90	0	50	90	

- A. Does the protein or RNA moiety act as the catalyst in this reaction? Explain which reactions helped you make your conclusion.
- B. Propose the function of spermidine in the last two reactions.

**Question 15.** The genome of the bacteriophage Q $\beta$  consists of about 4000 nucleotides of single-stranded RNA. Inside the *E. coli* host, replication of this genome requires a RNA-dependent RNA polymerase made up of phage and bacterial proteins. Interestingly, the replicase binds to a region in the center of the RNA genome, yet must start copying the 3' end of the RNA template to produce a new strand of RNA in the 5' to 3' direction. Researchers hypothesized that the presence of a predicted pseudoknot in the Q $\beta$  genome allowed the replicase to gain access to the 3' end of the RNA. To test their hypothesis, they measured the replication efficiency *in vitro* using wild-type replicase and different versions of the Q $\beta$  RNA containing mutations in a region key to the predicted pseudoknot formation. They specifically focused on an eight-nucleotide region from the center of the RNA that was complementary to the 3'-terminal hairpin. The wild-type and mutant sequences and replication data are given below.

Wild type 5'-UAAAGCAG-3'  
GUUUCGUC

Mutant A 5'-UAAACGAG-3'  
GUUUCGUC  
Mutant B 5'-UAAAGCAG-3'  
GUUUGCUC  
Mutant C 5'-UAAACGAG-3'  
GUUUGCUC

In vitro replication efficiency (relative to wild type)

Wild type: 100%

Mutant A: 1.6%

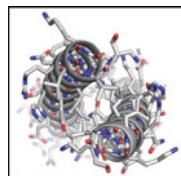
Mutant B: 0.6%

Mutant C: 42%

- A. Predict why the replication efficiency so low in Mutant A.
- B. Predict why the replication efficiency is restored to almost half of wild-type levels in Mutant C?
- C. Do you think the results support or refute the hypothesis that the presence of a pseudoknot affects replication?

Data adapted from Klovins and van Duin (1999. *J. Mol. Biol.* 294: 875–884.)

CHAPTER 6



# The Structure of Proteins

PROTEINS ARE POLYMERS. THAT IS, THEY ARE molecules that contain many copies of a smaller building block, covalently linked. The building blocks of proteins are  $\alpha$ -amino acids, of which there are 20 that occur regularly in the proteins of living organisms and that are specified by the genetic code. Some of these amino acids can undergo modification when already part of a protein, so the actual variety in proteins isolated from cells or tissues is somewhat greater.

## THE BASICS

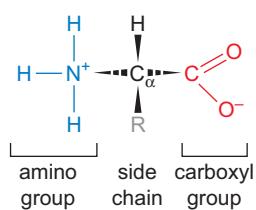
### Amino Acids

The  $\alpha$ -carbon ( $C_\alpha$ ) of an amino acid has four substituents (Fig. 6-1), distinct from each other except in the case of the simplest amino acid, glycine. An amino group, a carboxyl group, and a proton are three of these substituents on all of the naturally occurring amino acids. The fourth, often symbolized by R and sometimes called the “R group,” is the only distinguishing feature. The R-group is also called the “side chain,” for reasons that will be clear in the next section. Because its four substituents are distinct (except for glycine), the  $C_\alpha$  is a chiral center. Amino acids that occur in ordinary proteins all have the L-configuration at that center; D-amino acids are present in other kinds of molecules (including small protein-like polypeptides in microorganisms).

The properties of its R group determine the specific characteristics of an amino acid. The polarity of the group, which correlates with its solubility in water, is one critical property; size is another. It is useful to cluster the R groups of the 20 genetically encoded amino acids into the following categories: (1) neutral (i.e., uncharged) and nonpolar; (2) neutral and polar; and (3) charged (Fig. 6-2). The size (volume) of the side chain is of particular consequence for nonpolar amino acids because, as we shall see later, these side chains pack into the compact interior of a protein, and therefore the functional roles in proteins of glycine and alanine are quite different from those of phenylalanine and tryptophan. Note also that tryptophan, although largely nonpolar, has a hydrogen-bonding group that gives it a degree of polar character, and that tyrosine, although classified in Figure 6-2 as polar

## OUTLINE

- The Basics, 121
  - Importance of Water, 125
  - Protein Structure Can Be Described at Four Levels, 126
  - Protein Domains, 130
  - From Amino-Acid Sequence to Three-Dimensional Structure, 134
  - Conformational Changes in Proteins, 136
  - Proteins as Agents of Specific Molecular Recognition, 137
  - Enzymes: Proteins as Catalysts, 141
  - Regulation of Protein Activity, 142
- Visit Web Content for Structural Tutorials and Interactive Animations

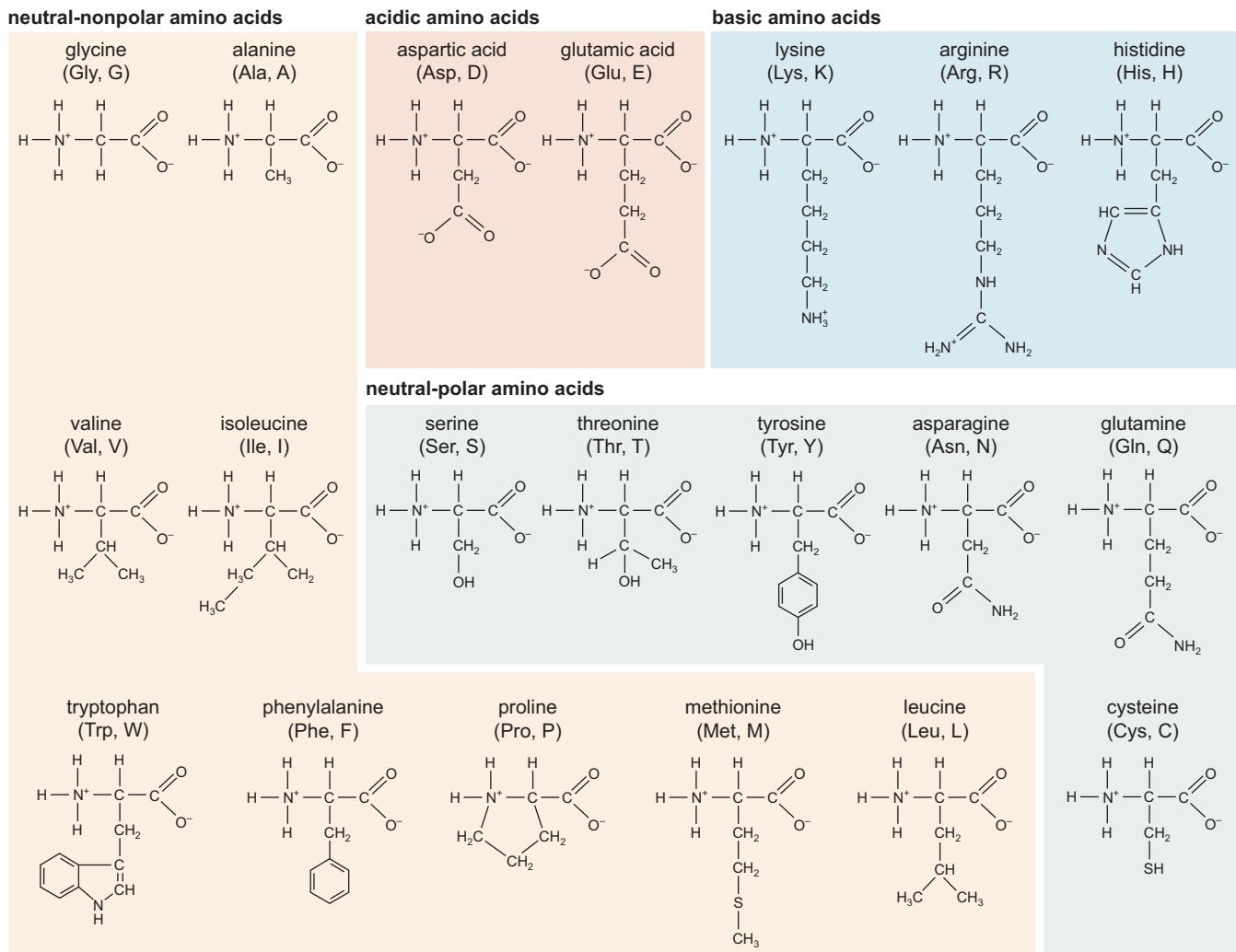


**FIGURE 6-1** Structural features of an amino acid.

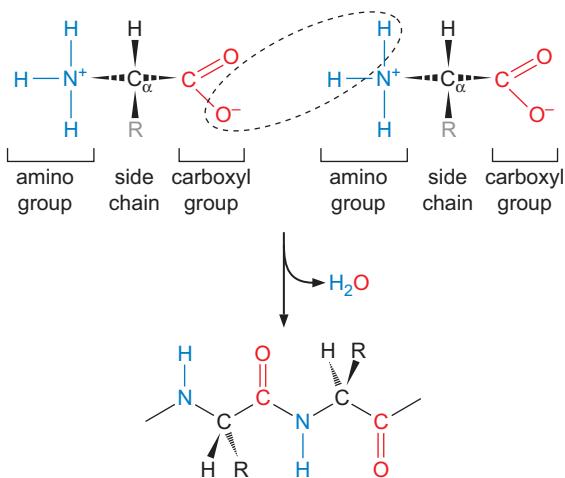
because of its OH group, is much less so than serine. In short, the boundaries between groups are less sharp than nomenclature might imply. The charged R groups are either negatively charged at neutral pH (aspartic acid and glutamic acid) or positively charged at neutral pH (lysine, arginine, and histidine). The  $\text{pK}_a$  of histidine is  $\sim 6.5$ , so even at neutral pH, histidine loses most of its charge. This property is particularly important for its role at the catalytic site of many enzymes.

### The Peptide Bond

**Peptide bonds** are the covalent links between amino acids in a protein. A peptide bond forms by a condensation reaction, with elimination of a water molecule (Fig. 6-3a). It is a special case of an amide bond. Each amino acid can form two such bonds, so that successive links of the same kind can create a linear (i.e., unbranched) **polypeptide chain**. Because formation of each peptide bond includes elimination of a water, the components of the chain are known as **amino acid residues**, or sometimes just “residues” when “amino acid” is evident from context. The peptide bond has partial



**FIGURE 6-2** The 20 naturally occurring amino acids in proteins. Commonly used abbreviations for amino acids, including the single-letter code, are shown in parentheses.

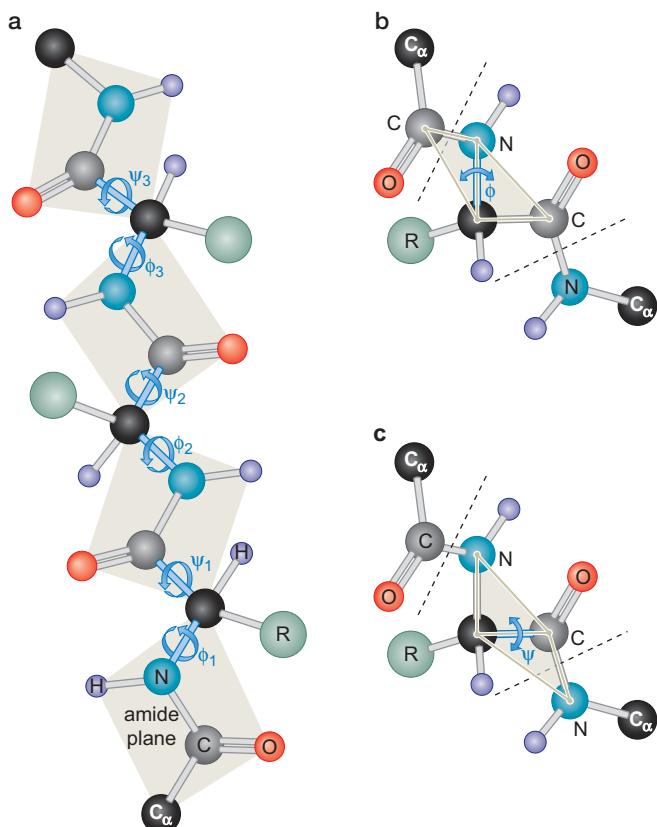


**FIGURE 6-3** Peptide-bond formation.

double-bond character; the carbonyl and amide components are nearly coplanar and almost always in a *trans* configuration (Fig. 6-4).

### Polypeptide Chains

The word **conformation** describes an arrangement of chemically bonded atoms in three dimensions, and we therefore speak of the conformation of a polypeptide chain, or more simply, of its “folded structure” or fold. If we follow the sequence of covalent linkages along a polypeptide chain,



**FIGURE 6-4** The backbone torsion angles  $\phi$  and  $\psi$ . (a) This diagram shows the swivel points of the peptide backbone. (b) The  $\phi$  torsion angle corresponds to the rotation about the  $\text{N}-\text{C}_{\alpha}$  bond; here the conformation corresponds to a value of  $\phi = 180^\circ$ . (c) The  $\psi$  torsion angle corresponds to the rotation about the  $\text{C}_{\alpha}-\text{C}$  bond; the conformation shown here represents  $\psi=0^\circ$ . (Adapted, with permission, from Kuriyan J. et al. 2012. *The molecules of life*. © Garland Science/Taylor & Francis LLC; Branden C. and Tooze J. 1999. *An introduction to protein structure*, p. 8, Fig. 1.6. © Garland Science/Taylor & Francis LLC.)

there are three bonds per amino acid residue—one that joins the NH group to the  $C_\alpha$ , another that joins the  $C_\alpha$  to the carbonyl, and finally the peptide bond to the next residue in the chain. The first two are single bonds with relatively free torsional rotation about them (Fig. 6-4). But the peptide bond has very little rotational freedom, because of its partial double-bond character. The polypeptide chain conformation is therefore specified by the values of the torsion angles about the first two backbone bonds of each residue, plus the torsion angles for each single bond in each side chain. The two backbone angles are conventionally denoted  $\phi$  and  $\psi$ . Many combinations of these two angles lead to atomic collisions, restricting the conformational freedom of a polypeptide chain to certain ranges of each angle (see Box 6-1, Ramachandran Plot).

### Three Amino Acids with Special Conformational Properties

Glycine, proline, and cysteine have special properties. Because its R group is just a proton, glycine is not chiral, and it has much more conformational freedom than any other amino acid. Conversely, proline, in which the side chain has a covalent bond with N as well as  $C_\alpha$  (making it, strictly speaking, an *imino* acid), has less conformational freedom than many other amino acids. Moreover, absence of the hydrogen-bonding potential of an NH group restricts its participation in secondary structures (see the section Protein Structure Can Be Described at Four Levels).

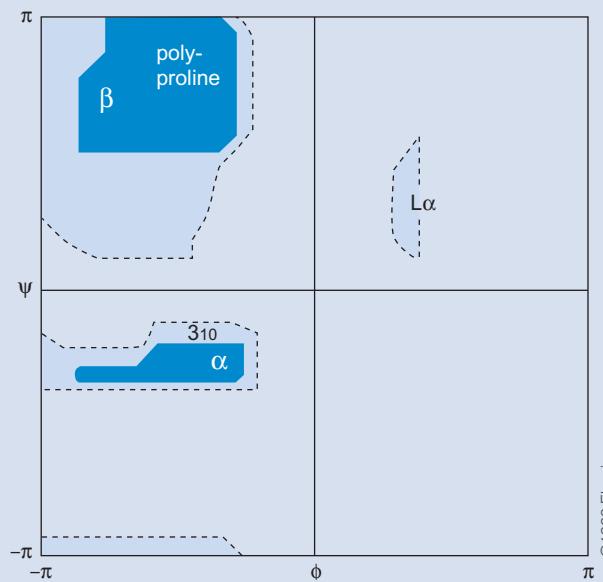
Cysteine, with a sulfhydryl group ( $-SH$ ) on its side chain, is the one amino acid that is sensitive to oxidation–reduction under roughly physiological conditions. Two cysteines, correctly positioned across from each other in a folded protein, can form a **disulfide bond** by oxidation of the two  $-SH$  groups to S—S (Fig. 6-5). (The resulting pair of amino acids, linked by the S—S covalent bridge, is sometimes called *cystine*.) Proteins on the cell surface and proteins secreted into the extracellular space are exposed

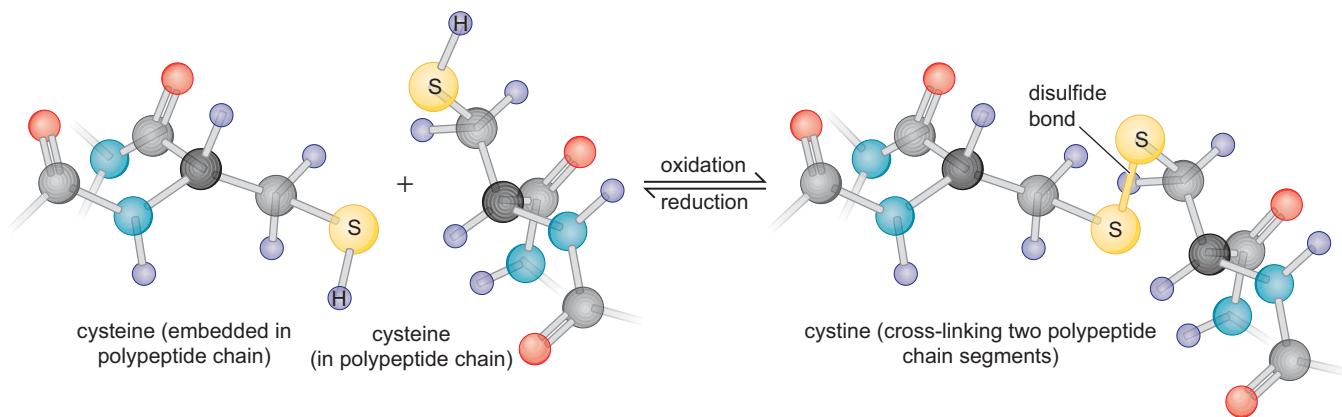
#### ► ADVANCED CONCEPTS

##### Box 6-1 Ramachandran Plot: Permitted Combinations of Main-Chain Torsion Angles $\phi$ and $\psi$

G.N. Ramachandran and coworkers (1963) studied all possible combinations of the torsion angles  $\phi$  and  $\psi$  (shown in Fig. 6-4) and determined which combinations avoided atomic collisions (“allowed”) and which combinations led to clashes (“forbidden”). The two-dimensional plot that shows the allowed and forbidden combinations is now known as a “Ramachandran plot” (Box 6-1 Fig. 1). The backbone conformations of regular secondary structures have the  $\phi$  and  $\psi$  values indicated: right-handed  $\alpha$  helix;  $\beta$  strand; polyproline helix (a threefold screw structure adopted preferentially by continuous stretches of proline);  $3_{10}$  helix (a helix with 3.3 residues per turn, closely related to the  $\alpha$  helix, which has 3.6 residues per turn); and left-handed  $\alpha$  helix,  $L\alpha$  (permitted for glycine only, because it has no side chain).

**BOX 6-1 FIGURE 1** The Ramachandran plot. The “allowed” areas are shown shaded in blue. (Modified, with permission, from Ramachandran G.N., et al. 1963. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* 7: 95–99.)





**FIGURE 6-5** Formation of the disulfide bond. (Adapted, with permission, from Kuriyan J. et al. 2012. *The molecules of life*. © Garland Science/Taylor & Francis LLC.)

to an environment with a redox potential that favors disulfide formation; most such proteins have disulfide bonds and no unoxidized cysteines. Living cells maintain a more reducing internal environment, and intracellular proteins very rarely have disulfide bonds.

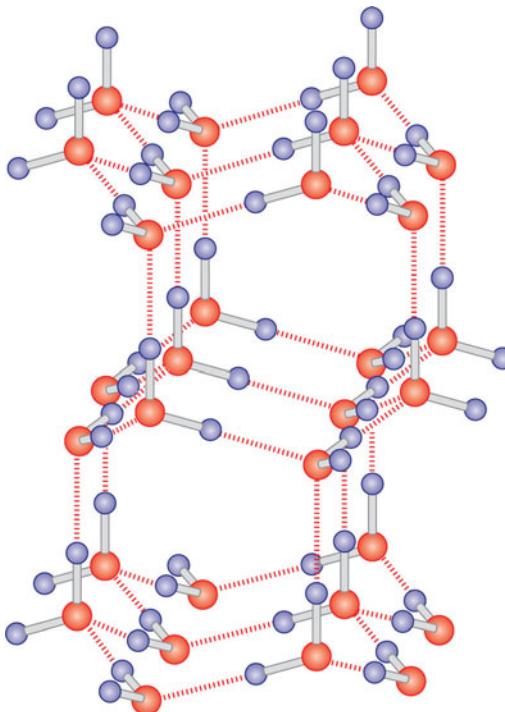
Disulfide bonds enhance the stability of a folded protein by adding covalent cross-links. They are particularly critical for stabilizing small, secreted proteins, such as some hormones, and for reinforcing the extracellular domains of membrane proteins, which face a much less controlled environment than do proteins that remain in the cell interior.

## IMPORTANCE OF WATER

All molecular phenomena in living systems depend on their aqueous environment. The importance of the distinction between polar and nonpolar amino acid side chains comes from their properties with respect to water as a solvent. Compare the side chains of aspartic acid and phenylalanine, which resemble acetic acid and toluene, respectively, linked to the peptide main chain. Acetic acid is very soluble in water; toluene is very insoluble. An aspartic acid side chain is therefore called **hydrophilic**, and a phenylalanyl side chain **hydrophobic**. Even hydrophilic side chains can have hydrophobic parts (e.g., the three methylene groups of a lysyl side chain).

Water is an extensively hydrogen-bonded liquid (Fig. 6-6). Each water molecule can donate two hydrogen bonds and accept two hydrogen bonds. The way in which a solute affects the hydrogen bonding of the surrounding water determines its hydrophilic or hydrophobic character. Hydrophobic molecules perturb the network of hydrogen bonds; hydrophilic molecules participate in it. Thus, it is more favorable for hydrophobic molecules to remain adjacent to each other (insolubility) than to disperse into an aqueous medium (solubility). The hydrophobic character of many amino acid side chains makes it favorable for them to cluster away from water, and the hydrophilic character of others allows them to project into water. The sequence of amino acids in a real protein has evolved so that these tendencies cause the polypeptide chain to fold up, sequestering residues of the former kind and exposing residues of the latter. Many of the huge number of possible sequences for an average-sized polypeptide chain cannot fold up in this way—if

**FIGURE 6-6** Water: the hydrogen-bonded structure of ice. In ice, each water molecule donates two hydrogen bonds (from its two protons [lavender]) to lone-pair electrons on an oxygen of its neighbor (red) and accepts two hydrogen bonds at lone pairs of its own oxygen. The hydrogen bonds are shown as dashed red lines. When ice melts, the network of hydrogen bonds fluctuates and breaks apart transiently, but individual water molecules retain (on average) most of the four hydrogen bonds with their neighbors. Thus, the structure of liquid water resembles a fluctuating and distorted version of the ice lattice shown here.

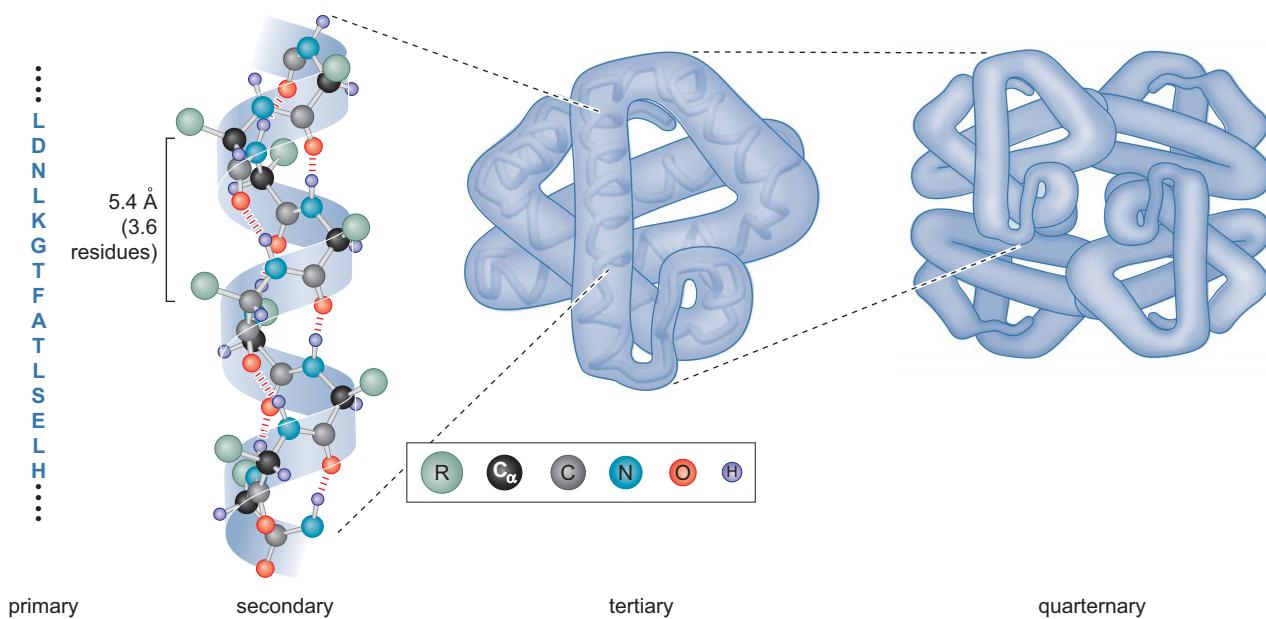


made, they either remain as randomly fluctuating, extended chains in solution (sometimes called **random coils**) or else they aggregate because the hydrophobic groups on one polypeptide chain cluster together with hydrophobic groups on other chains.

## PROTEIN STRUCTURE CAN BE DESCRIBED AT FOUR LEVELS

In analyzing and describing the structure of proteins, it is useful to distinguish four levels of organization (Fig. 6-7). The first level, the **primary structure** of a protein, is simply the sequence of amino acid residues in the polypeptide chain. As we have seen, the genetic code specifies the primary structure of a protein directly. The primary structure is thus just a one-dimensional (1D) string, specifying a pattern of chemical bonds; the remaining three levels depend on a protein's three-dimensional (3D) characteristics.

The **secondary structure** of a protein refers to the *local* conformation of its polypeptide chain—the 3D arrangement of a short stretch of amino acid residues. There are two very regular secondary structures found frequently in naturally occurring proteins, because these two local conformations are particularly stable ones for a chain of L-amino acids (Box 6-1). One of these is called the  **$\alpha$  helix** (Fig. 6-8a). The polypeptide backbone spirals in a right-handed sense around a helical axis, so that hydrogen bonds form between the main-chain carbonyl group of one residue and the main-chain amide group of a residue four positions further along in the chain. The other regular conformation is called a  **$\beta$  strand** (Fig. 6-8b). It is an extended conformation, in which the side chains project alternately to either side of the backbone, and the amide and carbonyl groups project laterally, also alternating. The backbone is not quite fully stretched, so that the strand has a slightly zigzag or pleated character. In folded proteins,  $\beta$  strands form

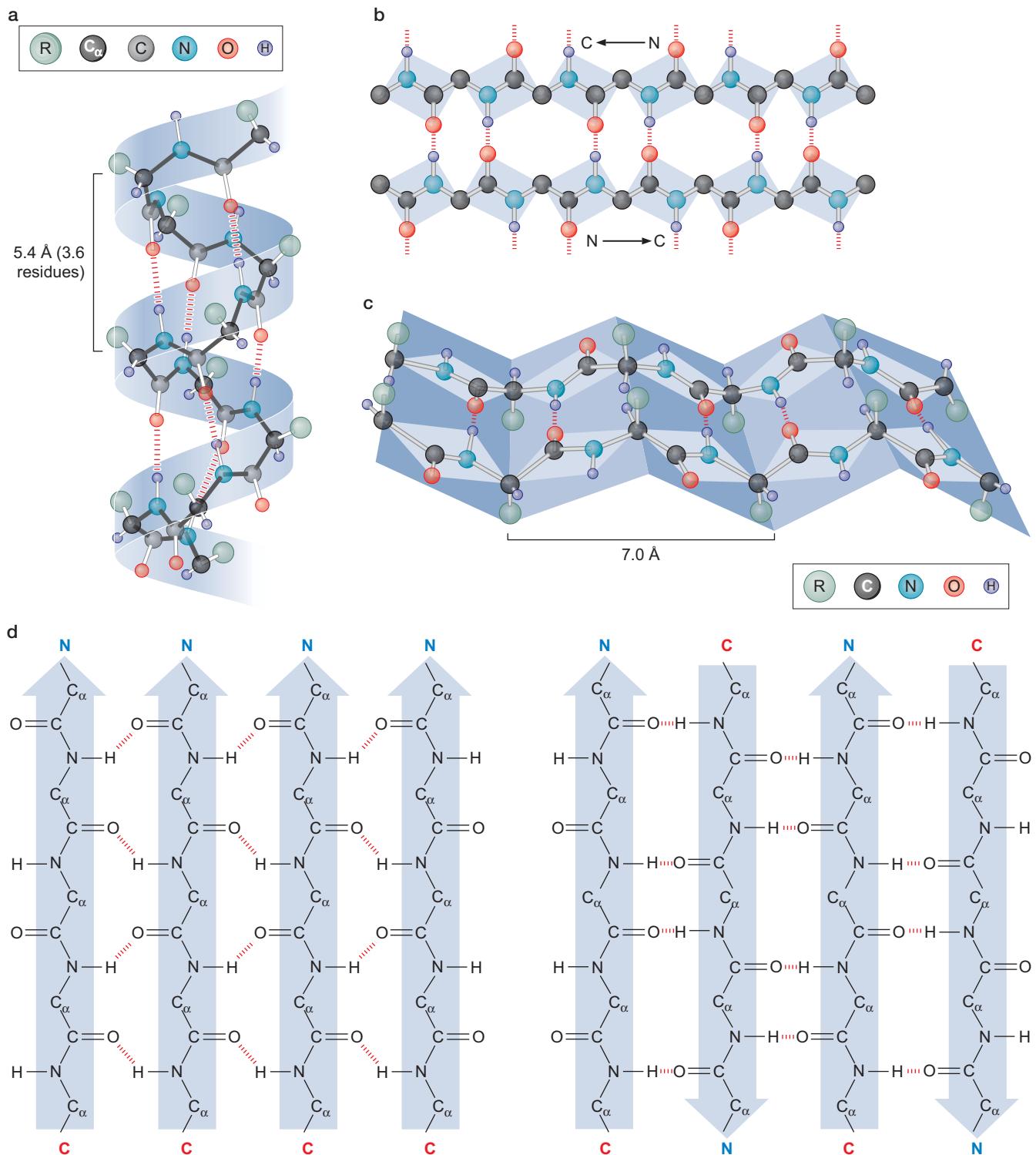


**FIGURE 6-7** Levels of protein structure, illustrated by hemoglobin. “Primary structure” refers to the sequence of amino acids in the polypeptide chain. The primary structure of a segment of a hemoglobin subunit is shown in single-letter code. “Secondary structure” refers to regular local structures, with repeated backbone hydrogen bonds. Shown here is a part of one of the long  $\alpha$  helices from the hemoglobin subunit. “Tertiary structure” refers to the folded structure of an entire polypeptide chain (or of a single domain of a multidomain protein). The drawing shows one of the four hemoglobin protein subunits. Dashed lines demarcate the segment of  $\alpha$  helix corresponding to the primary and secondary structures shown to the left. “Quaternary structure” refers to the arrangement of multiple protein subunits in a larger complex. Hemoglobin is a tetramer of two “ $\alpha$  chains” and two “ $\beta$  chains,” but the two kinds of chain have very similar tertiary structures, as can be seen in the drawing. (Modified from an illustration by Irving Geis. Rights owned by the Howard Hughes Medical Institute.)

sheets joined by main-chain hydrogen bonds. Either parallel or antiparallel hydrogen-bonding patterns are possible, sometimes called parallel or antiparallel  $\beta$ -pleated sheets, respectively. In real proteins, various mixed sheets are often found—rather than either strictly alternating strand directions or strictly unidirectional ones.

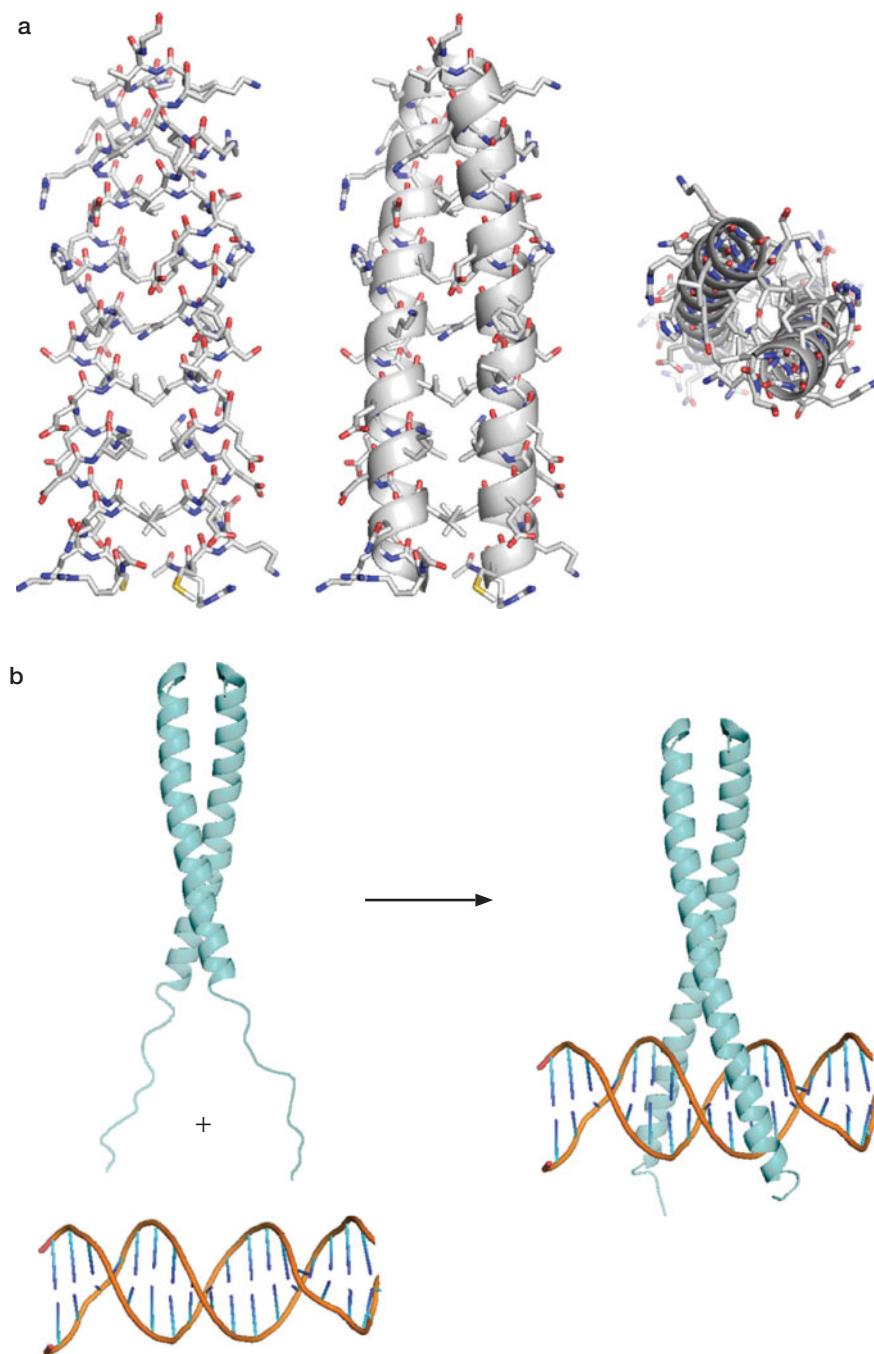
The **tertiary structure** of a protein refers to the usually compact, three-dimensionally folded arrangement that the polypeptide chain adopts under physiological conditions. Segments of the chain may be  $\alpha$  helices or  $\beta$  strands; the rest have less regular conformations (e.g., turns or loops between secondary-structure elements that allow these elements to pack tightly against each other). We will outline ways to describe and classify possible tertiary structures in a subsequent section. Usually, the stabilities of the secondary and tertiary structures of a polypeptide chain depend on each other.

Many proteins are composed of more than one polypeptide chain: **quaternary structure** refers to the way individual, folded chains associate with each other. We can distinguish cases in which there are a defined number of copies of a single type of polypeptide chain (generally called a “subunit” in this context, or a “protomer”) and cases in which there are defined numbers of each of more than one type of subunit. In simple cases, the way in which the subunits associate does not change how the individual



**FIGURE 6-8** Protein secondary structures. (a) An  $\alpha$  helix. Hydrogen bonds are represented by the series of broken red lines. (b)  $\beta$  sheets. Hydrogen bonds are represented by the series of broken red lines. On the top, a  $\beta$  sheet is shown from above. On the bottom, a  $\beta$  sheet is shown from the side. (c) A parallel  $\beta$  sheet, showing the hydrogen-bonding pattern, in which the chains run in the same amino-to-carboxyl direction. (d) An antiparallel  $\beta$  sheet, showing the hydrogen-bonding pattern, in which the chains run in opposite directions. (a, Modified from illustration by Irving Geis. b,c, Illustrations by Irving Geis. Rights owned by Howard Hughes Medical Institute. Not to be reproduced without permission. d, Adapted, with permission, from Branden C. and Tooze J. 1999. *Introduction to protein structure*, 2nd ed., p. 19, Fig 2.6a and p. 18, Fig 2.5b. © Garland Science/Taylor & Francis LLC.)

polypeptides fold. Often, however, the tertiary or even secondary structures of the components of a protein **oligomer** (i.e., a protein composed of a small number of subunits) depend on their association with each other. In other words, the individual subunits acquire secondary or tertiary structure only as they also acquire quaternary structure. One common example is the  $\alpha$ -helical coiled-coil: two (or sometimes three or even four) polypeptide chains, either identical or different, adopt  $\alpha$ -helical conformations and wrap very gently around each other (Fig. 6-9a). The individual chains are not, in general, stable as  $\alpha$  helices on their own—if the oligomeric interaction is lost, the separated helices unravel into disordered polypeptide chains.



**FIGURE 6-9** The yeast transcription factor GCN4. (a) Three views of the structure of the GCN4 coiled-coil. (Left) Representation that shows chemical bonds as sticks and atoms as junction, with carbon in gray, oxygen in red, and nitrogen in blue. The carboxyl termini of the two identical polypeptide chains are at the top. Note the ladder of hydrophobic side chains (mostly gray) at the interface between the two helices. (Center) Representation with polypeptide backbone as an idealized ribbon and side chains as sticks. Note that the two chains coil very gently around each other. (Right) The same representation as in the center, but viewed end-on from the top. (b) Structure of the GCN4 complex with DNA, illustrating the disorder-to-order transition of the so-called “basic region”—the segment amino-terminal to the coiled-coil, which, upon binding, folds into an  $\alpha$  helix in the major groove of DNA. Images prepared with PyMOL (Schrödinger, LLC).

## PROTEIN DOMAINS

### Polypeptide Chains Typically Fold into One or More Domains

Folding of a polypeptide chain creates an “inside” and an “outside” and thus generates **buried** and **exposed** amino acid side chains, respectively. If the polypeptide chain is too short, there are no conformations that bury enough hydrophobic groups to stabilize a folded structure. If the chain is too long, the complexity of the folding process is likely to generate errors. As a result of these restrictions, most stably folded conformations include between about 50 and 300 amino acid residues. Longer polypeptide chains generally fold into discrete modules known as **domains** (see Box 6-2, Glossary of Terms); each domain generally falls within the 50- to 300-residue range just mentioned. The structures of individual domains of such a protein are similar to the structures of smaller, single-domain proteins (Fig. 6-10a).

Each of the two or more domains of a folded polypeptide chain sometimes contains a continuous sequence of amino acid residues. Often,

#### ► ADVANCED CONCEPTS

##### **Box 6-2** Glossary of Terms

**Primary structure:** Amino acid sequence of a polypeptide chain.

**Secondary structure:** Elements of regular polypeptide-chain structure with main-chain hydrogen bonds satisfied. The secondary structures that occur frequently in proteins are the  $\alpha$  helix and the parallel and antiparallel  $\beta$  sheets.

**Tertiary structure:** The folded, three-dimensional conformation of a polypeptide chain.

**Quaternary structure:** Multi-subunit organization of an oligomeric protein or protein assembly.

**Domain:** A part of a polypeptide chain with a folded structure that does not depend for its stability on any of the remaining parts of the protein.

**Motif (sequence):** A short amino acid sequence with characteristic properties, often those suitable for association with a specific kind of domain on another protein. (Note that the term “domain” is sometimes incorrectly applied to such sequence motifs.)

**Motif (structural):** A domain substructure that occurs in many different proteins, often having some characteristic amino acid sequences properties (e.g., the helix-turn-helix motif in many DNA-recognition domains).

**Topology (or fold):** The structure of most protein domains can be represented schematically by the connectivity in three dimensions of their constituent secondary-structural elements and the packing of those elements against each other. Jane Richardson introduced “ribbon diagrams,” such as those in many of the figures in this chapter, as convenient ways to visualize the fold of a domain (see the caption to Fig. 6-10). Not all folds are found in naturally occurring proteins (e.g., knotted folds are not found), and some folds are more common than others.

**Homologous domains (or proteins):** Domains (or proteins) that derive from a common ancestor. They necessarily have the same fold, and they often (but not always) have recognizably similar amino acid sequences.

**Homology modeling:** Modeling the structure of a domain based on that of a homologous domain.

**Ectodomain:** The part of a single-pass membrane protein that lies on the exterior side of the cell membrane.

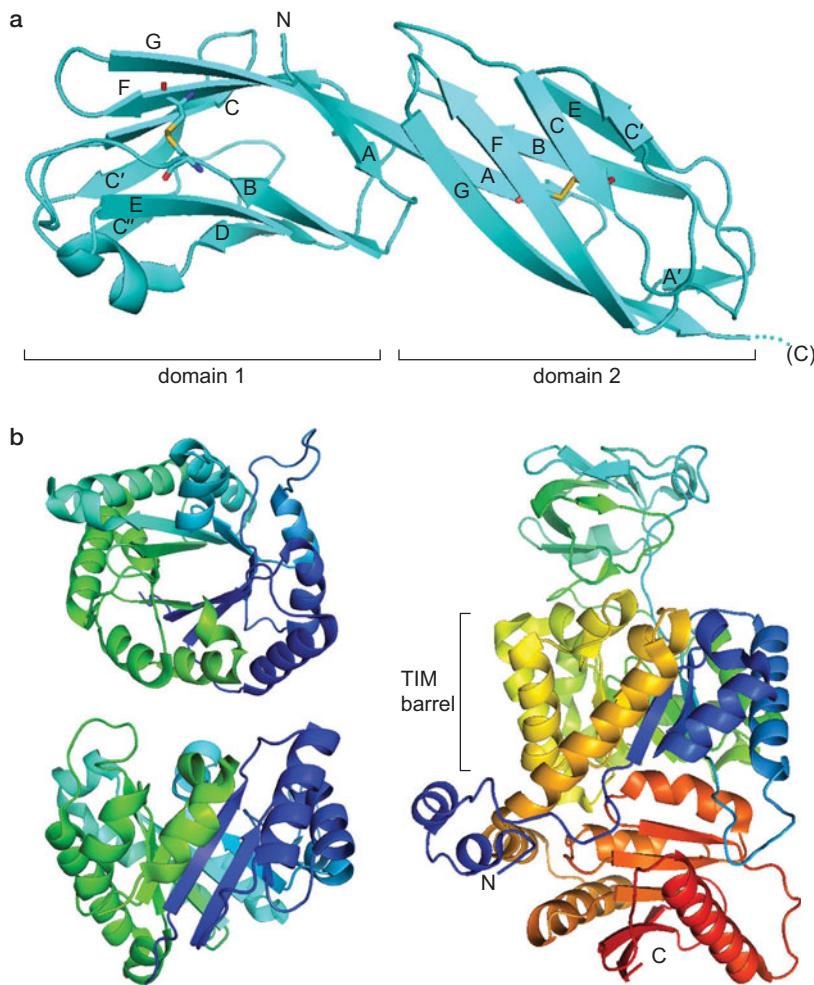
**Glycosylation:** Addition of a chain, sometimes branched, of one or more sugars (glycans) to a protein side chain. The glycans can be N-linked (attached to the side-chain amide of asparagine) or O-linked (attached to the side-chain hydroxyl of serine or threonine).

**Denaturation:** Unfolding a protein or a domain of a protein, either by elevated temperature or by agents such as urea, guanidinium hydrochloride, or strong detergent (“denaturants”).

**Chaperone:** A protein that increases the probability of native folding of another protein, usually by preventing aggregation or by unfolding a misfolded polypeptide chain so that it can “try again” to fold correctly.

**Active site (or catalytic site):** The site on an enzyme that binds the substrate(s), often in a configuration resembling the transition state of the reaction catalyzed.

**Allosteric regulation:** Control of affinity or of the rate of an enzymatic reaction by a ligand that binds at a site distinct from that of the substrate(s). The mechanism of allosteric regulation often involves a change in quaternary structure—that is, a reorientation or repositioning of subunits with respect to each other.



**FIGURE 6-10 Protein domains.** Polypeptide chains are shown here schematically as “ribbons”—a representation, introduced by Jane Richardson, that emphasizes the role of secondary structural elements in the folded conformation of a domain:  $\alpha$  helices are curled ribbons;  $\beta$  strands are gently curved arrows, pointing toward the carboxyl terminus. Intervening loops between secondary structural elements are shown as “worms.” (a) Two of the four domains of the protein CD4, which is found on the surface of certain T-cells and macrophages. Each of these domains is a  $\beta$ -sandwich with an immunoglobulin fold (see Box 6-3); the  $\beta$  strands of each domain are designated by letters in the order in which they follow in the polypeptide chain. Each domain has a single disulfide bond, shown in a stick representation with bonds to sulfur atoms in yellow. (b) Two enzymes: triose phosphate isomerase (TIM; left) and pyruvate kinase (PK; right). The figure shows one monomer of the TIM dimer. The TIM subunit is the prototype of a domain called a “TIM barrel”—a short cylinder in which the eight strands that form the inner barrel alternate with helices that cover the periphery. The two views are along the barrel axis (top) and from the side (bottom). The colors run from dark blue at the amino terminus to green at the carboxyl terminus. PK folds into three domains. The central domain is a TIM barrel (compare with the side view of TIM). The “rainbow” colors run from dark blue at the amino terminus to red at the carboxyl terminus. The light blue domain at the top folds from residues that follow strand 3 of the TIM barrel. The orange-red domain at the bottom contains residues carboxy-terminal to the last TIM-barrel helix. The comparison of TIM and PK shows that a domain found as an isolated unit in one protein can join with additional domains in another protein. Moreover, one or more of those additional domains can fold from a polypeptide chain “inserted” between secondary structural elements of the principal domain. Images prepared with PyMOL (Schrödinger, LLC).

however, at least one of the domains folds from two (or more) noncontiguous segment(s), and the intervening part of the chain forms a distinct domain (Fig. 6-10b). The intervening domain then looks like an insertion into the domain that folds from the flanking segments.

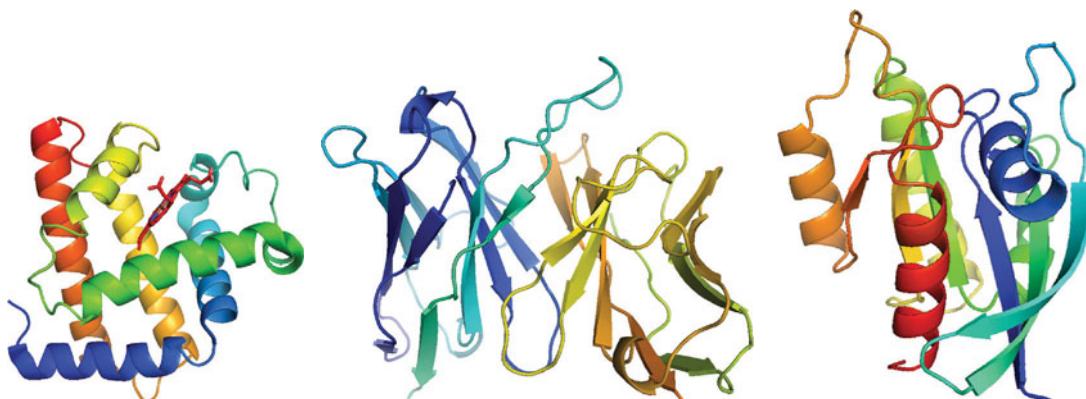
### Basic Lessons from the Study of Protein Structures

The large number of domain structures that have been determined experimentally allows us to draw the following conclusions. First, most hydrophobic side chains are, indeed, buried, and most polar side chains are exposed. Second, if a functional group that can donate or accept a hydrogen bond is buried, it almost always has a hydrogen-bonding partner. The reason for this property is easy to grasp, when we recall that were the polar group exposed on the domain surface, it would make a similar hydrogen bond with water (which can both donate and accept). If the hydrogen bond were missing in the folded conformation, a favorable energetic contribution would have been lost when water was stripped away from that group as the polypeptide chain folded. Even hydrophobic amino acid residues have two hydrogen-bonding groups, an NH and a CO, in their peptide backbone. These hydrogen bonds are also satisfied in folded structures, in considerable part by formation of secondary structures. Both  $\alpha$  helices and  $\beta$  sheets satisfy the main-chain hydrogen bonds of all of the residues within them.

Fulfilling main-chain hydrogen bonding is probably an important reason for the prevalence of regular secondary structures, even within compactly folded protein domains. As a result, it is useful to classify the observed domain structures according to the kinds of secondary structures present within them. We observe that even a relatively short polypeptide chain could, in principle, have an astronomically large number of folded conformations. Only a restricted number of these appear in the large catalog of known 3D protein structures. These not only have a substantial proportion of their amino acid residues in  $\alpha$  helices or  $\beta$  sheets (rather than in irregular loops, which would be much less likely to allow main-chain hydrogen bonding), but also have a relatively simple 3D folding pattern. For example, the Ig domains in CD4 (Fig. 6-10a) are composed of two  $\beta$  sheets—a  **$\beta$  sandwich**—with four or five strands in one sheet and four in the other. Although there would be many ways for the polypeptide chain to pass from one of these eight or nine strands to the next, the observed pattern is one in which the chain makes either a sharp turn within one sheet, linking two adjacent strands, or passes across the top or bottom of the domain to the other sheet. One very important property of all known domain structures is that the chain does not form a knot—that is, if you imagined pulling on its ends, the whole thing would open into a straight line.

### Classes of Protein Domains

Classifications of protein domains allow simple, summary descriptions. One widely used classification hierarchy, embodied in a database called CATH, starts with separation of proteins into classes according to their principal secondary structures (mostly  $\alpha$  helix, mostly  $\beta$  strand, a mixture of the two, and a fourth class for the usually small domains with very little secondary structure). The most important levels in the classification hierarchy are **fold** (also called **topology**) and **homology**. The fold class takes into account not only the secondary structures, but also how the chain passes from one helix or strand to another. The diagrams in Figure 6-11 illustrate this point. A group of homologous proteins are ones with sequence similarities great enough to assume that they have a common evolutionary origin. An unanswered question concerns the likelihood that all domains of a given fold class have a common origin—for very complex domains, a common origin seems intuitively reasonable.



**FIGURE 6-11** Examples of the three principal classes of fold. (Left) An all  $\alpha$ -helical protein (myoglobin). (Center) A heterodimer of two all  $\beta$ -strand domains (the variable region of an immunoglobulin—see Box 6-3). (Right) A mixed  $\alpha$ - and  $\beta$ -domain (the small GTPase, Ras). Colors in each domain run from dark blue at the amino terminus to red at the carboxyl terminus. Images prepared with PyMOL (Schrödinger, LLC).

## Linkers and Hinges

The links between two domains of a folded protein can be very short, allowing a tight and rigid interface between them, or quite long, allowing considerable flexibility. Some proteins have extremely long flexible linkers, because their function within a cell requires that the domains at either end interact over long and variable distances. The amino acid sequences of long linkers generally lack large hydrophobic groups, which their extendable, flexible conformation cannot sequester from water, and have other simplified features.

We summarize our discussion of domains and four levels of protein structure with the illustration of an antibody (immunoglobulin) molecule, described in Box 6-3, The Antibody Molecule as an Illustration of Protein Domains.

## Post-Translational Modifications

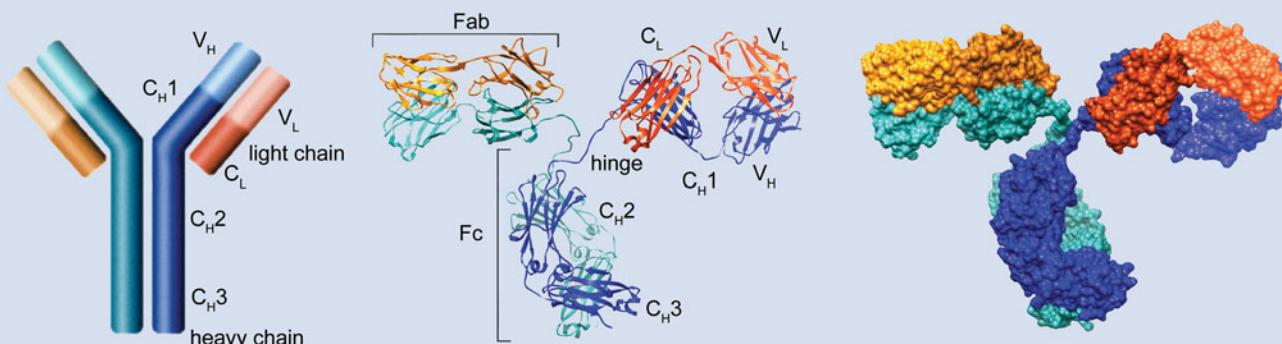
Various modifications of amino acid side chains, introduced following emergence of the polypeptide chain from a ribosome, can modulate the

### ► ADVANCED CONCEPTS

#### Box 6-3 The Antibody Molecule as an Illustration of Protein Domains

Circulating antibodies are immunoglobulin G (IgG) molecules, which contain two identical heavy chains and two identical light chains. The light chains have a variable domain ( $V_L$ ) and a constant domain ( $C_L$ ); the heavy chains, a variable domain ( $V_H$ ) and three constant domains ( $C_{H1}$ ,  $C_{H2}$ , and  $C_{H3}$ ). Thus, there are a total of 12 independent domains.  $V_H$  and  $V_L$  are “variable,” because there is a large combinatorial library of genes that encode them and because somatic mutations occur in the selected gene during the course of an immune response. The variable domains determine specific affinity for antigen. The  $C_{H1–3}$  and  $C_L$  domains are “constant,” because a much smaller number of these domains are linked with one of the many variable domains during maturation of an antibody-producing cell and because they are not prone to somatic mutation. The domains pair in the assembled heterotetramer as shown in

Box 6-3 Figure 1:  $V_H$  with  $V_L$  and  $C_{H1}$  with  $C_L$ , forming an Fab (“antigen-binding”) fragment;  $C_{H2}$  and  $C_{H3}$  of one heavy chain with  $C_{H2}$  and  $C_{H3}$  of the other, respectively, forming an Fc fragment. Controlled proteolytic attack selectively cleaves the hinge, allowing preparation of both the Fab and Fc moieties. Each of the domains has a similar, “Ig-domain” fold, illustrated also in Figure 6-11 as an example of an all- $\beta$  domain. The short link (“elbow”) between variable and constant domains has restricted flexibility. The much longer link (hinge) between  $C_{H2}$  and  $C_{H3}$  of each of the heavy chains has much greater flexibility, allowing the antigen binding sites (called “complementarity determining regions”) at the tip of each Fab to orient and reorient according to the relative positions of their cognate sites on the antigen.



**BOX 6-3 FIGURE 1** Three different representations of IgG. The left panel is a schematic diagram of the “Y-like” pattern of association among the four chains of IgG. In the center, a “ribbon” diagram emphasizes the IgG secondary structure. And, in the right panel, a surface rendering shows that side chains of folded proteins pack efficiently to fill the hydrophobic interior of the protein. Images prepared with PyMOL (Schrödinger, LLC) and UCSF Chimera.

structure and function of a protein. One of the most important is glycosylation—addition of one or more sugars (“glycans”) to an asparagine side chain or to a serine or threonine side chain. This modification generally takes place in the endoplasmic reticulum of eukaryotic cells, and it is therefore a nearly universal characteristic of the ectodomains of cell-surface proteins and of secreted proteins. Proteins bearing glycans are called glycoproteins. Enzymes that transfer glycans to asparagine side chains recognize a short sequence motif, Asn-X-Ser/Thr, where X can be any amino acid residue.

Phosphorylation of serine, threonine, tyrosine, or histidine side chains is another widespread modification, critical for intracellular regulation. Phosphorylation of the first three residues occurs largely in eukaryotic cells; phosphorylation of the last is more common in prokaryotes.

## FROM AMINO-ACID SEQUENCE TO THREE-DIMENSIONAL STRUCTURE

---

### Protein Folding

The amino acid sequence of a domain determines its stable, folded structure. This generalization is an important part of the central dogma of molecular biology, because it means that the nucleotide sequence of a translated gene specifies not only the amino acid sequence of the protein it encodes, but also the 3D structure and function of that protein. A classic experiment concerning refolding of an unfolded protein in the laboratory first established this point (see Box 6-4, Three-Dimensional Structure of a Protein Is Specified by Its Amino Acid Sequence [Anfinsen Experiment]). It also showed that a polypeptide chain can fold correctly without any additional cellular machinery.

The Anfinsen refolding experiment relies on several key points. First, a protein purified from cells or tissues can be unfolded in solution into a random coil. This unfolding is often called **denaturation**, and it is generally accomplished by exposing the protein to high concentrations of certain solutes called **denaturants** (e.g., urea or guanidinium hydrochloride). If the protein is an enzyme, it loses its catalytic activity. If it has a specific binding property (e.g., recognition of a site on DNA), it loses that specificity. That is, almost all of the functional properties of proteins depend on their folded structures. In the case of the protein that Anfinsen and colleagues used in the experiments described in Box 6-4, complete unfolding also required reducing its four disulfide bonds. Second, careful removal of the denaturant allows the protein to fold again. This process is not always very efficient in the laboratory, for many reasons. Cells have enzymes known as **folding chaperones** that can unfold a misfolded protein and allow it to “try again.” Some of these chaperones also sequester the unfolded protein to prevent aggregation with other proteins in the cell, but they do not in any way specify the correct final structure of their substrate protein. Third, measurement of enzymatic activity is a good monitor of correct refolding of ribonuclease, the protein used in Anfinsen’s experiments. That is, recovery of activity is a good way to follow accumulation of the refolded enzyme in its **native** (i.e., functional) conformation.

Another conclusion from experiments such as Anfinsen’s is that the native structure of a protein is the most stable conformation that its polypeptide chain can adopt, given the particular sequence of amino acids in that chain. In physical chemistry, one would say that the native structure has the lowest free energy of any possible conformation.

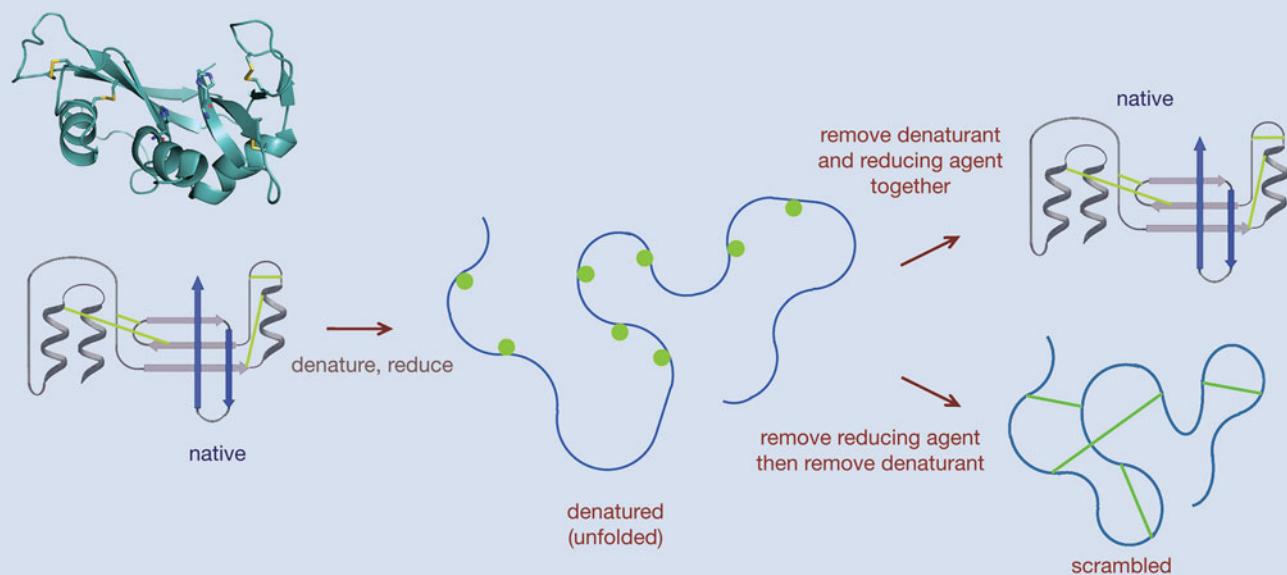
## ► KEY EXPERIMENTS

### Box 6-4 Three-Dimensional Structure of a Protein Is Specified by Its Amino Acid Sequence (Anfinsen Experiment)

In the early 1960s, Christian Anfinsen and coworkers carried out a classic series of experiments, showing that the amino acid sequence of a protein is sufficient to determine its correctly folded structure and that no external folding “machinery” is necessary. This conclusion is fundamental to our understanding of how the nucleotide sequence of a gene ultimately encodes the information needed to specify protein function.

Ribonuclease A is an enzyme that cleaves the phosphodiester backbone of RNA. The enzyme is active when folded into its native conformation but is inactive when unfolded by a denaturant, such as urea or guanidinium hydrochloride at concentrations of 2–5 M. The 124-residue protein has eight cysteines, which form four disulfide bonds (see Box 6-4 Fig. 1). These disulfides can be reduced to sulfhydryls by adding a reducing agent, such as  $\beta$ -mercaptoethanol. Anfinsen and coworkers found that if they unfolded ribonuclease A in the presence of  $\beta$ -mercaptoethanol and then removed both the denaturant and the reducing

agent by dialysis, they could recover a high level of enzymatic activity. Assuming that only a properly folded enzyme can catalyze hydrolysis of phosphodiester bonds, recovery of activity showed that the polypeptide chain contains all the information needed to dictate the folded structure. When Anfinsen et al. first removed the  $\beta$ -mercaptoethanol, allowing disulfide bonds to re-form, and then dialyzed away the denaturant, they failed to detect activity. Eight cysteines can pair in 105 distinct ways. Formation of disulfide bonds in the presence of denaturant might be expected to allow cysteines to pair randomly, leading primarily to forms with scrambled disulfide bonds rather than to the unique pairing found in the native protein. Thus, oxidation of the unfolded ribonuclease A should yield less than 1% of the activity recovered by oxidizing and refolding at the same time. This expectation agrees with the observations, strengthening the fundamental conclusion that only when the native, noncovalent contacts can form will each cysteine find its proper partner.



**BOX 6-4 FIGURE 1** The Anfinsen experiment. Ribonuclease A is represented on the upper left as a ribbon diagram showing the tertiary structure of the enzyme (here the disulfide bonds are shown in yellow). The corresponding schematic below depicts the various secondary structure elements and the locations of the four disulfide bonds. Reducing the disulfides in the presence of a denaturant unfolds the polypeptide chain. Removal of the reducing agent in the presence and in the absence of denaturant leads to two quite different outcomes, as described in the text. In the schematic, the disulfide bonds are represented as green lines and the cysteines as green circles.

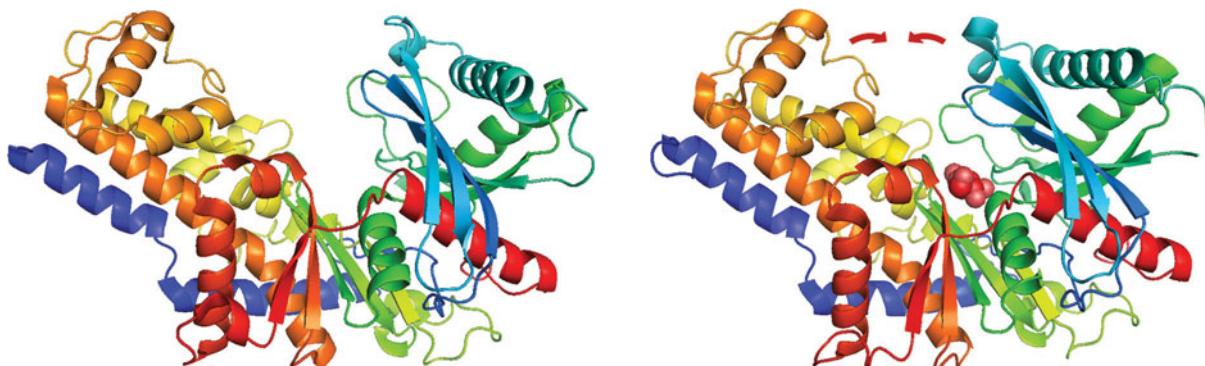
### Predicting Protein Structure from Amino Acid Sequence

In principle, if amino acid sequence determines the folded structure of a protein, it should be possible to devise a computational method for doing the same thing. But in practice, the computational task is daunting. The following comments illustrate why. First, we might imagine that a computer could calculate the stability (free energy) of every possible conformation of the

polypeptide chain and then pick the one that corresponds to a minimum. It is indeed possible to compute the various forces between atoms in a protein that determine its stability—hydrogen bonds, hydrophobic contacts, and so on. But consider a small protein of 100 amino acid residues and imagine that each residue can have only three configurations (e.g., helix, strand, and other). Then the number of possible conformations is roughly  $3^{100}$  or  $10^{47}$ , an astronomical figure, ruling out this strategy. Second, we might try to simulate the process of protein folding, by some sort of dynamic calculation. Efforts to do so are starting to work, for small proteins and with advanced computational resources; the answers are good approximations for some purposes, but not yet adequate for understanding all aspects of function. Such an approach is not likely in the near future to be a practical way to predict structures of complex proteins just from their amino acid sequences. Third, if we already know the structure of a similar, homologous protein, we might consider starting with it as a first approximation and computationally changing the amino acid residues to match the new protein we wish to understand. Computations of this kind, known as **homology modeling**, have become relatively practical. Their reliability obviously depends on the similarity of the two proteins in question and on the desired accuracy of the prediction.

## CONFORMATIONAL CHANGES IN PROTEINS

The folded (or unfolded) conformation of a protein under particular conditions is the one with the lowest free energy. If the environment of the protein changes, however, the most stable conformation can also change. We have seen one example—the unfolding and refolding of ribonuclease in response to adding and removing a very high concentration of urea. Much less drastic changes in the environment of a protein can also induce functionally important, conformational shifts. For example, when presented with its substrate, glucose, the single-domain enzyme hexokinase closes up around it (Fig. 6-12). Formation of energetically favorable contacts with the substrate makes the closed structure more stable than the open one, shifting the position of a dynamic equilibrium from mostly open to mostly closed.



**FIGURE 6-12** Domain closure in the enzyme hexokinase. The two lobes of hexokinase, an enzyme that transfers a phosphate to glucose, close up on each other (red arrows) when the substrate (glucose) binds. (Left) Enzyme before binding glucose. (Right) After binding glucose (shown in surface representation, red, in the catalytic cleft of the enzyme). The polypeptide chain is in rainbow colors from blue (amino terminus) to red (carboxyl terminus). Note that the folded chain traverses back and forth twice between the two lobes. Images prepared with PyMOL (Schrödinger, LLC).

Interaction of two proteins with each other can cause one or both partners to undergo a conformational change. Sometimes, the interacting part of one of the partners is unstructured (disordered and flexible) until it associates with the other partner. In other words, the properly folded conformation is stable only in the presence of its target, which can be DNA or RNA as well as another protein. The  $\alpha$  helices in the dimeric coiled-coil of the yeast transcription factor GCN4 are stable only when associated with each other. When bound to DNA, an additional segment of the protein forms an  $\alpha$  helix in the DNA major groove, but the same part of the protein is unstructured when GCN4 is not associated with its DNA-binding site (Fig. 6-9b).

## PROTEINS AS AGENTS OF SPECIFIC MOLECULAR RECOGNITION

---

### Proteins That Recognize DNA Sequence

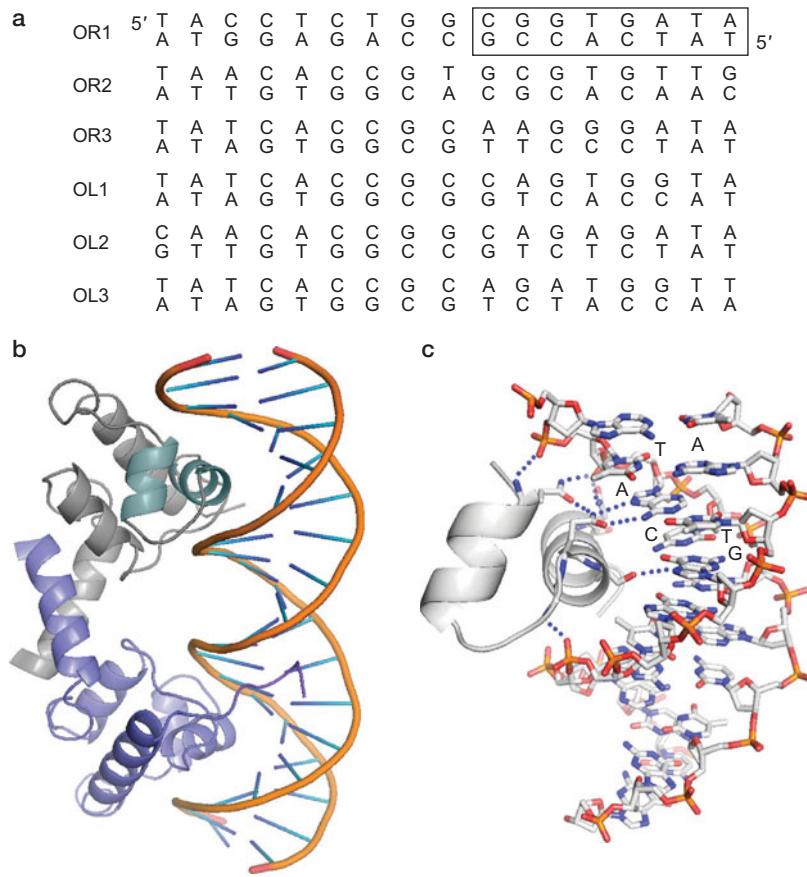
Regulation of gene expression depends on proteins that bind short DNA segments having a specific nucleotide sequence. We consider here several examples that illustrate some of the principles of protein structure and interaction described above.

*i. GCN4* We have already described GCN4 to illustrate how the folding of a protein sometimes depends on its interactions with other proteins (the other chain of the dimer, in the case of the GCN4 coiled-coil segment) or with a target (a DNA site). GCN4 binds tightly to DNA only when the sequence of bases at the contact site is the correct one. Because the  $\alpha$  helices in the major groove fit snugly, their side chains need to be complementary—in their shapes, their polarity, and their hydrogen-bond donor and acceptor properties—to the DNA surface. These  $\alpha$  helices also have several arginine and lysine residues, which anchor them to the DNA backbone by forming salt bridges with phosphates, reinforcing the snug fit.

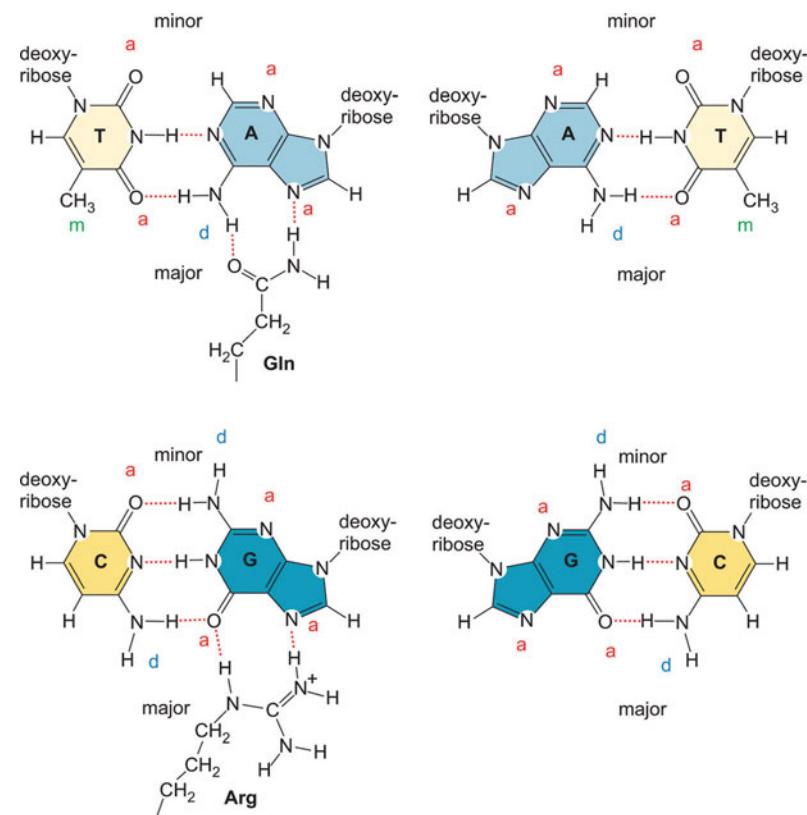
*ii. The Bacteriophage  $\lambda$  Repressor* The repressor of bacteriophage  $\lambda$  has six binding sites on the bacteriophage genome, which all have related but slightly different sequences; the exact sequence of each of the sites determines its affinity for the repressor (Fig. 6-13a). The protein is a symmetric dimer, and the sites have approximately symmetric (**palindromic**) sequences. Each subunit of the protein has two folded domains: an amino-terminal, DNA-binding domain and a carboxy-terminal, dimerization domain.

The DNA-binding domain of  $\lambda$  repressor is a compact bundle of five  $\alpha$  helices (Fig. 6-13b). Unlike GCN4, this domain does not undergo any major structural changes when it associates with DNA. Two of its helices (the second and third) form a structural motif, known as a **helix-turn-helix**, seen in many other DNA-binding proteins, especially those from prokaryotes. The way this motif fits against the DNA double helix allows the second of the two helices, sometimes called the **recognition helix**, to fit into the major groove of DNA and to present several of its side chains to the exposed edges of the base pairs (Fig. 6-13c). The major-group edge of each base pair presents a characteristic pattern of hydrogen-bond donor and acceptor groups; the A:T and T:A base pairs also present the hydrophobic surface of a thymine methyl group (Fig. 6-14). The hydrogen-bonding and nonpolar contact properties of side chains on the  $\lambda$ -repressor recognition helix match those of the base sequence recognized. Contacts between the protein and the

**FIGURE 6-13** DNA recognition by the repressor of bacteriophage  $\lambda$ . (a) The nucleotide sequences of the six DNA sites (“operators”) in the  $\lambda$  genome that bind the  $\lambda$  repressor. Each site is approximately a “palindrome”—the sequence of bases is the same (with some deviations) when read 5' to 3' from either end. The right-hand “half site” of the top sequence (OR1) and the left-hand half site of the bottom sequence (OL3) correspond to the best consensus of all the half sites. Because the overall length is an odd number (17 base pairs), the central base pair is necessarily an exception to a perfect palindrome. (b) The DNA-binding (amino-terminal) domain of  $\lambda$  repressor, bound to operator DNA. Each subunit is a cluster of five  $\alpha$  helices. Two of these (in light blue on the upper subunit) form a helix-turn-helix motif; the first of the two bridges from one side of the major groove to the other, and the second lies in the groove and nearly parallel to its principal direction. (c) Polar interactions (hydrogen bonds and salt bridges) between residues in the helix-turn-helix motif and DNA (both backbone and bases). The protein fits snugly in the major groove only when the base-pair contacts match the groups on the protein that lie opposite them. Images prepared with PyMOL (Schrödinger, LLC).



**FIGURE 6-14** Properties of DNA base pairs in the major and minor grooves. The four DNA base pairs, with labels on groups in the major and minor groove that can determine specific contacts: a, hydrogen-bond acceptor; d, hydrogen-bond donor; m, methyl group (van der Waals contact). Hydrogen bonds are shown as dotted lines. In the major groove, each of the four base pairs presents a distinct pattern: T:A, m-a-d-a; A:T, a-d-a-m; C:G, d-a-a; G:C, a-a-d. Two particular examples of amino acid side-chain complementarity are shown with the T:A and C:G pairs. These two modes of base-pair recognition (pointed out in 1976 by Seeman, Rosenberg, and Rich) do occur with some frequency, but most cases of specific DNA recognition involve a more complex set of contacts. In the minor groove, T:A and A:T present the same pattern of potential contact (a-a); likewise, C:G and G:C (a-d-a). Thus, sequence-specific DNA recognition usually involves major-groove contacts.

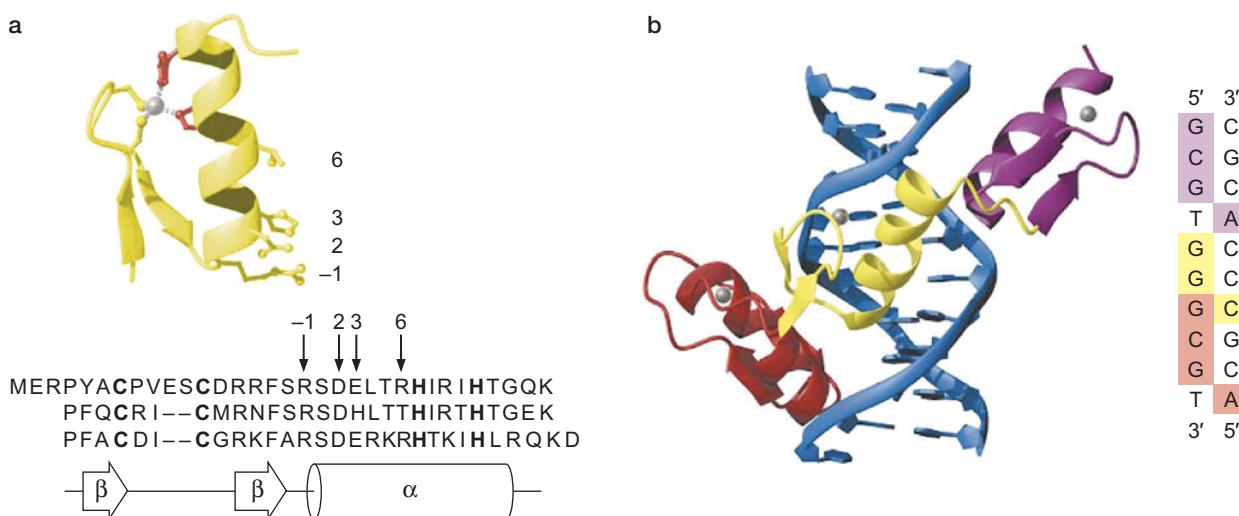


sugar–phosphate backbone of DNA position and orient the recognition helix side chains to ensure this complementarity.

The complementarity of protein side chains and DNA bases differs in an important way from the complementarity of the two bases in a DNA base pair. Each DNA base has a unique complementary base, such that their hydrogen bonding is consistent with the geometry of an undistorted double helix. In contrast, there are several ways in which proteins recognize a particular base or even a particular sequence of bases. Moreover, as illustrated by the different sequences of the repressor-binding sites, the same protein structure can adjust slightly to create complementarity with a slightly altered base sequence (at some cost in affinity). Thus, there is no “code” for DNA recognition by proteins—just a set of recurring themes, such as the presentation of protein side chains by an  $\alpha$  helix inserted into the DNA major groove.

The  $\lambda$  repressor illustrates a general feature of proteins that recognize specific DNA sequences: they have relatively small DNA-binding domains, usually linked to one or more additional domains with distinct functions, such as oligomerization or interaction with other proteins.

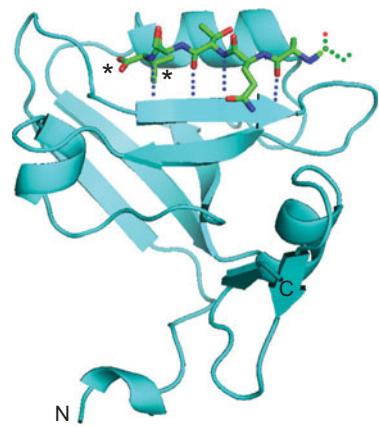
**iii. Zinc-Finger Proteins** The most abundant DNA-recognition domain in many eukaryotes is a small module known as a **zinc finger** (Fig. 6-15a). These domains generally occur in tandem, with short linker segments between them. The linkers are flexible; when the proteins bind DNA, they become ordered. The approximately 30 residues of each zinc finger are barely enough to create a hydrophobic core, and the zinc ion in the center is necessary to hold together the folded domain. Two cysteines and two



**FIGURE 6-15 Zinc-finger motifs.** (a) The Cys2His2 zinc finger motif and the Zif268 finger sequences. Shown at the top is a ribbon diagram of finger 2, including the two cysteine side chains (yellow) and two histidine side chains (red) that coordinate the zinc ion (silver sphere). The side chains of key residues make base contacts in the major groove of the DNA (numbers identify their position relative to the start of the recognition helix). Shown below is the amino acid sequence alignment of the three fingers from Zif268 with the conserved cysteines and histidines in boldface. Secondary structure elements are indicated at the bottom of the diagram. (b) To the left is the Zif268–DNA complex, showing the three zinc fingers of Zif268 bound in the major groove of the DNA. Fingers are spaced at 3-bp intervals; DNA (blue); fingers 1 (red), 2 (yellow), and 3 (purple); the coordinated zinc ions (silver spheres). The DNA sequence of the Zif268 binding site on the right is color-coded to indicate base contacts for each finger. (Reproduced, with permission, from Pabo C.O. et al. 2001. *Annu. Rev. Biochem.* **70**: 313–340, Fig. 6-15a is Fig. 1 on p. 315; Fig. 6-15b is Fig. 2 on p. 316. © Annual Reviews.)



**FIGURE 6-16** The LEF-1 protein bound to DNA. Image prepared with PyMOL (Schrödinger, LLC).



**FIGURE 6-17** Peptide recognition. Specific recognition of the carboxy-terminal segment of a protein by a PDZ domain—a repeating module that associates with the carboxy-terminal, cytoplasmic “tails” of membrane proteins. Principal contacts are in pockets (asterisks for the carboxyl group and the nonpolar side chain of the carboxy-terminal valine) and through addition to the antiparallel  $\beta$  sheet in the domain (foreground) by several residues of the ligand that precede the valine (dotted black lines represent  $\beta$ -sheet hydrogen bonds). Image prepared with PyMOL (Schrödinger, LLC).

histidines coordinate the  $Zn^{2+}$ . Because intracellular proteins do not have disulfide bonds,  $Zn^{2+}$  coordination often serves the same stabilizing purpose for very small domains. When zinc fingers bind DNA, the short  $\alpha$  helix lies in the major groove, and successive zinc fingers in a tandem array contact successive sets of base pairs—roughly 3 bp per zinc finger, with some overlap (Fig. 6-15b). There is considerable regularity in the pattern of base-pair contacts: residues  $-1, 2, 3$ , and  $6$  of the helix are the most likely to contact one or more base pairs (Fig. 6-15a). Because of this regularity—and the way in which tandem zinc fingers wind into the DNA major groove—proteins can be designed to recognize relatively long sequences of base pairs. Moreover, libraries of individual modules are now available to make designed proteins specific for DNA sequences 12–18 bp in length.

**iv. Lymphocyte Enhancer Factor-1 (LEF-1)** Contacts with base pairs in the major groove of DNA are not the only way to create base sequence specificity. The base sequence does not uniquely specify the pattern of hydrogen-bond contacts in the minor groove, because A:T and T:A look the same in this respect, as do G:C and C:G (Fig. 6-14), but base sequence also influences the propensity for the DNA double helix to bend or twist—that is, to adopt conformations that deviate from an ideal Watson–Crick double helix. This sensitivity to the influence of base sequence on the propensity of DNA to bend and twist is sometimes called “indirect readout,” to distinguish it from the sequence specificity provided by direct polar and nonpolar contacts with base pairs. Lymphocyte Enhancer Factor-1 (LEF-1), which regulates T-cell gene expression in concert with several other factors, is a three-helix bundle that fits into the substantially widened minor groove of bent DNA (Fig. 6-16). Most of the amino acid side chains that face into the minor groove are nonpolar, and one of them inserts part way between two adjacent base pairs, helping to stabilize the nearly  $90^\circ$  bend in the DNA axis. The bend brings proteins bound upstream and downstream of LEF-1 closer together: It has been called an “architectural protein” for this reason, because part of its role is to enhance contacts between other DNA-bound transcription factors.

### Protein–Protein Interfaces

Protein–protein interfaces tend to be even more exquisitely complementary than protein–DNA interfaces. The reason is that the former generally involve considerable hydrophobic surface, whereas the latter are largely polar. Water, which is both a donor and acceptor, can bridge gaps between hydrogen-bonding groups at a DNA–protein interface, but a gap between nonpolar surfaces at a protein interface would leave either a hole or an isolated water—both very unfavorable. As we have seen, a transcription factor such as  $\lambda$  repressor can bind DNA targets with a modest range of sequences, each deviating slightly from a consensus. The same is not true for most protein interfaces. For some transcription factors, alternative pairing of structurally homologous subunits *does* occur, to increase combinatorial diversity. The relevant complementary surfaces are conserved in such cases, which probably arise from gene duplication at some point in the evolutionary history of the protein.

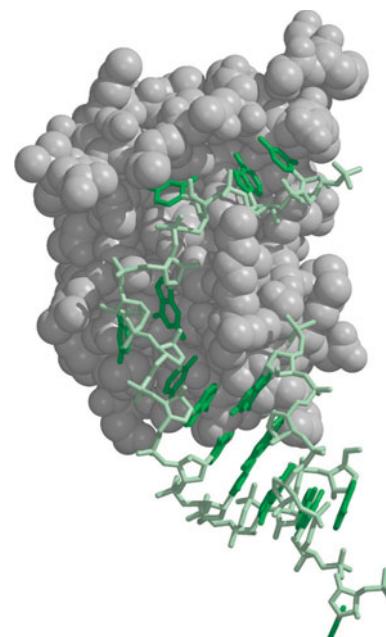
Specific protein recognition can depend on association of prefolded, matching surfaces of two subunits, such as occurs in formation of a hemoglobin tetramer (Fig. 6-7), or on cofolding of two polypeptide chains, as in GCN4 dimerization (Fig. 6-9a), or on docking of an unstructured segment onto the recognition surface of a partner protein (Fig. 6-17). In this last sort of interaction, the segment in question adopts a defined structure in

the complex—that is, its correctly folded conformation is stable only in the presence of the target surface. Binding sometimes depends on a post-translational modification such as phosphorylation or acetylation, so that the interaction can be switched on or off by signals from other cellular processes. The docked segment of polypeptide chain often has a recognizable amino acid sequence motif. Association of this kind is particularly common in the assembly of protein complexes that regulate transcription, probably because it allows considerable variability in longer-range organization. Either the unstructured segment or the domain that binds it, or both, may be embedded in a larger unstructured region with a relatively polar, “low-complexity” amino acid composition (i.e., having many repeated instances of the same, polar residue). These low-complexity regions impart long-range flexibility, so the spacing between the specific interactions can vary, and the same assembly can adapt to different circumstances (e.g., to different arrays of sites on DNA).

### Proteins That Recognize RNA

Unlike DNA, RNA can have a great variety of local structures, and tertiary interactions stabilize well-defined 3D conformations, as in tRNA. Protein–RNA interactions are therefore in some respects rather like protein–protein interactions. The shape of the RNA and the way interacting groups (e.g., phosphates or 2'-hydroxyl groups or bases) distribute on its surface are critical determinants of specificity. Two prefolded structures can associate, as in binding of a tRNA to the enzyme that transfers an amino acid to its 5' end, or one or both partners can have little or no fixed structure until the complex forms.

The RNA-recognition motif (RRM; also known as the ribonuclear protein [RNP] motif) is a sequence that characterizes a domain involved in specific RNA recognition. The RRM sequence of 80–90 amino acids folds into a four-stranded antiparallel  $\beta$  sheet and two  $\alpha$  helices that pack against it. This arrangement gives the domain a characteristic split  $\alpha\beta$  topology. An example of this common domain is found in the U1A protein that interacts with the U1 small nuclear RNA (snRNA), both components of the machinery that splices RNA transcripts (Chapter 14). The structure of the U1A:U1snRNA complex, shown in Figure 6-18, shows that the shape of the RNA-binding surface of U1A is specific for this particular RNA.



**FIGURE 6-18** Structure of spliceosomal protein:RNA complex: U1A binds hairpin II of U1 snRNA. The protein is shown in gray; the U1 snRNA is shown in green. (Oubridge C. et al. 1994. *Nature* 372: 432.) Image prepared with MolScript, BobScript, and Raster 3D.

### ENZYMES: PROTEINS AS CATALYSTS

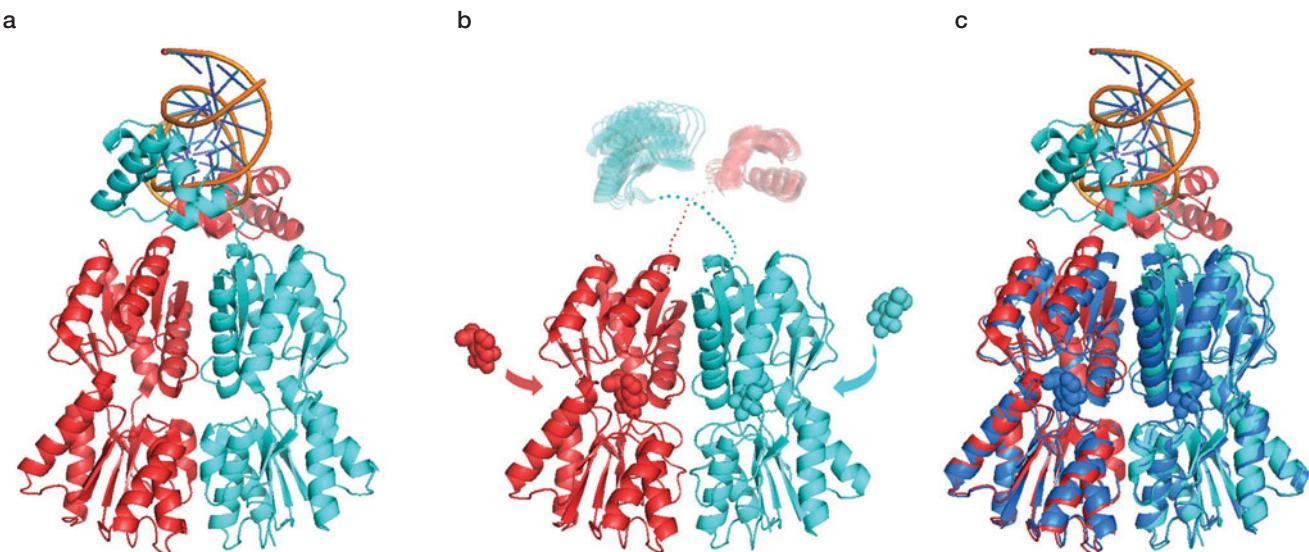
One of the most important roles for proteins in cells is to catalyze biochemical reactions. Almost all processes that go on in a cell—from transformation of nutrients for generating energy to polymerization of nucleotides for synthesis of DNA and RNA—require catalysis (i.e., enhancement of their rates), because the spontaneous reaction rates are far too slow to support normal cellular activity and survival. Most catalysts in living systems are proteins (enzymes); RNA is a catalyst for certain very ancient reactions (ribozymes).

The barrier to a chemical reaction is formation of a high-energy arrangement of the reactants, known as the **transition state**. Because the transition state has a structure intermediate between those of the reactants and the products, some distortion of the reactants is necessary to reach it. A reaction can be accelerated—often very dramatically—by reducing the energy needed to distort the reactants into their transition-state configurations. Most enzymes do so by having an **active site**—usually a pocket or groove—that

is complementary in shape and interaction properties (e.g., hydrogen bonds and nonpolar contacts) to the transition state of the reaction. The favorable contacts that form when the reactants associate with the active site compensate to some extent for the distortion they undergo to do so. The precision with which evolution of an enzyme structure molds its active site imparts great specificity to this process. For example, enzymes that catalyze polymerization of deoxyribonucleotides into DNA cannot, in general, catalyze polymerization of ribonucleotides into RNA, because the 2'-hydroxyl of the ribose would collide with atoms in the active site of the polymerase.

## REGULATION OF PROTEIN ACTIVITY

We have seen that interaction with other molecules—both small molecules such as the substrates of an enzyme or macromolecules such as proteins and nucleic acids—can induce proteins to undergo conformational changes. Molecules that bind a protein (or any other target) in a defined way are known as **ligands**. Ligands can regulate the activity of a protein (e.g., an enzyme) by stabilizing a particular state. For example, if binding of a ligand to an enzyme stabilizes a conformation in which the active site is blocked, the ligand will have turned off the activity of that enzyme. The binding site for the inhibitory ligand need not overlap the active site—it need only



**FIGURE 6-19** Allosteric regulation of Lac repressor DNA binding. (a) DNA-bound conformation of dimeric Lac repressor. A short DNA segment, representing the specific binding site (“operator”), is at the top of the figure. The amino-terminal, DNA-binding domain, with a helix-turn-helix recognition motif, interacts with base pairs in the major groove. The body of the protein has a site, located between its two domains, that accommodates molecules related to lactose; the site is empty in the DNA-bound conformation shown here. The two identical repressor subunits are in red and cyan, respectively. (b) Binding of an inducer molecule (any of a variety of galactosides, illustrated in surface rendering, both outside the repressor, as if about to bind, and also at the specific binding site within each repressor subunit) causes the two domains in the body of the repressor to change orientation with respect to each other. As a result, the hinge segments between the DNA-binding domains and the body of the protein become disordered, with the domains themselves now loosely tethered and unable to bind tightly to operator sites. (c) Superposition of the DNA-bound and induced conformations, to show how one of the domains of the repressor shifts with respect to the other. DNA-bound subunits are colored as in panel a; the induced repressor dimer is in dark blue. Images prepared with PyMOL (Schrödinger, LLC).

be such that ligand binding lowers the energy of a conformation in which the reactants cannot reach the active site or in which the active site no longer has the right configuration. Conversely, ligand binding at a remote site might favor a conformation in which the active site is available to substrate and complementary to the transition state of the reaction; the ligand would then be an activator. This kind of regulation is known as **allosteric regulation** or allostery, because the structure of the ligand (its “steric” character) is different from (Greek *allo-*) the structure of any of the reactants.

The Lac repressor (which inhibits expression of the bacterial gene encoding  $\beta$ -galactosidase, an enzyme that hydrolyzes  $\beta$ -galactosides such as lactose) is a good example of allosteric regulation in control of transcription (Fig. 6-19). Lac repressor is a dimer. The dimer has two distinct conformations—one when bound to a specific DNA site (known as its **operator**) and another when bound to an inhibitory metabolite (known as its **inducer**). Because the operator-bound repressor blocks RNA polymerase from synthesizing  $\beta$ -galactosidase mRNA and because a high concentration of the inducer favors a conformation that does not bind well to DNA, the inducer can change DNA affinity and hence influence gene regulation, even though its binding site is at some distance from the DNA-contacting surface of the repressor. Even more complicated allosteric switches are possible, with multiple ligands and multiple binding sites. Allosteric regulation often involves quaternary-structure changes, as in the transition between the two dimer conformations of Lac repressor.

## SUMMARY

---

Proteins are linear chains of amino acids, joined by peptide bonds (“polypeptide chains”). The 20 L-amino acids specified by the genetic code include nine with nonpolar (hydrophobic) side chains, six with polar side chains that do not bear a charge at neutral pH, two with acidic side chains (negatively charged at neutral pH), and three with basic side chains (positively charged at neutral pH, or partially so in the case of histidine). Peptide bonds have partial double-bond character; torsion angles for the N-C $\alpha$  and C $\alpha$ -(C=O) bonds specify the three-dimensional conformation of a polypeptide-chain backbone. Three amino acids have special conformational properties: glycine is nonchiral, with greater conformational freedom than the others; proline (technically, an imino acid) has a covalent bond between side chain and amide, restraining its conformational freedom; and cysteine, with a sulphydryl group on its side chain, can undergo oxidation to form a disulfide bond with a second cysteine, cross-linking a folded polypeptide chain or two neighboring polypeptide chains. The reducing environment of a cell interior restricts disulfide-bond formation to oxidizing organelles and the extracellular milieu.

Protein structure is traditionally described at four levels: primary (the sequence of amino acids in the polypeptide chain—the one level determined directly by the genetic code), secondary (local, repeated backbone conformations, stabilized by main-chain hydrogen bonds—principally  $\alpha$  helices and  $\beta$  strands), tertiary (the folded, three-dimensional conformation of a polypeptide chain), and quaternary (association of folded polypeptide chains in a multisubunit assembly). At the tertiary level, polypeptide chains fold

into one or more independent domains, which would fold similarly even if excised from the rest of the protein. The structure of a domain can usefully be specified by the way in which its component secondary-structure elements (helices and strands) pack together in three dimensions. Linkers between domains of a multidomain polypeptide chain can be long and flexible or short and stiff. The aqueous environment and the diverse set of naturally occurring amino acids are together critical for the conformational stability of folded domains and of the interfaces between them that create quaternary structure. Nonpolar side chains cluster away from water into the closely packed, hydrophobic core of a folded domain, and any sequestered hydrogen-bonding groups, which lose a hydrogen bond with water, must have a protein-derived partner. Secondary-structure elements satisfy the latter requirement for the main-chain amide and carbonyl groups, thus accounting for their importance in describing and classifying domain structures.

The sequence of amino acids in a polypeptide chain specifies whether and how it will fold. This property allows the genetic code to determine not merely primary structure, but other levels as well, and hence to dictate protein function. The various noncovalent interactions within a correctly folded domain (and in extracellular domains, the covalent disulfide bonds) create a global free-energy minimum (conformation of greatest stability), so that the chain can reach its native conformation spontaneously. Changes in the environment of a protein, including post-translational modifications of one or more of its side chains or binding of ligands, may alter the position of this free-energy minimum and

induce a conformational change. The array of amino acid side chains on the surface of a folded protein, and sometimes even in a segment of unfolded polypeptide chain, can also specify how it recognizes a protein or nucleic-acid partner or a

small-molecule ligand. Proteins are thus the key agents of specific molecular recognition, both within a cell and between cells, as well as the specific catalysts of chemical reactions (enzymes).

## BIBLIOGRAPHY

---

### Books

- Branden C. and Tooze J. 1999. *Introduction to protein structure*, 2nd ed. Garland Publishing, New York.
- Kuriyan J., Konforti B., and Wemmer D. 2012. *The molecules of life*. Garland Publishing, New York.
- Pauling L. 1960. *The nature of the chemical bond*, 3rd ed. Cornell University Press, Ithaca, New York.
- Petsko G.A. and Ringe D. 2003. *Protein structure and function (primers in biology)*. New Science Press, Waltham, Massachusetts.
- Williamson M. 2012. *How proteins work*. Garland Publishing, New York.

### Protein Structure Can Be Described at Four Levels

Richardson J.S. 1981. The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* **34**: 167–339.

### From Amino Acid Sequence to Three-Dimensional Structure

Anfinsen C.B. 1973. Principles that govern the folding of protein chains. *Science* **181**: 223–230.

## QUESTIONS

---

**MasteringBiology**®

For instructor-assigned tutorials and problems, go to *MasteringBiology*.

For answers to even-numbered questions, see Appendix 2: Answers.

**Question 1.** What is the bond that can form between two cysteines in secreted proteins? Why does this bond not ordinarily form in intracellular proteins? How does this interaction differ from the interactions that can occur between other amino acid side chains?

**Question 2.** Give an example of two amino acid side chains that can interact with each other through an ionic bond at neutral pH. See Chapter 3 for a review of ionic bonds.

**Question 3.** A mutation that occurs in DNA can cause an amino acid substitution in the encoded protein. Amino acid substitutions are described as conservative when the amino acid in the mutated protein has chemical properties similar to those of the amino acid it has replaced. Referring to Figure 6-2, identify four different examples of pairs of amino acids that could be involved in conservative substitutions.

**Question 4.** Peptide bond formation is an example of a condensation reaction. Explain what this statement means and why peptide bond formation is also referred to as a dehydration reaction.

**Question 5.** Describe how a  $\beta$  strand differs from a  $\beta$  sandwich.

**Question 6.** The oxygen-binding proteins hemoglobin and myoglobin differ in that hemoglobin functions as a tetramer in red blood cells, whereas myoglobin functions as a monomer in muscle cells. The globular structure of myoglobin and each hemoglobin monomer involves eight  $\alpha$ -helical segments. Is it the primary, secondary, tertiary, or quaternary structure that differs most between these two proteins? Explain.

**Question 7.** For the following amino acids, suggest whether they are more likely to be found buried or exposed in a stably folded

protein domain: phenylalanine, arginine, glutamine, methionine. Explain your answers.

**Question 8.** You treat a protein with the denaturant urea. For each interaction or bond below, state if the interaction or bond is disrupted by the urea treatment.

- A. Ionic bonds.
- B. Hydrogen bonds.
- C. Disulfide bonds.
- D. Peptide bonds.
- E. van der Waals interactions.

**Question 9.** From what you learned about the structure of DNA in Chapter 4, explain why Gcn4 interacts with DNA in the major groove rather than in the minor groove. Describe the importance of arginines and lysines in the interaction between Gcn4 and DNA.

**Question 10.** Predict the effect of substituting one or more of the conserved cysteines or histidines in a Cys<sub>2</sub>His<sub>2</sub> zinc finger with alanine. Explain your answer.

**Question 11.** Describe the unusual features of the interaction of LEF-1 with DNA.

**Question 12.** How do enzymes enhance the rate of a reaction?

**Question 13.** Consider a ligand that is structurally similar to the substrate of an enzyme and that binds tightly in the active site, excluding the normal substrate. What is the difference between such a “competitive inhibitor” and an allosteric inhibitor?

**Question 14.** A translation initiation factor, called Tif3 or eIF4B, in yeast cells has the following sequence of elements in its polypeptide chain: an amino-terminal domain containing

an RNA recognition motif (RRM), a central segment with a seven-fold repeated sequence rich in basic and acidic amino acid residues, and a carboxy-terminal region with no evident homology to known motifs or domains.

**A. Describe the significance of RRMs.**

The modularity of the protein suggested the following series of experiments, to analyze the roles of its different parts. Cells with a deletion of the *TIF3* gene do not grow at 37°C, but grow normally at 30°C. By adding a gene that encodes a fragment of the protein, it is possible to assay for complementation—the capacity of that fragment to confer wild-type growth at 37°C. The results of such experiments are shown in the table below, in which ++, +, and – indicate the degree of growth/complementation.

Tif3 protein	Growth at 37°C
Full-length	++
RRM + first three repeats	++
RRM + first repeat	–
All seven repeats + carboxy-terminal segment	+

**B. From these data, which region of the protein is required for wild-type growth at 37°C?**

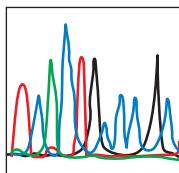
A further set of experiments involved an in vitro translation assay, using an extract from the strain lacking the *TIF3* gene. The reaction was initiated by adding purified Tif3 protein or one of its truncated forms. The results are in the table below, which shows the percentage of in vitro translation relative to the reaction with full-length Tif3.

Tif3 protein	Translation Activity (%)
Full-length	100
RRM + first three repeats	43
RRM + first repeat	0
All seven repeats + carboxy-terminal segment	0

**C. How do these results compare with the complementation results in part B? Do they modify your conclusion from the genetic experiments?**

Data adapted from Niederberger et al. (1998. *RNA* **4**: 1259–1267).

*This page intentionally left blank*



# Techniques of Molecular Biology

THE LIVING CELL IS AN EXTRAORDINARILY complicated entity, producing thousands of different macromolecules and harboring a genome that ranges in size from a million to billions of base pairs. Understanding how the genetic processes of the cell work requires a variety of challenging experimental approaches. These include methods for separating individual macromolecules from the myriad mixtures found in the cell and for dissecting the genome into manageable sized segments for manipulation and analysis of specific DNA sequences, as well as the use of suitable model organisms in which the tools of genetic analysis are available, as will be discussed in Appendix 1. The successful development of such methods has been one of the major driving forces in the field of molecular biology during the last several decades, as well as one of its greatest triumphs.

Recently, it has become possible to apply molecular approaches to the large-scale analysis of the full complement of DNA, RNA, and proteins in the cell. These genomic and proteomic approaches and the rapidly increasing number of genome sequences becoming available make it possible to undertake large-scale comparisons of the genomes of different organisms or to identify all of the phosphorylated proteins in a particular cell type.

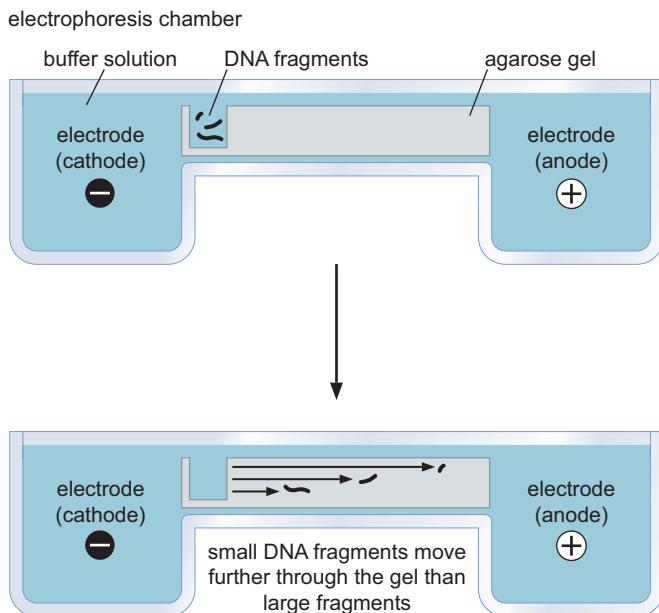
In this chapter, we provide a brief introduction to some of the modern methods that permit biologists to investigate the function of individual proteins as well as to perform large-scale analysis of genomes and proteomes. As we shall see, these methods often depend on, and were developed from, an understanding of the properties of biological macromolecules themselves. For example, the base-pairing characteristics of DNA and RNA gave rise to the development of techniques of hybridization that now permit whole-genome identification of gene expression. Insight into the activities of DNA polymerases, restriction endonucleases, and DNA ligases gave birth to the techniques of DNA cloning and the polymerase chain reaction (PCR), which allow scientists to isolate essentially any DNA segment—even some from extinct life-forms—in unlimited quantities.

This chapter is divided into five parts: methods for the analysis of DNA and RNA; the large-scale analysis of genomic DNA; analysis of proteins; large-scale analysis of proteins; and, finally, we describe the analysis of nucleic acid–protein interactions, approaches that help us to explore how these separate components come together and interact to facilitate the inner workings of the cell.

## O U T L I N E

- Nucleic Acids: Basic Methods, 148
  - Genomics, 168
  - Proteins, 173
  - Proteomics, 179
- Nucleic Acid–Protein Interactions, 182
  - Visit Web Content for Structural Tutorials and Interactive Animations

**FIGURE 7-1** DNA separation by gel electrophoresis. The figure shows a gel from the side in cross section. The “well” into which the DNA mixture is loaded is indicated at the left, at the top of the gel. This is also the end at which the cathode of the electric field is located, the anode being at the bottom of the gel. As a result, the DNA fragments, which are negatively charged, move through the gel from the top to the bottom. The distance each DNA travels is inversely related to the size of the DNA fragment, as shown. (Adapted, with permission, from Micklos D.A. and Freyer G.A. 2003. *DNA science: A first course*, 2nd ed., p. 114. © Cold Spring Harbor Laboratory Press.)



## NUCLEIC ACIDS: BASIC METHODS

### Gel Electrophoresis Separates DNA and RNA Molecules according to Size

We begin by discussing the separation of DNA and RNA molecules by the technique of **gel electrophoresis**. Linear DNA molecules separate according to size when subjected to an electric field through a **gel matrix**—an inert, jelly-like porous material. Because DNA is negatively charged, when subjected to an electrical field in this way, it migrates through the gel toward the positive pole (Fig. 7-1). DNA molecules are flexible and occupy an effective volume. The gel matrix acts as a sieve through which DNA molecules pass; large molecules (with a larger effective volume) have more difficulty passing through the pores of the gel and thus migrate through the gel more slowly than do smaller DNAs. This means that once the gels have been electrophoresed or “run” for a given time, molecules of different sizes are separated because they have moved different distances through the gel.

After electrophoresis is complete, the DNA molecules can be visualized by staining the gel with fluorescent dyes like **ethidium**, which binds to DNA and intercalates between the stacked bases (see Chapter 4, Fig. 4-28). The stained DNA molecules appear as “bands” that each reveal the presence of a population of DNA molecules of a specific size.

Two alternative kinds of gel matrices are used: **polyacrylamide** and **agarose**. Polyacrylamide has high resolving capability but can separate DNAs over only a narrow size range. Thus, electrophoresis through polyacrylamide can resolve DNAs that differ from each other in size by as little as a single base pair but only with molecules of up to several hundred (just under 1000) base pairs. Agarose has less resolving power than polyacrylamide but can separate DNA molecules of up to tens, and even hundreds, of kilobases.

Very long DNAs are unable to penetrate the pores even in agarose. Instead, they snake their way through the matrix with one end leading the way and the other end trailing behind. As a consequence, DNA molecules above a certain size (30–50 kb) migrate to a similar extent and cannot be resolved. These very long DNAs can be resolved from one another, however, if the electric field is

applied in pulses that are oriented orthogonally to each other. This technique is known as **pulsed-field gel electrophoresis** (Fig. 7-2). Each time the orientation of the electric field changes, the DNA molecule, which is snaking its way through the gel, must reorient to the direction of the new field. The larger the DNA, the longer it takes to reorient. Pulsed-field gel electrophoresis can be used to determine the size of large genomic DNAs, even entire bacterial chromosomes and chromosomes of lower eukaryotes, such as fungi—DNA molecules of up to several megabases in length.

Electrophoresis separates DNA molecules not only according to their molecular weight, but also according to their shape and topological properties (see Chapter 4). A circular DNA molecule that is relaxed or nicked migrates more slowly than does a linear molecule of equal mass. In addition, as we have seen, supercoiled DNAs, which are compact and have a small effective volume, migrate more rapidly during electrophoresis than do less supercoiled or relaxed circular DNAs of equal mass.

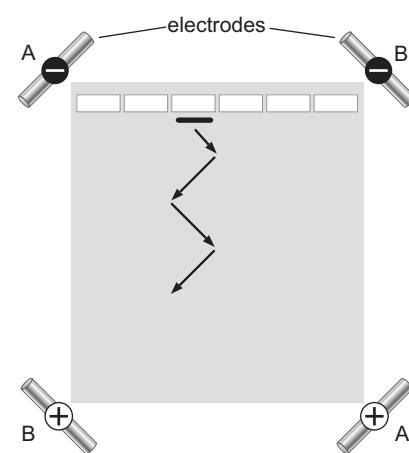
Electrophoresis is used to separate RNAs as well. Linear double-stranded DNAs have a uniform secondary structure, and their rate of migration during electrophoresis is proportional to their molecular weight. Like DNAs, RNAs have a uniform negative charge. But RNA molecules are usually single-stranded and have, as we have seen (Chapter 5), extensive secondary and tertiary structures, which influences their electrophoretic mobility. To eliminate this variable, RNAs can be treated with reagents, such as glyoxal, that react with the RNA to prevent the formation of base pairs (glyoxal forms adducts with amino groups in the bases, thereby preventing base pairing). Glyoxylated RNAs are unable to form secondary or tertiary structures and hence migrate with a mobility that is approximately proportional to their molecular weight. As we shall see later, electrophoresis is used in a similar way to separate proteins on the basis of their size.

### Restriction Endonucleases Cleave DNA Molecules at Particular Sites

Most naturally occurring DNA molecules are much larger than can readily be managed, or analyzed, in the laboratory. For example, chromosomes are extremely long single DNA molecules that can contain thousands of genes and more than 100 Mb of DNA (see Chapter 8). If we are to study individual genes and individual sites on DNA, the large DNA molecules found in cells must be broken into manageable fragments. This can be done using **restriction endonucleases** that cleave DNA at particular sites by recognizing specific sequences.

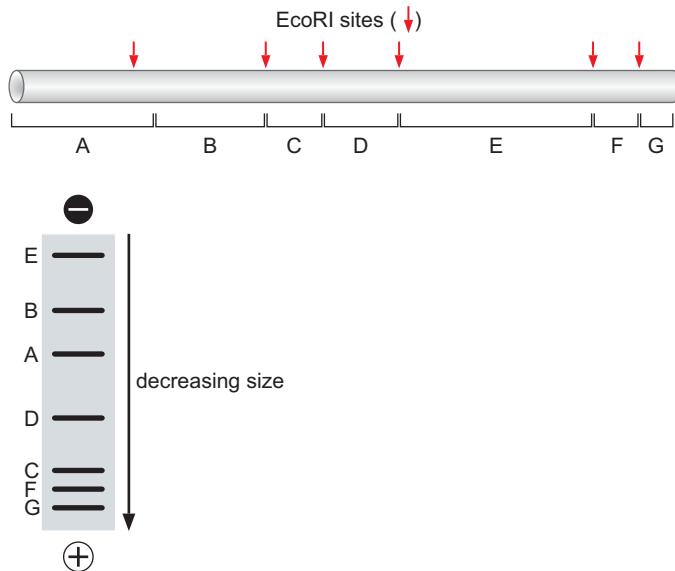
Restriction enzymes used in molecular biology typically recognize short (4–8 bp) target sequences, usually palindromic, and cut at a defined position within those sequences. Consider one widely used restriction enzyme, EcoRI, so named because it was found in certain strains of *Escherichia coli* and was the first (I) such enzyme found in that species. This enzyme recognizes and cleaves the sequence 5'-GAATTC-3'. (Because the two strands of DNA are complementary, we need specify only one strand and its polarity to describe a recognition sequence unambiguously.)

This hexameric sequence (like any other) would be expected to occur once in every 4 kb on average. (This is because there are four possible bases that can occur at any given position within a DNA sequence, and thus the chances of finding any given specific 6-bp sequence is 1 in  $4^6$ .) Consider a linear DNA molecule with six copies of the GAATTC sequence: EcoRI would cut it into seven fragments in a range of sizes reflecting the distribution of those sites in the molecule. Subjecting the EcoRI-cut DNA to electrophoresis through a gel separates the seven fragments from each other on the



**FIGURE 7-2** Pulsed-field gel electrophoresis. In this figure, the agarose gel is shown from above with a series of sample wells at the top of the gel. A and B represent two sets of electrodes. These are switched on and off alternately, as described in the text. When A is on, the DNA is driven toward the bottom right corner of the gel, where the anode of that pair is situated. When A is switched off and B is switched on, the DNA moves toward the bottom left corner. The arrows thus show the path followed by the DNA as electrophoresis proceeds. (Adapted, with permission, from *Watson J., et al. Molecular cloning: A laboratory manual*, 3rd ed., Fig. 5-7. © Cold Spring Harbor Laboratory Press.)

**FIGURE 7-3** Digestion of a DNA fragment with endonuclease EcoRI. At the top is shown a DNA molecule and the positions within it at which EcoRI cleaves. When the molecule, digested with that enzyme, is run on an agarose gel, the pattern of bands shown is observed.

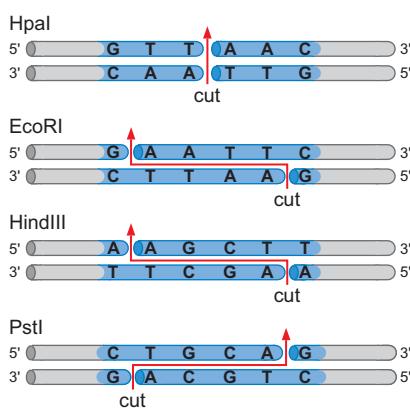


basis of their different sizes (Fig. 7-3). Thus, in the experiment shown, EcoRI has dissected the DNA into specific fragments, each corresponding to a particular region of the molecule.

Cleaving the same DNA molecule with a different restriction enzyme—for example, HindIII, which also recognizes a 6-bp target but of a different sequence (5'-AAGCTT-3')—cuts the molecule at different positions and generates fragments of different sizes. Thus, the use of multiple enzymes allows different regions of a DNA molecule to be isolated. It also allows a given molecule to be identified based on the characteristic series of patterns when the DNA is digested with a set of different enzymes.

Other restriction enzymes such as Sau3A1 (which is found in the bacterium *Staphylococcus aureus*) recognize tetrameric sequences (5'-GATC-3') and thus cut DNA more frequently, approximately once every 250 bp. At the other extreme are enzymes that recognize octomeric sequences such as NotI, which recognizes the octameric sequence 5'-GC<sub>6</sub>CCGC-3' and cuts, on average, only once every 65 kb (Table 7-1). Of note, some restriction enzymes are sensitive to methylation. That is, methylation of a base (or bases) within a recognition sequence inhibits enzyme activity at that site.

Restriction enzymes differ not only in the specificity and length of their recognition sequences but also in the nature of the DNA ends they generate. Thus, some enzymes, such as HpaI, generate flush or “blunt” ends; others, such as EcoRI, HindIII, and PstI, generate staggered ends (Fig. 7-4). For example, EcoRI cleaves covalent (phosphodiester) bonds between G and A at staggered positions on each strand. The hydrogen bonds between the 4 bp between these cut sites are easily broken to generate 5' protruding ends of 4 nucleotides in length (Fig. 7-5). Note that these ends are complementary



**FIGURE 7-4** Recognition sequences and cut sites of various endonucleases. As shown, different endonucleases not only recognize different target sites but also cut at different positions within those sites. Thus, molecules with blunt ends or with 5' or 3' overhanging ends can be generated.

**TABLE 7-1** Some Restriction Endonucleases and Their Recognition Sequences

Enzyme	Sequence	Cut Frequency <sup>a</sup>
Sau3A1	5'-GATC-3'	0.25 kb
EcoRI	5'-GAATTC-3'	4 kb
NotI	5'-GC <sub>6</sub> CCGC-3'	65 kb

<sup>a</sup>Frequency = 1/4<sup>n</sup>, where n is the number of base pairs in the recognition sequence.

to each other. They are said to be “sticky” because they readily anneal through base pairing to each other or to other DNA molecules cut with the same enzyme. This is a useful property that we consider in our discussion of DNA cloning.

### DNA Hybridization Can Be Used to Identify Specific DNA Molecules

As we saw in Chapter 4, the capacity of denatured DNA to reanneal (i.e., to re-form base pairs between complementary strands) allows for the formation of hybrid molecules when homologous, denatured DNAs from two different sources are mixed with each other under the appropriate conditions of ionic strength and temperature. This process of base pairing between complementary single-stranded polynucleotides is known as **hybridization**.

Many techniques rely on the specificity of hybridization between two DNA molecules of complementary sequence. For example, this property is the basis for detecting specific sequences within complicated mixtures of nucleic acids. In this case, one of the molecules is a **probe** of defined sequence—either a purified fragment or a chemically synthesized DNA molecule. The probe is used to search mixtures of nucleic acids for molecules containing a complementary sequence. The probe DNA must be labeled so that it can be readily located once it has found its target sequence. The mixture being probed is typically either separated by size on a gel or distributed as a library of clones (see later discussion).

There are two basic methods for labeling DNA. The first involves adding a label to the end of an intact DNA molecule. Thus, for example, the enzyme polynucleotide kinase adds the  $\gamma$ -phosphate from ATP to the 5'-OH group of DNA. If that phosphate is radioactive, this process labels the DNA molecule to which it is transferred.

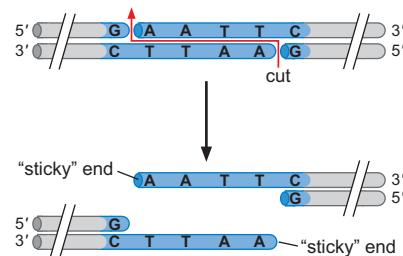
Labeling by incorporation (the other mechanism) involves synthesizing new DNA in the presence of a labeled precursor. This approach is often performed by using PCR with a labeled precursor, or even by hybridizing short random hexameric oligonucleotides to DNA and allowing a DNA polymerase to extend them. The labeled precursors are most commonly nucleotides modified with either a fluorescent moiety or radioactive atoms. Typically, the fluorescent moiety need only be attached to the base of one of the four nucleotides used as precursors for DNA synthesis ( $\sim 25\%$  of labeling is generally sufficient for most purposes).

DNA labeled with fluorescent precursors can be detected by illuminating the DNA sample with appropriate wavelength UV light and monitoring the longer-wavelength light that is emitted in response. Radioactively labeled precursors typically have radioactive  $^{32}\text{P}$  or  $^{35}\text{S}$  incorporated into the  $\alpha$ -phosphate of one of the four nucleotides. This phosphate is retained in the product DNA (see Chapter 9). Radioactive DNA can be detected by exposing the sample of interest to X-ray film or by photomultipliers that emit light in response to excitation by the  $\beta$  particles emitted from  $^{32}\text{P}$  and  $^{35}\text{S}$ .

There are many ways that hybridization is used in the identification of specific DNA or RNA fragments. The two most common are described later.

### Hybridization Probes Can Identify Electrophoretically Separated DNAs and RNAs

It is often desirable to monitor the abundance or size of a particular DNA or RNA molecule in a population of many other similar molecules. For



**FIGURE 7-5** Cleavage of an EcoRI site. EcoRI cuts the two strands within its recognition site to give 5' overhanging ends. These are called “sticky” ends—they readily adhere to other molecules cut with the same enzyme because they provide complementary single-strand ends that come together through base pairing.

example, this can be useful when determining the amount of a specific mRNA that is expressed in two different cell types or the length of a restriction fragment that contains the gene being studied. This type of information can be obtained using blotting methods that localize specific nucleic acids after they have been separated by electrophoresis.

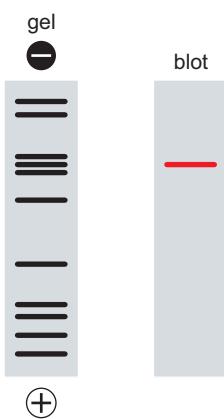
Suppose that the yeast genome has been cleaved with the restriction enzyme EcoRI and the investigator wants to identify or know the size of the fragment that contains the gene of interest. When stained with ethidium bromide, the thousands of DNA fragments generated by cutting the yeast genome are too numerous to resolve into discretely visible bands, and they look like a smear centered around  $\sim 4$  kb. The technique of **Southern blot hybridization** (named after its inventor Edward Southern) will identify within the smear the size of the particular fragment containing the gene of interest.

In this procedure, the cut DNA is separated by gel electrophoresis, and the gel is soaked in alkali to denature the double-stranded DNA fragments. These fragments are then transferred from the gel to a positively charged membrane to which they adhere, creating an imprint, or “blot,” of the gel. During the transfer process, the DNA fragments are bound to the membrane in positions that mirror their corresponding positions in the gel after electrophoresis. After DNAs of interest are bound to the membrane, the charged membrane is incubated with a mixture of nonspecific DNA fragments to saturate all of the remaining binding sites on the membrane. Because the DNA in this mixture is randomly distributed on the membrane and, if chosen properly, will not contain the sequence of interest (e.g., from a different organism than the probe DNA), it will not interfere with subsequent detection of a specific gene.

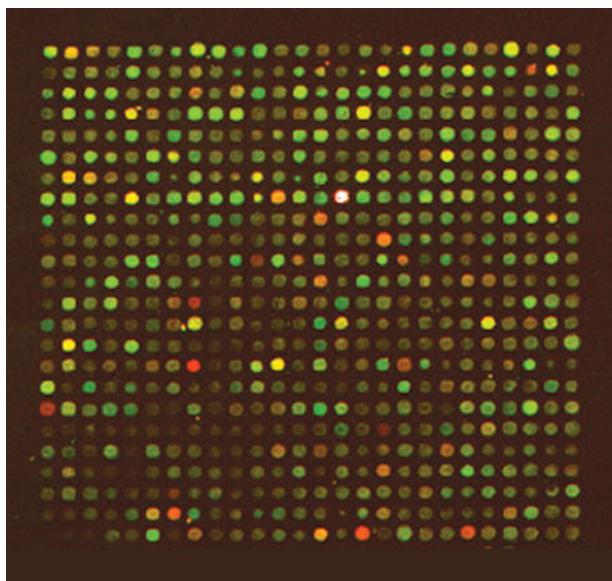
The DNA bound to the membrane is then incubated with probe DNA containing a sequence complementary to a sequence within the gene of interest. Because all of the nonspecific binding sites on the membrane are occupied with unrelated DNA, the only way that the probe DNA can associate with the membrane is by hybridizing to any complementary DNA present on the membrane. This probing is performed under conditions of salt concentration and temperature close to those at which nucleic acids denature and renature. Under these conditions, the probe DNA will hybridize tightly to only its exact complement. Often the probe DNA is in high molar excess compared with its immobilized target on the filter, thereby favoring probe hybridization rather than reannealing of the denatured DNA on the blot. In addition, the immobilization of the denatured DNA on the filter tends to interfere with renaturation. A variety of films or other media sensitive to the light or electrons emitted by the labeled DNA can detect where on the blot the probe hybridizes. For example, when a radioactively labeled probe DNA is used and X-ray film is exposed to the filter and then developed, an **autoradiogram** is produced in which the pattern of exposure on the film corresponds to the position of the hybrids on the blot (Fig. 7-6).

A similar procedure called **northern blot hybridization** (to distinguish it from Southern blot hybridization) can be used to identify a particular mRNA in a population of RNAs. Because mRNAs are relatively short (typically  $<5$  kb), there is no need for them to be digested with any enzymes (there are only a limited number of specific RNA-cleaving enzymes). Otherwise, the protocol is similar to that described for Southern blotting. The separated mRNAs are transferred to a positively charged membrane and probed with a probe DNA of choice. (In this case, hybrids are formed by base pairing between complementary strands of RNA and DNA.)

An investigator might perform northern blot hybridization to ascertain the amount of a particular mRNA present in a sample rather than its size. This measure is a reflection of the level of expression of the gene that



**FIGURE 7-6** A Southern blot. DNA fragments, generated by digestion of a DNA molecule by a restriction enzyme, are run out on an agarose gel. Once stained, a pattern of fragments is seen. When transferred to a filter and probed with a DNA fragment homologous to just one sequence in the digested molecule, a single band is seen, corresponding to the position on the gel of the fragment containing that sequence.



**FIGURE 7-7** Microarray grid comparing expression patterns in two tissues (muscles and neurons) in *Caenorhabditis elegans*. Each circle in the grid contains a short DNA segment from the coding region of a single gene in the *C. elegans* genome. RNA was extracted from muscles and neurons, and labeled with fluorescent dyes (red and green, respectively). Thus, the red circles indicate genes expressed in muscle, whereas the green circles reflect genes expressed in neurons. The yellow circles indicate genes expressed in both cell types. It is clear that the two samples express distinct sets of genes. (Courtesy of Stuart Kim, Stanford University.)

encodes that mRNA. Thus, for example, one might use northern blot hybridization to ask how much more mRNA of a specific type is present in a cell treated with an inducer of the gene in question compared with an uninduced cell. As another example, northern blot hybridization might be performed to compare the relative levels of a particular mRNA (and hence the expression level of the gene in question) among different tissues of an organism. Because an excess of DNA probe is used in these assays, the amount of hybridization is related to the amount of mRNA present in the original sample, allowing the relative amounts of mRNA to be determined.

The principles of Southern and northern blot hybridization also underlie microarray analysis, which we consider in the Genomics section of this chapter. The availability of complete sequence information has enabled development of this “reverse hybridization” experiment. A microarray is constructed by attaching several hundred to thousands of known DNA sequences to a solid surface, typically a glass or plastic slide (Fig. 7-7). Each sequence is derived from a different gene in the organism under study. When describing microarray analysis, the terms used are the reverse of their use in Southern or northern analysis. In microarray analysis, the fixed, unlabeled sequences are called the “probes,” because these are known DNA sequences, whereas the “target” is composed of amplified, labeled cDNAs generated from the total RNA from a cell or tissue. When target sequences are hybridized to the array of probe DNAs, the intensity of the hybridization signal to each DNA species in the array is a measure of the level of expression of the gene in question.

### Isolation of Specific Segments of DNA

Much of the molecular analysis of genes and their function requires the separation of specific segments of DNA from much larger DNA molecules and their selective amplification. Isolating a large amount of a single pure DNA molecule facilitates the analysis of the information encoded in that particular DNA molecule. Thus, the DNA can be sequenced and analyzed, or it can be cloned and expressed to allow the study of its protein product.

The ability to purify specific DNA molecules in significant quantities allows them to be manipulated in various other ways as well. For example,

recombinant DNA molecules can be created and used to alter the expression of a particular gene (e.g., by fusing its coding sequence to a heterologous promoter). Alternatively, purified DNA sequences can be recombined to generate DNAs that encode so-called **fusion proteins**—that is, hybrid proteins made up of parts derived from different proteins. The techniques of DNA cloning and amplification by PCR have become essential tools in asking questions regarding the control of gene expression, maintenance of the genome, and protein function.

## DNA Cloning

The ability to construct recombinant DNA molecules and maintain them in cells is called **DNA cloning**. This process typically involves a vector that provides the information necessary to propagate the cloned DNA in the replicating host cell. Key to creating recombinant DNA molecules are the restriction enzymes that cut DNA at specific sequences and other enzymes that join the cut DNAs to one another. By creating recombinant DNA molecules that can be propagated in a host organism, a particular DNA fragment can be both purified from other DNAs and amplified to produce large quantities.

In the remainder of this section, we describe how DNA molecules are cut, recombined, and propagated. We then discuss how large collections of such hybrid molecules, called **libraries**, can be created. In a library, a common vector carries many alternative inserts. We describe how libraries are made and how specific DNA segments can be identified and isolated from them.

Once DNA is cleaved into fragments, it typically needs to be inserted into a vector for propagation. That is, the DNA fragment must be inserted into a second DNA molecule (the vector) to be replicated in a host organism. The most common host used to propagate DNA is the bacterium *E. coli*. Vector DNAs typically have three characteristics.

1. They contain an origin of replication that allows them to replicate independently of the chromosome of the host. (Note that some yeast vectors also require a centromere.)
2. They contain a selectable marker that allows cells that contain the vector (and any attached DNA) to be readily identified.
3. They have unique sites for one or more restriction enzymes. This allows DNA fragments to be inserted at a defined point within the vector such that the insertion does not interfere with the first two functions.

Many common vectors are small (~3 kb) circular DNA molecules called **plasmids**. These molecules were originally derived from extrachromosomal circular DNA molecules that are found naturally in many bacteria and single-cell eukaryotes (Appendix 1). In many cases (although not in yeast), these DNAs carry genes encoding resistance to antibiotics. Thus, naturally occurring plasmids already have two of the characteristics desirable for a vector: they can propagate independently in the host, and they carry a selectable marker. A further benefit is that these plasmids are sometimes present in multiple copies per cell. This increases the amount of DNA that can be isolated from a population of cells. Naturally occurring plasmids typically are restricted in the amount of DNA that can be carried (typically limited to 1–10 kb). For the cloning and propagation of large fragments, typically used in genomic analysis and for DNA sequencing as we discuss later, various artificial vector constructs have been created, for example, bacterial and yeast artificial chromosomes (BACs and YACs) that can accommodate from 120 to >500 kb of DNA.

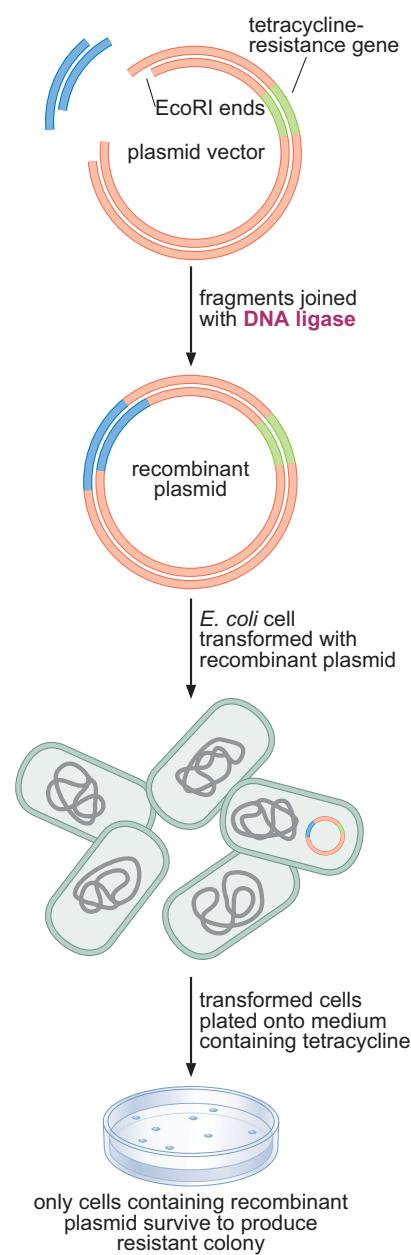
Inserting a fragment of DNA into a vector is generally a relatively simple process (Fig. 7-8). Suppose that a plasmid vector has a unique recognition site for EcoRI. The vector is prepared by digesting it with EcoRI, which linearizes the plasmid. Because EcoRI generates protruding 5' ends that are complementary to each other (see Fig. 7-5), the sticky ends are capable of reannealing to re-form a circle with two nicks. Treatment of the circle with the enzyme **DNA ligase** and ATP would seal the nicks to re-form a covalently closed circle. The target DNA is prepared by cleaving it with a restriction enzyme, in this case with EcoRI, to generate potential insert DNAs. Vector DNA is mixed with an excess of insert DNAs cleaved by EcoRI under conditions that allow sticky ends to hybridize. DNA ligase is then used to link the compatible ends of the two DNAs. Adding an excess of the insert DNA relative to the plasmid DNA ensures that the majority of vectors will reseal with insert DNA incorporated (Fig. 7-8).

Some vectors not only allow the isolation and purification of a particular DNA but also drive the expression of genes within the insert DNA. These plasmids are called **expression vectors** and have transcriptional promoters, derived from the host cell, immediately adjacent to the site of insertion. If the coding region of a gene (without its promoter) is placed at the site of insertion in the proper orientation, then the inserted gene will be transcribed into mRNA and translated into protein by the host cell. Expression vectors are frequently used to express heterologous or mutant genes to assess their function. They can also be used to produce large amounts of a protein for purification. In addition, the promoter in the expression vector can be chosen such that expression of the insert is regulated by the addition of a simple compound to the growth media (e.g., a sugar or an amino acid) (see Chapter 18 for a discussion of transcriptional regulation in prokaryotes). This ability to control when the gene will be expressed is particularly useful if the gene product is toxic.

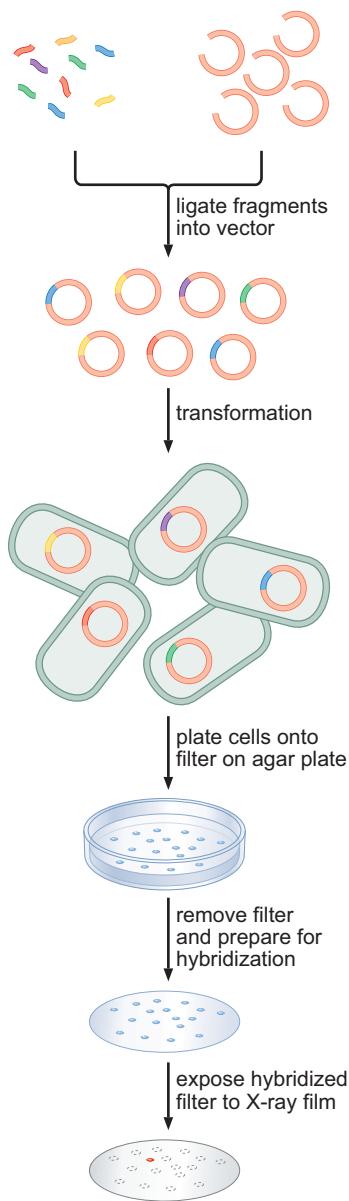
### Vector DNA Can Be Introduced into Host Organisms by Transformation

Propagation of the vector with its insert DNA is achieved by introducing the recombinant DNA into a host cell by transformation. As we discussed in Chapter 2, **transformation** is the process by which a host organism can take up DNA from its environment. Some bacteria, but not *E. coli*, can do this naturally and are said to have **genetic competence**. *E. coli* can be rendered competent to take up DNA, however, by treatment with calcium ions. Although the exact mechanism for DNA uptake is not known, it is likely that the  $\text{Ca}^{2+}$  ions shield the negative charge on the DNA, allowing it to pass through the cell membrane. Thus, calcium-treated *E. coli* cells are said to be competent to be transformed. An antibiotic to which the plasmid imparts resistance is then included in the growth medium to select for the growth of cells that have taken up the plasmid DNA—these cells are called **transformants**. Cells harboring the plasmid will be able to grow in the presence of the antibiotic, whereas those lacking it will not.

Transformation is a relatively inefficient process. Only a small percentage of the DNA-treated cells take up the plasmid. It is this low efficiency of transformation that makes it necessary to use selection with the antibiotic. The inefficiency of transformation also ensures that, in most cases, each cell receives only a single molecule of DNA. This property makes each transformed cell and its progeny a carrier of a unique DNA molecule. Thus, transformation effectively purifies and amplifies one DNA molecule away from all other DNAs in the transforming mixture.



**FIGURE 7-8 Cloning in a plasmid vector.** A fragment of DNA, generated by cleavage with EcoRI, is inserted into the plasmid vector linearized by that same enzyme. Once ligated (see text), the recombinant plasmid is introduced into bacteria by transformation (see text). Cells containing the plasmid can be selected by growth on the agar plates that contain growth media including antibiotic to which the plasmid confers resistance. (Adapted, with permission, from Micklos D.A. and Freyer G.A. 2003. *DNA science: A first course*, 2nd ed., p. 129. © Cold Spring Harbor Laboratory Press.)



**FIGURE 7-9** Construction and probing of a DNA library. To construct the library, genomic DNA and vector DNA, digested with the same restriction enzyme, are incubated together with ligase. The resulting pool or library of hybrid vectors (each vector carrying a different insert of genomic DNA, represented in a different color) is then introduced into *E. coli*, and the cells are plated onto a filter placed over agar medium. Once colonies have grown, the filter is removed from the plate and prepared for hybridization: cells are lysed, the DNA is denatured, and the filter is incubated with a labeled probe. The clone of interest is identified by autoradiography.

### Libraries of DNA Molecules Can Be Created by Cloning

A **DNA library** is a population of identical vectors that each contains a different DNA insert (Fig. 7-9). To construct a DNA library, the target DNA (e.g., human genomic DNA) is digested with a restriction enzyme that gives a desired average insert size, ranging from <100 bp to more than a megabase (for very large insert sizes, the DNA is typically incompletely cut with a restriction enzyme). The cleaved DNA is then mixed with the appropriate vector cut with the same restriction enzyme in the presence of ligase. This creates a large collection of vectors with different DNA inserts.

Different kinds of libraries are made using insert DNA from different sources. The simplest are derived from total genomic DNA cleaved with a restriction enzyme; these are called **genomic libraries**. This type of library is most useful when generating DNA for sequencing a genome. If, on the other hand, the objective is to clone a DNA fragment encoding a particular gene, then a genomic library can be used efficiently only when the organism in question has relatively little non-coding DNA. For an organism with a more complex genome, this type of library is not suitable for this task because many of the DNA inserts will not contain coding DNA sequences.

To enrich for coding sequences in the library, a **cDNA library** is created. This is made as shown in Figure 7-10. Instead of starting with genomic DNA, mRNAs are converted into DNA sequences. The process that allows this is called **reverse transcription** and is performed by a special DNA polymerase (reverse transcriptase) that can make DNA from an RNA template (see Chapter 12). When treated with reverse transcriptase, mRNA sequences are converted into double-stranded DNA copies called **cDNAs** (for “copy DNAs”). From this point on, construction of the library follows the same strategy as does construction of a genomic library—the cDNA products and vector are treated with the same restriction enzyme, and the resulting fragments are then ligated into the vector.

To isolate individual inserts from a library, host cells (usually *E. coli*) are transformed with the entire library. Each transformed cell contains only a single vector with its associated insert DNA. Thus, each cell that propagates after transformation will contain multiple copies of just one of the possible clones from the library. The colony produced from cells carrying any cloned sequence of interest can be identified and the DNA retrieved. There are various ways to identify the clone. For example, as we describe later, hybridization with a unique DNA or RNA probe can identify a colony of cells that include a particular insert DNA.

### Hybridization Can Be Used to Identify a Specific Clone in a DNA Library

When attempting to clone a gene, a common step is to identify fragments of that gene among clones in a library. This can be achieved using a DNA probe whose sequence matches part of the gene of interest. Such a probe can be used to identify the particular colony of cells harboring clones containing that region of the gene, as we now describe.

The process by which a labeled DNA probe is used to screen a library is called *colony hybridization*. A typical cDNA library will have thousands of different inserts, each contained within a common vector (see above). After transformation of a suitable bacterial host strain with the library, the cells are plated out on Petri dishes containing solid growth medium (usually agar; see Appendix 1). Each cell grows into an isolated colony of cells, and each cell within a given colony contains the same vector and insert from the library (there are typically a few hundred colonies per dish).

The same type of positively charged membrane filter used in the Southern and northern blotting techniques is used here to secure small amounts of DNA for probing. In this case, pieces of the membrane are pressed on top of the dish of colonies, and imprints of cells (including the DNA contained within the cells) from each colony are lifted onto the filter (note that some cells from each colony remain on the plate). Thus, the filter retains a sample of each DNA clone positioned on the filter in a pattern that matches the pattern of colonies on the plate. This ensures that once the desired clone has been identified by probing the filter, the colony of cells carrying that clone can be readily identified on the plate and the plasmid containing the appropriate insert DNA can be purified from these cells.

Probing of the filters is performed as follows: the filters are treated under conditions that cause the cells on the membrane to break open and the DNA to leak out and bind to the filter at the same location as the cells from which the DNA was derived. The filters can then be incubated with the labeled probe under the same conditions that were used in the northern and Southern blotting experiments.

As we mentioned above and discuss in Appendix 1, bacteriophage (particularly  $\lambda$ ) have also been modified for use as vectors. When libraries are made using a phage vector, they can be screened in much the same way as just described for the screening of plasmid libraries. The difference is that the plaques formed by growth of the phage on bacterial lawns are screened rather than colonies (see Appendix 1).

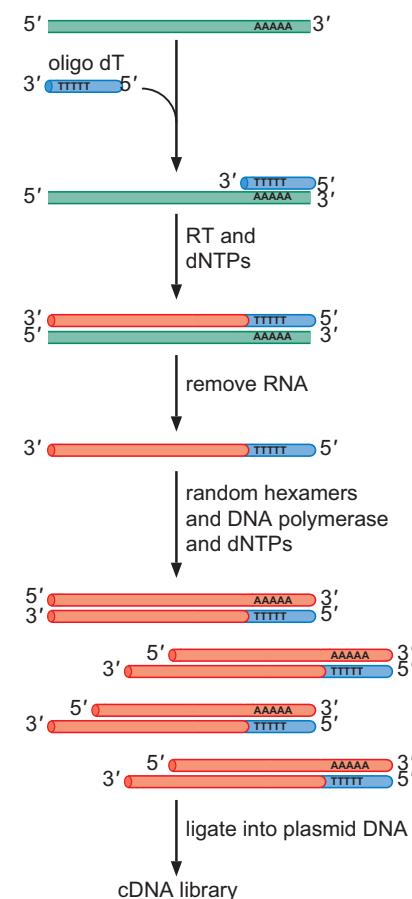
### Chemical Synthesis of Defined DNA Sequences

Many believe that the modern era of molecular biology was launched by the development of methods for the chemical synthesis of short, custom-designed segments of single-stranded DNA (ssDNA), known as **oligonucleotides**. The most common methods of chemical synthesis are performed on solid supports using machines that automate the process. The precursors used for nucleotide addition are chemically protected molecules called **phosphoamidines** (Fig. 7-11). In contrast to the direction of chain growth used by DNA polymerases (see Chapter 9), growth of the DNA chain is by addition to the 5' end of the molecule.

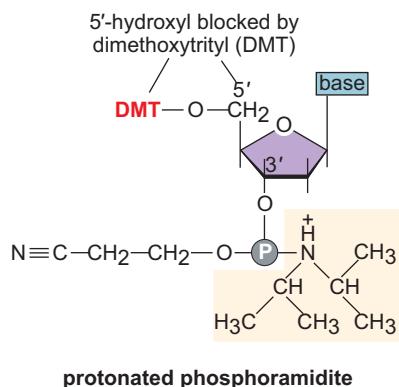
Chemical synthesis of DNA molecules 10–100 bases long is efficient and accurate. It is a routine procedure: an investigator can simply program a DNA synthesizer to make any desired sequence by typing the base sequence into a computer controlling the machine. But as the synthetic molecules get longer, the final product is less uniform because of the inherent failures that occur during any cycle of the process. Thus, molecules >100 nucleotides long are difficult to synthesize in the quantity and with the accuracy desirable for most molecular analysis.

The short ssDNA segments that can readily be made, however, are well suited for many purposes. For example, a custom-designed oligonucleotide harboring a mismatch to a segment of cloned DNA can be used to create a directed mutation in that cloned DNA. This method, called **site-directed mutagenesis**, is performed as follows: the oligonucleotide is hybridized to the cloned DNA fragment and used to prime DNA synthesis with the cloned DNA as template. In this way, a double-strand molecule with one mismatch is made. The two strands are separated, and the strand with the desired mismatch is amplified further.

Custom-designed oligonucleotides can be used in this manner to introduce recognition sequences for restriction enzymes, which can be used to create various recombinant DNAs, such as fusions between the coding regions of two different genes or the fusion of a promoter from one gene



**FIGURE 7-10 Construction of a cDNA library.** The RNA-dependent DNA polymerase reverse transcriptase (RT) transcribes RNA into DNA (copy, or cDNA). In the first step (first-strand synthesis), oligos of poly-T sequence serve as primers by hybridizing to the poly-A tails of the mRNAs. (cDNA libraries are typically made from eukaryotic cells whose mRNA have poly-A tails at their 3' ends; see Chapter 19.) Reverse transcriptase extends the dT primer to complete a DNA copy of the mRNA template. The product is a duplex composed of one strand of mRNA and its complementary strand of DNA. The RNA strand is removed by treatment with base (NaOH), and the remaining single-stranded DNA now serves as template for the second step (second-strand synthesis). Short random sequences of DNA usually ~6 bp long (called random hexamers) serve as primers by hybridizing to various sequences along the copy DNA template. These primers are then extended by DNA polymerase to create double-stranded DNA products that can be cloned into a plasmid vector (see Fig. 7-8) to create a cDNA library.



**FIGURE 7-11** Protonated phosphoramidite. As shown, the 5'-hydroxyl group is blocked by the addition of a dimethoxytrityl protecting group.

with the coding region of another. Alternatively, introduced mutations can change the sequence encoding a particular amino acid in a gene. By comparing the properties of the resulting mutant protein to the wild-type protein, researchers can test the importance of the specific amino acid for that protein's function.

Custom-designed oligonucleotides are critical in PCR, which we describe next, and are an indispensable feature of the DNA-sequencing strategies that we describe later. Therefore, a common feature in designing experiments to construct new molecular clones of genes, to detect specific DNAs, to amplify DNAs, and to sequence DNAs is to design and have synthesized a short synthetic ssDNA oligonucleotide of the desired sequence.

### The Polymerase Chain Reaction Amplifies DNAs by Repeated Rounds of DNA Replication In Vitro

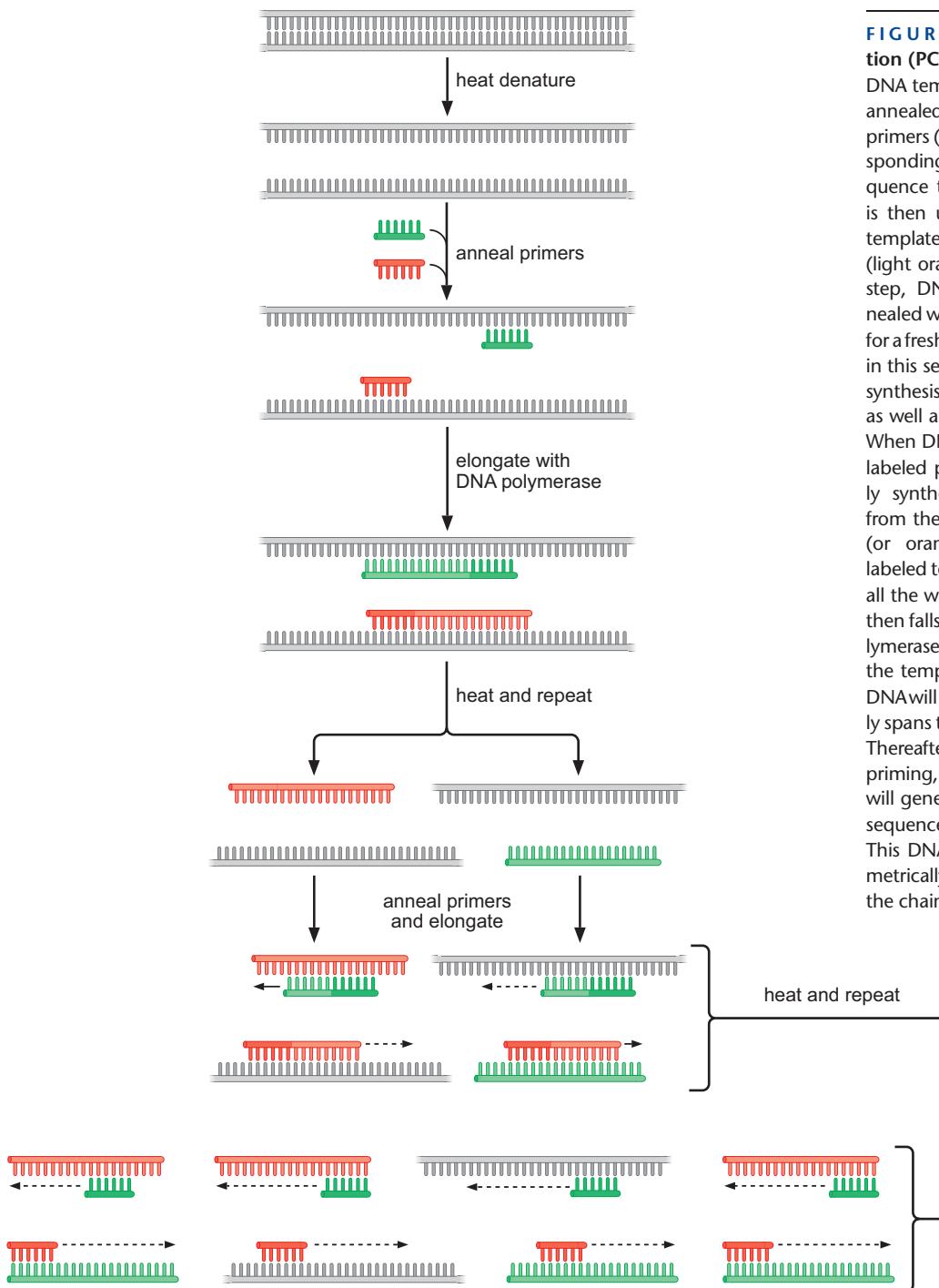
The game-changing method for amplifying particular segments of DNA, distinct from cloning and propagation within a host cell, is the **polymerase chain reaction (PCR)**. This procedure is performed entirely biochemically, that is, *in vitro*. PCR uses the enzyme DNA polymerase that directs the synthesis of DNA from deoxynucleotide substrates on a single-stranded DNA template. As you will in Chapter 9, DNA polymerase adds nucleotides to the 3' end of a custom-designed oligonucleotide when it is annealed to a longer template DNA. Thus, if a synthetic oligonucleotide or primer is annealed to a single-strand template that contains a region complementary to the oligonucleotide, DNA polymerase can use the oligonucleotide as a primer and elongate its 3' end to generate an extended region of double-stranded DNA.

How is this enzyme and reaction exploited to amplify specific DNA sequences? Two synthetic, single-strand oligonucleotides are synthesized. One is complementary in sequence to the 5' end of one strand of the DNA to be amplified, and the other is complementary to the 5' end of the opposite strand (Fig. 7-12). The DNA to be amplified is then denatured, and the oligonucleotides are annealed to their target sequences. At this point, DNA polymerase and deoxynucleotide substrates are added to the reaction, and the enzyme extends the two primers. This reaction generates double-stranded DNA over the region of interest on *both* strands of DNA. Thus, two double-stranded copies of the starting fragment of DNA are produced in this, the first, cycle of the PCR.

Next, the DNA is subjected to another round of denaturation and DNA synthesis using the same primers. (Note that only the sequence between the primers is, in fact, precisely amplified.) This process generates four copies of the fragment of interest. In this way, additional repeated cycles of denaturation and primer-directed DNA synthesis amplify the region between the two primers in a geometric manner (2, 4, 8, 16, 32, 64, etc.). Thus, a fragment of DNA that was originally present in vanishingly small amounts is amplified into a large quantity of a double-stranded DNA (see Fig. 7-12; Box 7-1, Forensics and the Polymerase Chain Reaction). Indeed, after 20 to 30 cycles of PCR a DNA sequence that is undetectable among millions of others (e.g., one sequence in the entire human genome) can be readily identified as a single band on an agarose DNA gel.

### Nested Sets of DNA Fragments Reveal Nucleotide Sequences

We next consider how nucleotide sequences are determined. We first describe “classical” methods of DNA sequencing that were used to



**FIGURE 7-12** Polymerase chain reaction (PCR). In the first step of the PCR, the DNA template is denatured by heating and annealed with synthetic oligonucleotide primers (dark orange and dark green) corresponding to the boundaries of the DNA sequence to be amplified. DNA polymerase is then used to copy the single-stranded template by extension from the primers (light orange and light green). In the next step, DNA is once again denatured, annealed with primers, and used as a template for a fresh round of DNA synthesis. Note that in this second cycle, the primers can prime synthesis from the newly synthesized DNAs as well as from the original template DNA. When DNA polymerase extends the green-labeled primer that had annealed to newly synthesized (orange-labeled) template from the previous round of DNA synthesis (or orange-labeled primer from green-labeled template), the polymerase proceeds all the way to the end of the template and then falls off (in the figure [bottom], the polymerases have not yet reached the end of the templates). Thus, in this second cycle, DNA will have been synthesized that precisely spans the DNA sequence to be amplified. Thereafter, further rounds of denaturation, priming, and DNA synthesis (not shown) will generate DNAs that correspond to the sequence interval set by the two primers. This DNA will increase in abundance geometrically with each subsequent cycle of the chain reaction.

determine the nucleotide sequences of individual genes. We then discuss how this method was automated for the sequencing of entire genomes. Finally, we consider “next-generation” sequencing methods that are now used to produce personalized genomes.

It is now possible to determine the entire sequence of nucleotides for a genome, as has now been done for organisms ranging in complexity from bacteria to man. This allows us to find any specific sequence with great rapidity and accuracy (as discussed later in this chapter).

The underlying principle of conventional DNA sequencing is based on the separation, by size, of nested sets of DNA molecules. Each of the DNA

## TECHNIQUES

### Box 7-1 Forensics and the Polymerase Chain Reaction

Imagine being in a forensic laboratory and having a DNA sample from a suspected criminal. We want to determine whether the suspect's DNA contains a polymorphism that is present in DNA found at the scene of the crime. Polymorphisms are alternative DNA sequences (alleles) found in a population of organisms at a common, homologous region of the chromosome, such as a gene. A polymorphism can be as simple as alternative, single-base-pair differences at the same site in the chromosome among different members of the population or differences in the length of a simple nucleotide repeat sequence such as CA (see Chapter 9). What we want to do is amplify DNA surrounding and including the site of the polymorphism so that we can subject it to nucleotide sequencing (discussed later) and

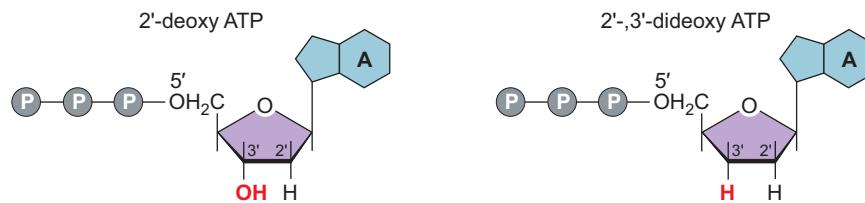
determine if there is a match to the sequence found in the crime scene sample. The nucleotide sequence of the amplified DNA helps to determine (along with checks for additional polymorphisms) whether the two DNA samples match. This approach to defining the DNA sequence is called "DNA profiling" or "DNA fingerprinting," intended as an analogy between identification using DNA and identification using conventional fingerprinting techniques. DNA profiling was first used in 1985 (the U.S. Federal Bureau of Investigation [FBI] began using the technique in 1988) and since that time has become widely used in the analysis of crime scene evidence, both to convict and to exonerate suspected individuals (see, e.g., The Innocence Project; [www.innocenceproject.org](http://www.innocenceproject.org)).

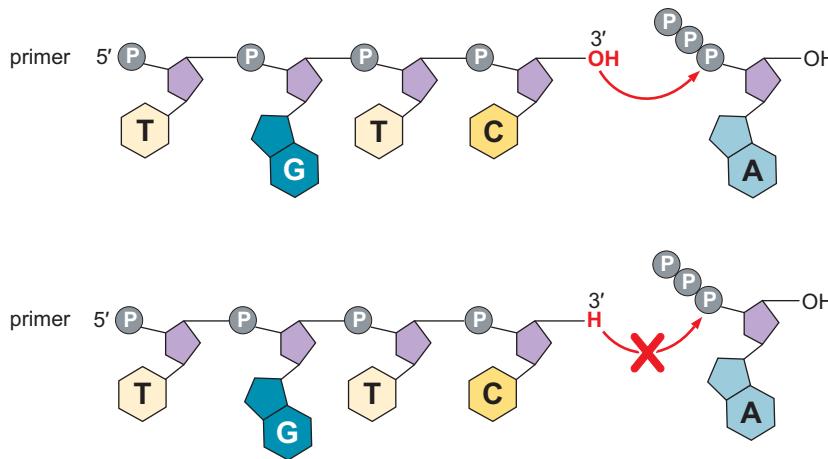
molecules starts at a common 5' end and terminates at one of many alternative 3' end points. Members of any given set have a particular type of base at their 3' ends. Thus, for one set, the molecules all end with a G, for another a C, for a third an A, and for the final set a T. Molecules within a given set (e.g., the G set) vary in length depending on where the particular G at their 3' end lies in the sequence. Each fragment from this set therefore indicates where there is a G in the DNA molecule from which they were generated. How these fragments are generated is discussed later (and is shown in Fig. 7-15).

The most commonly used procedure, which uses **chain-terminating nucleotides** and in vitro DNA synthesis, is the foundation for the original automation of DNA sequencing. In the chain-termination method, DNA is copied by DNA polymerase from a DNA template starting from a fixed point specified by hybridization of an oligonucleotide primer. DNA polymerase uses 2'-deoxynucleoside triphosphates as substrates for DNA synthesis, and DNA synthesis occurs by extending the 3' end. (The chain-termination method relies on the principles of enzymatic synthesis of DNA, which is discussed in Chapter 9.) The chain-termination method uses special, modified substrates called 2',3'-dideoxynucleotides (ddNTPs), which lack the 3'-hydroxyl group on their sugar moiety as well as the 2'-hydroxyl (Fig. 7-13). DNA polymerase will incorporate a 2',3'-dideoxynucleotide at the 3' end of a growing polynucleotide chain, but once incorporated, the lack of a 3'-hydroxyl group prevents the addition of further nucleotides, causing elongation to terminate (Fig. 7-14).

Now suppose that we "spike" (delete or add) a cocktail of the nucleotide substrates with the modified substrate 2',3'-dideoxyguanosine triphosphate (ddGTP) at a ratio of one ddGTP molecule to 100 2'-deoxy-GTP molecules (dTTP). This will cause DNA synthesis to abort at a frequency of one in 100 times the DNA polymerase encounters a C on the template strand (Fig. 7-15a). Because all of the DNA chains commence growth from the

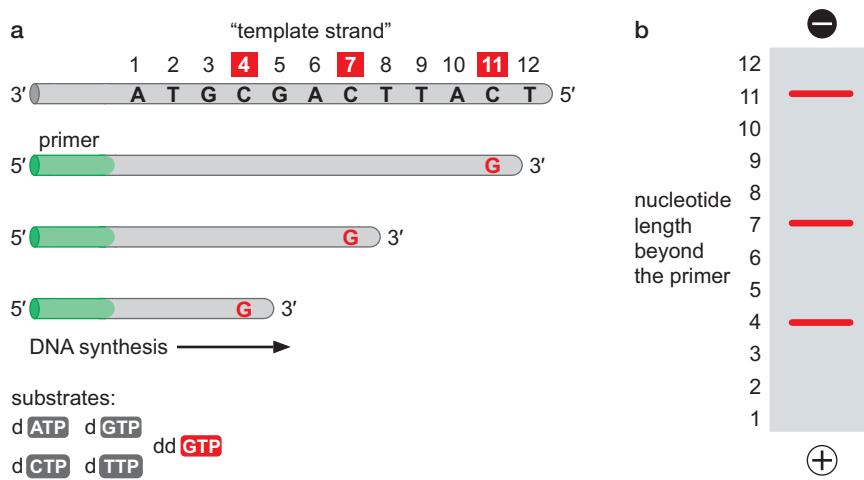
**FIGURE 7-13** Dideoxynucleotides used in DNA sequencing. On the left is 2'-deoxy ATP. This can be incorporated into a growing DNA chain and allow another nucleotide to be incorporated directly after it. On the right is 2',3'-dideoxy ATP. This can be incorporated into a growing DNA chain, but once in place it blocks further nucleotides being added to the same chain.



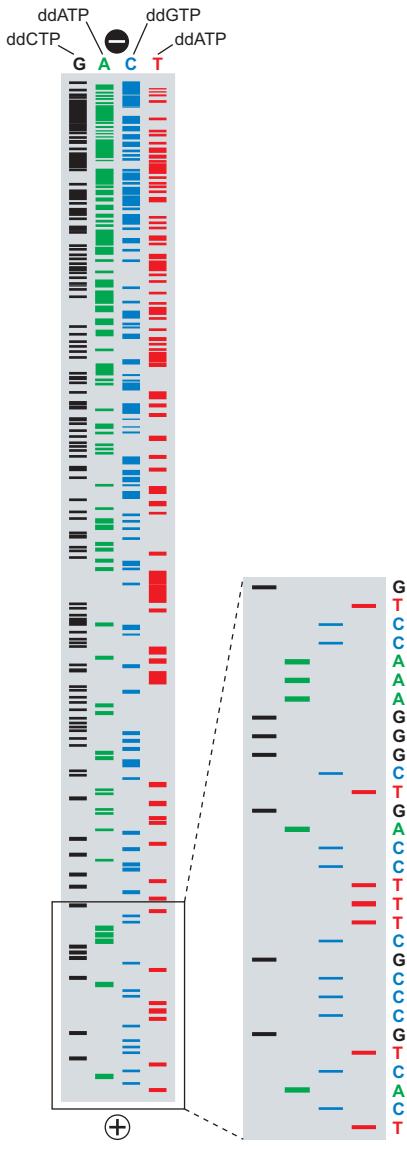


**FIGURE 7-14** Chain termination in the presence of dideoxynucleotides. The top illustration shows a DNA chain being extended at the 3' end with addition of an adenine nucleotide onto the previously incorporated cytosine. The presence of dideoxycytosine in the growing chain (shown at the bottom) blocks further addition of incoming nucleotides as described in the text.

same point, the chain-terminating nucleotides will generate a nested set of polynucleotide fragments, all sharing the same 5' end but differing in their lengths and hence their 3' ends. The length of the fragments therefore specifies the position of Cs in the template strand. The fragments can be labeled at their 5' ends either by the use of a radioactively labeled primer or a primer that has been tagged with a fluorescent adduct, or at their 3' ends with fluorescently labeled derivatives of ddGTP. Upon electrophoresis through a polyacrylamide gel, the nested set of fragments yields a ladder of fragments, each rung of the ladder representing a C on the template strand (Fig. 7-15b). If we similarly spike DNA synthesis reactions with ddCTP, ddATP, and ddTTP, then in toto we will generate four nested sets of fragments, which together provide the full nucleotide sequence of the DNA. To read that sequence, the fragments generated in each of the four reactions are resolved on a polyacrylamide gel (Fig. 7-16).



**FIGURE 7-15** DNA sequencing by the chain-termination method. As described in the text, chains of different length are synthesized in the presence of dideoxynucleotides. The length of the chains produced depend on the sequence of the DNA template and which dideoxynucleotide is included in the reaction. (a, top) The sequence of the template. In this reaction, all bases are present as deoxynucleotides, but G is present in the dideoxy form as well. Thus, when the elongating chain reaches a C in the template, it will, in some fraction of the molecules, add the ddGTP instead of dGTP. In those cases, chains terminate at that point. (b) Fragments separated on a polyacrylamide gel. The lengths of fragments seen on the gel reveal the positions of cytosines in the template DNA being sequenced in the reaction described.



**FIGURE 7-16** DNA-sequencing gel. The lengths of DNA chains, terminated with the dideoxynucleotide indicated at the top of each lane, are determined by resolving on a polyacrylamide gel, as shown. Reading the gel from bottom to top gives the 5'-to-3' sequence.

As we shall see later, this conceptually simple approach, developed initially to sequence short, defined DNA fragments, has undergone a series of technical adaptations and improvements that allow the analysis of whole genomes (see Box 7-2, Sequenators Are Used for High-Throughput Sequencing).

### Shotgun Sequencing a Bacterial Genome

The bacterium *Haemophilus influenzae* was the first free-living organism to have a complete genome sequence and assembly. It was a logical choice because it has a small, compact genome that is composed of just 1.8 million base pairs (Mb) of DNA (less than 1/1000th the size of the human genome). The *H. influenzae* genome was sheared into many random fragments with an average size of 1 kb. These pieces of genomic DNA were cloned into a plasmid DNA vector to create a library. DNA was prepared from individual recombinant DNA colonies and separately sequenced on Sequenators using the dideoxy method discussed above. This method is called “shotgun” sequencing. Random recombinant DNA colonies are picked, processed, and sequenced. To ensure that every single nucleotide in the genome was captured in the final genome assembly, 30,000–40,000 separate recombinant clones were sequenced. A total of ~20 Mb of raw genome sequence was produced ( $600\text{ bp} \times 33,000$  different colonies = 20 Mb of total DNA sequence). This is called **10 × sequence coverage**. In principle, every nucleotide in the genome should have been sequenced 10 times.

This method might seem tedious, but it is considerably faster and less expensive than the techniques that were originally envisioned. One early strategy called for systematically sequencing every defined restriction DNA fragment on the physical map of the bacterial chromosome. A drawback of this procedure is that most of the known restriction fragments are larger than the amount of DNA sequence information generated in a single reaction. Consequently, additional rounds of digestion, mapping, and sequencing would be required to obtain a complete sequence for any given defined region of the genome. These additional steps of cloning and restriction mapping are considerably more time-consuming than the repetitive automated sequencing of random DNA fragments. In other words, the computer is much faster at assembling random DNA sequences than the time required to clone and sequence a complete set of restriction fragments spanning a bacterial genome.

The approximately 30,000 sequencing reads derived from random genomic DNA fragments are directly entered into the computer, and programs are used to assemble overlapping DNA sequences. This process is conceptually similar to the assembly of a giant dense crossword puzzle in which the determined words give clues to the overlapping but unknown words. Random DNA fragments are “assembled” based on matching sequences. The sequential assembly of such short DNA sequences ultimately leads to a single continuous assembly, also called a contig (see Fig. 7-18).

### The Shotgun Strategy Permits a Partial Assembly of Large Genome Sequences

From our preceding discussion, it is obvious that sequencing short 600-bp DNA fragments is incredibly fast and efficient. In fact, the automated sequencing machines are so efficient that they far surpass our ability to assemble and annotate the raw DNA sequence information. In other

## ► KEY EXPERIMENTS

### Box 7-2 Sequenators Are Used for High-Throughput Sequencing

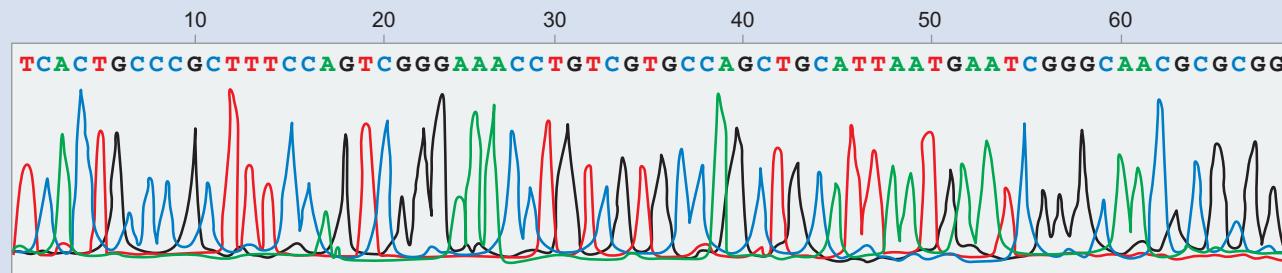
When the sequencing of the human genome was first envisioned, it seemed like a daunting, virtually hopeless enterprise. After all, the complete human genome consists of a staggering 3 billion ( $3 \times 10^9$ ) base pairs, and the early methods for determining the nucleotide sequence of even short DNA fragments were quite tedious. In the 1980s and early 1990s, an individual researcher could produce only a few hundred base pairs, perhaps 500 bp, of DNA sequence in a day or two of concentrated effort. Several technical innovations since then have greatly accelerated the speed and reliability of DNA sequencing.

As we described in the preceding section, the chain-termination method produces nested sets of DNAs that differ in size by just a single nucleotide. Initially, large polyacrylamide gels were used to fractionate these nested DNAs (see Fig. 7-16). However, in recent years, cumbersome gels have been replaced by short columns, which permit the resolution of nested DNAs in just 2–3 h. These short, reusable columns permit the fractionation of DNA fragments ranging from 700 bp to 800 bp, similar to the capacity of the far more cumbersome polyacrylamide gels that they have replaced.

A major technical advance in DNA sequencing came from the use of **fluorescent chain-terminating nucleotides**. In principle, it is possible to label each of the nested DNAs from a fragment with a single “color.” The color of each nested DNA depends on the identification of the last nucleotide. For

example, DNAs ending with a T residue at position 50 in the template DNA might be labeled red, whereas those nested DNAs ending with a G residue at position 51 might be labeled black. Thus, each nested DNA has a unique size and color. As they are fractionated on the sequencing columns based on size, fluorescent sensors detect the color of each nested DNA (Box 7-2 Fig. 1). In this way, a single column produces 600–800 bp of DNA sequence after less than 3 h of size separation.

Automated sequencing machines—**Sequenators**—were developed that have 384 separate fractionation columns. In principle, these machines can generate more than 200,000 nucleotides (200 kb) of raw DNA sequence in just a few hours. In a 9-h day, each machine can produce three sequencing “runs” and more than one-half a megabase (500 kb) of sequence information. A cluster of 100 such machines could generate the equivalent of one human genome,  $3 \times 10^9$  bp, in just 2 mo. There are currently five major sequencing centers in the United States and the United Kingdom. Each contains large clusters of automated DNA-sequencing machines. Together, these five centers produce a staggering  $60 \times 10^9$  bp of raw DNA sequence information per year. This corresponds to the equivalent of 20 human genomes per year! But as we shall see later, this is child’s play when compared with the next-generation Sequenators that routinely produce the equivalent of a complete human genome in a single run of just a few hours.



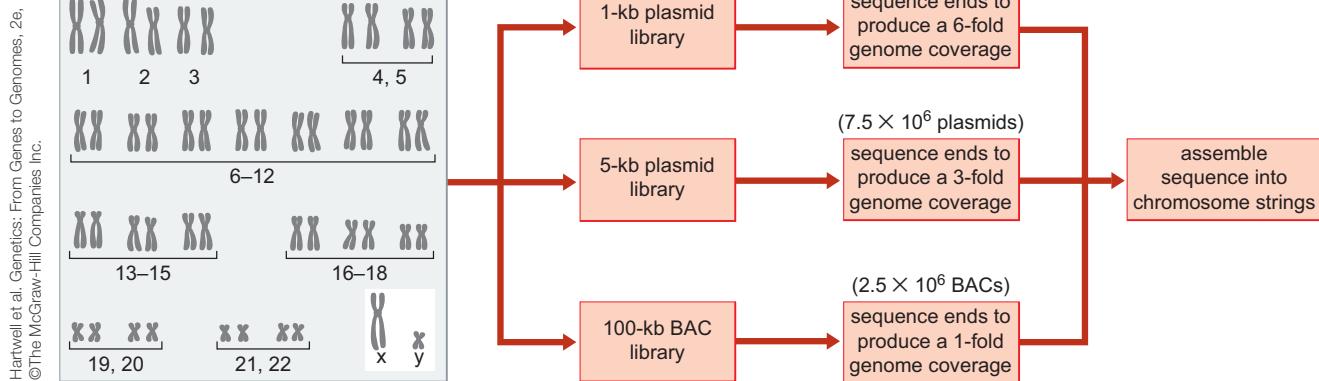
**BOX 7-2 FIGURE 1** DNA sequence readout. In this reaction, as described in the text, fluorescently end-labeled dideoxynucleotides are used, and the chains are separated by column chromatography. The profile of positions of As is represented in green, Ts in red, Gs in black, and Cs in blue.

words, the rate-limiting step in determining the complete DNA sequence of complex genomes, such as the human genome, is the analysis of the data, rather than the production of the data per se. This problem is rapidly becoming even more severe as the methods for sequencing are becoming increasingly faster and more powerful. It is now possible to generate several billion base pairs (gigabase pairs, Gb) of DNA sequence information in one “run” on an automated machine (see the section entitled The \$1000 Human Genome Is within Reach). We now consider how the shotgun-sequencing method used to determine the complete sequence of the *H. influenzae* genome was adapted for much larger and complicated animal genomes.

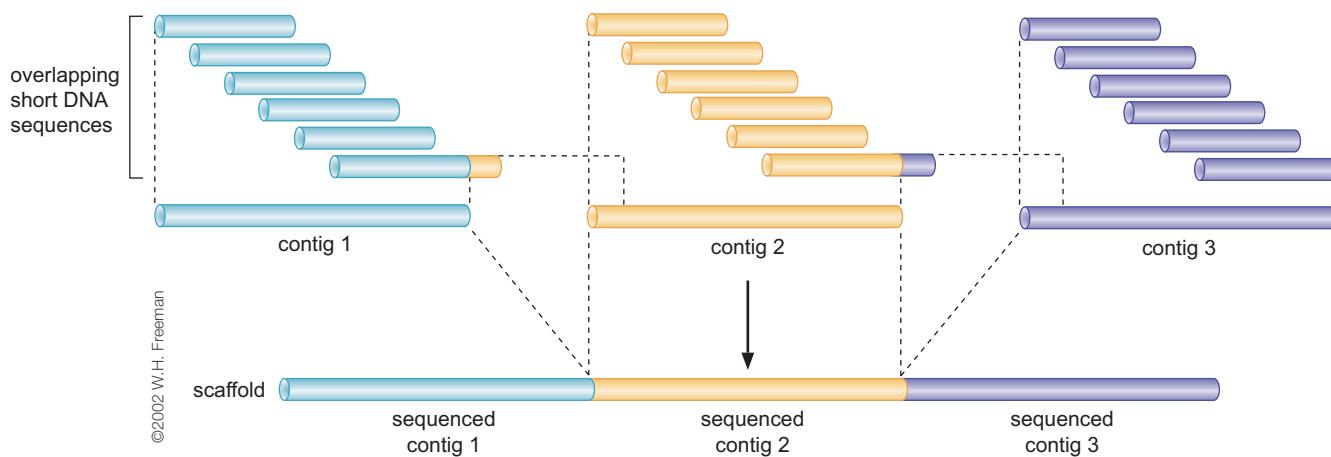
The average human chromosome is composed of 150 Mb. Thus, the 600 bp of DNA sequence provided by a typical sequencing reaction represents only 0.0004% of a typical chromosome. Consequently, to determine the complete sequence of the chromosome, it is necessary to generate a large number of sequencing reads from many short DNA fragments (Fig. 7-17). To achieve this goal, DNA is prepared from each of the 23 chromosomes that constitute the human genome and then sheared into small fragments by passage through small-gauge pressurized needles. The collection of small fragments, each derived from individual chromosomes, is then reduced into pools. Typically, two or three pools are constructed for fragments of differing (increasing) sizes—for example, fragments of 1, 5, or 100 kb in length. These fragments are then randomly cloned into bacterial plasmids as we described above to make libraries.

Recombinant DNA, containing a random portion of a human chromosome, can be rapidly isolated from bacterial plasmids and then quickly sequenced using automated sequencing machines. To ensure that every sequence is sampled in the complete chromosome, an average of 2 million random DNA fragments are processed. With an average of 600 bp of DNA sequence per fragment, this procedure produces more than 1 billion base pairs (1 Gb) of sequence data, or nearly 10 times the amount of DNA in a typical chromosome. As discussed above for the sequencing of the bacterial chromosome, by sampling about 10 times the amount of sequence in a chromosome, we can be confident that every portion of the chromosome will be captured.

The process of producing “shotgun” recombinant libraries and huge excesses of random DNA-sequencing reads seems very wasteful. However, a cluster of 100 384-column automated sequencing machines can generate 10-fold coverage of a human chromosome in just a few weeks. This approach is considerably faster than the methods involving the isolation of known regions within the chromosome and sequentially sequencing a known set of staggered DNA fragments. Thus, the key technological insight that facilitated the sequencing of the human genome was the reliance on automated **shotgun sequencing** and the subsequent use of computers to assemble the different pieces. The combination of automated sequencing machines and



**FIGURE 7-17** Strategy for construction and sequencing of whole-genome libraries. Contiguous sequences are determined for the shotgun sequencing of the short genomic DNA fragments. Contigs are extended by the use of end sequences derived from the larger fragments carried in the 5-kb and 100-kb insert clones as described in the text. (Adapted, with permission, from Hartwell L. et al. 2003. *Genetics: From genes to genomes*, 2nd ed., Fig. 10-13. © McGraw-Hill.)



**FIGURE 7-18** Contigs are linked by sequencing the ends of large DNA fragments. For example, one end of a random 100-kb genomic DNA fragment might contain sequence matches within contig 1, whereas the other end matches sequences in contig 2. This places the two contigs on a common scaffold. (Adapted, with permission, from Griffiths A.J.F. et al. 2002. *Modern genetics*, 2nd ed., Fig. 9-29b. © W.H. Freeman.)

computers proved to be a potent one–two punch that led to the completion of the human genome sequence years earlier than originally planned.

Sophisticated computer programs have been developed that assemble the short sequences from random shotgun DNAs into larger contiguous sequences called **contigs**. Sequences or “reads” that contain identical sequences are assumed to overlap and are joined to form larger contigs (Fig. 7-18). The sizes of these contigs depend on the amount of sequence obtained—the more sequence, the larger the contigs and the fewer gaps in the sequence.

Individual contigs are typically composed of 50,000–200,000 bp. This is still far short of a typical human chromosome. However, such contigs are useful for analyzing compact genomes. For example, the *Drosophila* genome contains an average of one gene every 10 kb, thus a typical contig has several linked genes. Unfortunately, more complex genomes often contain considerably lower gene densities (see Chapter 8). Because the human genome contains an average of one gene every 100 kb, a typical contig is often insufficient to capture an entire gene, let alone a series of linked genes. We now consider how relatively short contigs are assembled into larger **scaffolds** that are typically 1–2 Mb in length.

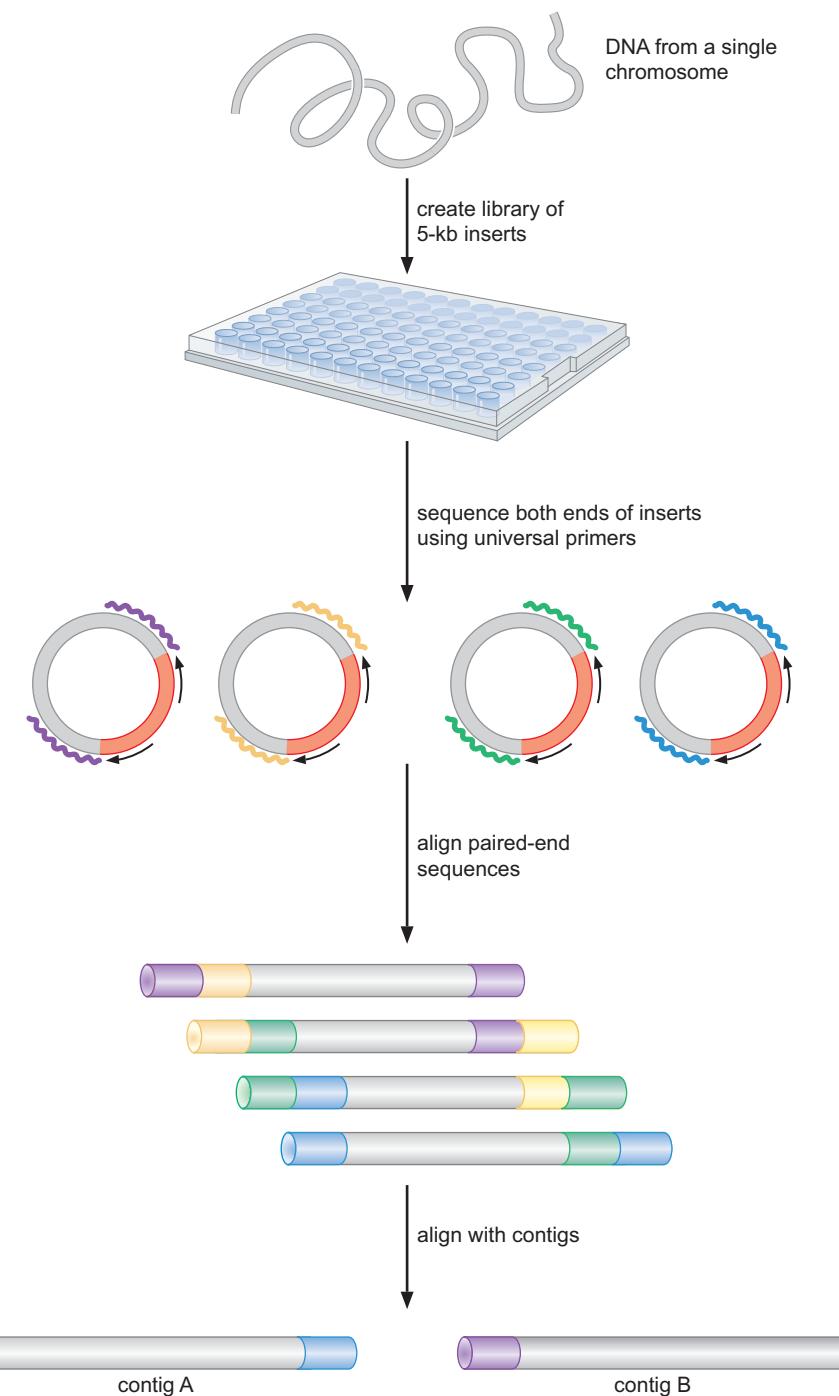
### The Paired-End Strategy Permits the Assembly of Large-Genome Scaffolds

A major limitation to producing larger contigs is the occurrence of repetitive DNAs (see Chapter 8). Such sequences complicate the assembly process because random DNA fragments from unlinked regions of a chromosome or genome might appear to overlap because of the presence of the same repetitive DNA sequence. One method that is used to overcome this difficulty is called **paired-end sequencing**. This is a simple technique that has produced powerful results (see Fig. 7-19).

In addition to producing shotgun DNA libraries composed of short DNA fragments, the same genomic DNA is also used to produce recombinant libraries composed of larger fragments, typically between 3 and 100 kb in length. Consider a DNA sample from a single human chromosome. Some of the DNA is used to produce 1-kb fragments, whereas another aliquot of

the same sample is used to produce 5-kb fragments. The end result is the construction of two libraries, one with small inserts and a second with larger inserts (see Fig. 7-17).

Universal primers are made that anneal at the junction between the plasmid and both sides of the large inserted DNA fragment. Individual runs will produce ~600 bp of sequence information at each end of the random insert. A record is kept of what end sequences are derived from the same inserted fragment. One end might align with sequences contained within contig A, whereas the other end aligns with a different contig, contig B. Contigs A and B are now assumed to derive from the same region of the chromosome



**FIGURE 7-19** A “shotgun” library containing random genomic DNA inserts of 5 kb in length. Each well on the plate contains a different insert. Sequences 600 bp in length are determined for both ends of each genomic DNA (color coded). These paired-end sequences are used to align different contigs. In this example, the 5-kb genomic DNA fragment with the blue sequences contains matching sequences with contig A and contig B.

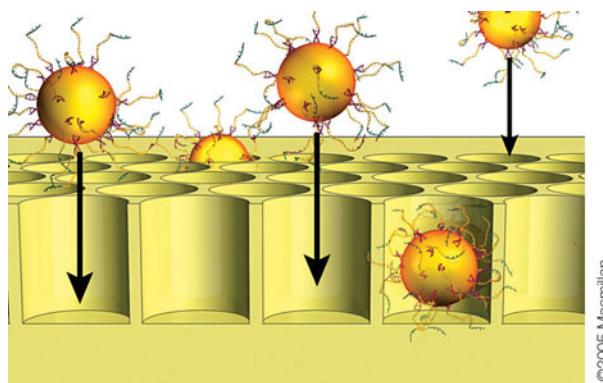
because they share sequences with a common 5-kb fragment. Because most repetitive DNA sequences are less than 2 or 3 kb in length, the “paired-end” sequences from the 5-kb insert are sufficient to span contigs interrupted by repetitive DNAs.

The preceding results usually produce contigs that are <500 kb in length. To obtain long-range sequence data, on the order of several megabases or more, it is necessary to obtain paired-end sequence data from large DNA fragments that are at least 100 kb in length. These can be obtained using a special cloning vector called a **BAC (bacterial artificial chromosome)** that can accommodate very large inserts, up to hundreds of kilobases of DNA. The principle of how these are used to produce long-range sequence information is the same as that described for the 5-kb inserts. Primers are used to obtain 600-bp sequencing reads from both ends of the BAC insert. These sequences are then aligned to different contigs, which can then be assigned to the same scaffold by virtue of sharing sequences from a common BAC insert. The use of BACs often permits the assignment of multiple contigs into a single scaffold of several megabases (see Fig. 7-18).

### The \$1000 Human Genome Is within Reach

The sequencing of the first two human genomes (one from the National Institutes of Health and the other from a private company) cost more than \$300 million. There is now a campaign to use nanotechnology to produce rapid and inexpensive genome sequencing. The goal is to make the technology sufficiently rapid, simple, and inexpensive to permit the sequencing of individual genomes for clinical diagnosis. The first generation of high-throughput, nanotechnology sequencing machines is now available.

The 454 Life Sciences sequencing machine generates up to 400 Mb of sequence information in a 4-h “run.” The basic principle is very clever. Small fragments of DNA (genomic, cDNA, etc.) are mixed with small beads. The mixture is sufficiently dilute so that a single DNA molecule binds to a single bead. Next, the DNA-containing beads are dispersed on a silicon plate consisting of 400,000 regularly spaced picoliter-sized wells. The small size of the wells ensures that each one captures no more than a single bead. PCR is performed directly on the bead-tethered DNAs to amplify each DNA molecule (Fig. 7-20). Thus, a homogeneous population of DNA molecules is created in each well, which is then used as a template for an additional round of DNA synthesis. Sequencing is performed in stepwise fashion with the plate being separately exposed to dATP, dGTP, dCTP, and dTTP sequentially, with a washing cycle between each pulse of



©2005 Macmillan

**FIGURE 7-20** Cartoon of individual pores in the 454 sequencing apparatus. Each pore contains a small bead with an amplified DNA sequence. Sequential rounds of sequencing are detected by the release of pyrophosphate and light. Further description of the method is given in the text. (Reprinted, with permission, from Margulies M. et al. 2005. *Nature* **437**: 376–380, Fig. 1a. © Macmillan.)

deoxynucleotide substrate. The incorporation of a deoxynucleotide depends on the presence of the complementary base in the template and results in the liberation of pyrophosphate. This release promotes an enzymatic reaction that produces pulses of light, which are detected by a microprocessor attached to a computer. The light pulses indicate which nucleotide is incorporated in each well during each round of synthesis, thereby producing the sequence of the DNA contained in all 400,000 wells. Sequential addition of each nucleotide is continued until ~200–250 bases of sequence have been determined from each DNA fragment.

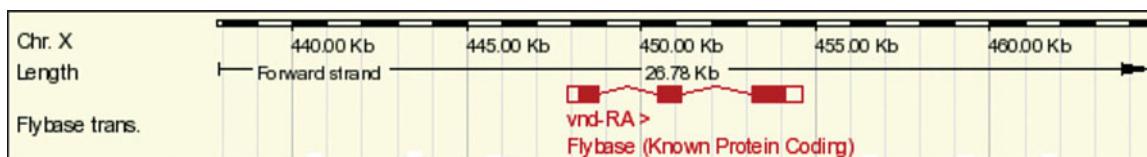
454 sequencing has produced the complete genome of the lead author of this textbook (for some reason, the company seems less interested in the genomes of the other authors). At 100 Mb of genome sequence per “run,” complete  $1 \times$  coverage of Watson’s genome required just 30 runs (2–3 wk on one machine). If started now (at the time of this writing), the total cost would be about \$10,000–\$30,000, a small fraction of the cost of the first human genome sequence. The sequence information is not necessarily sufficient to produce a *de novo* genome assembly. Rather, the finished human genome sequence produced by the National Institutes of Health is used as a template for comparison. Each of the 200–250-bp sequence reads produced by 454 sequencing are identified on the finished genome until Watson’s variants of every gene are identified. Thus, the meaning of sequencing a human genome has shifted. Because we have a finished whole-genome sequence assembly in hand, new genomes require only short sequencing reads to obtain a comprehensive atlas of an individual’s unique genetic composition.

The next generation of sequencing machines is approaching the goal of the \$1000 genome. Illumina has produced a machine that can generate hundreds of millions of sequencing reads of ~200 bp per run. The basic principle is similar to that seen for the 454 Life Sciences sequencing machine. The difference is that individual DNA molecules are attached to a glass slide. Limited PCR amplification is performed to produce approximately 1000 copies per DNA molecule. Sequential DNA synthesis reactions are performed and detected by the release of pyrophosphate. The Illumina Sequencers routinely produce several gigabase pairs of DNA sequence information in a single run. A variety of “next-gen” high-throughput sequencing methods are being developed, including ion semiconductor sequencing, which detects the hydrogen ion released upon incorporation of a nucleotide during DNA synthesis.

## GENOMICS

---

Before the advent of whole-genome sequencing, investigators were severely limited in the scope of DNA sequence comparisons. At best, they could look at the DNA sequences of just a few individual genes among a small set of organisms. With the advent of powerful, automated DNA sequencing machines, it is now possible to obtain complete information regarding the organization and genetic composition of entire genomes. In fact, as of this writing, nearly 200 different genomes have been sequenced and assembled. It is therefore possible to compare the complete genetic composition of many different microbes, plants, and animals. In this section, we consider the basic methods that are used for the annotation of genomes—that is, the use of both experimental and computational methods for the identification of every gene (including intron–exon structure; see Chapter 14) and associated regulatory sequences within a complex genome.



**FIGURE 7-21** Structure of the *vnd* locus in *Drosophila*. An ~25-kb interval on the X chromosome that contains the *vnd* gene. The *vnd* transcription unit contains three exons and two introns. The unfilled portions of the 5' (left) and 3' (right) exons indicate non-coding sequences that do not contribute to the final protein product. FlyBase is the standardized database that is used to analyze the *Drosophila* genome.

### Bioinformatics Tools Facilitate the Genome-Wide Identification of Protein-Coding Genes

Genome sequence assemblies correspond to contiguous blocks of millions of sequential As, Gs, Cs, and Ts encompassing every chromosome of the organism in question. They are large, tedious, and uninformative unless “annotated.” As described in the next few pages, **annotation** is the systematic identification of every stretch of genomic DNA that contains protein-coding information or non-coding sequences that specify regulatory RNAs such as microRNAs (miRNAs; see Chapter 20). The detailed intron–exon structure of every transcription unit is identified, and in cases in which the genome in question corresponds to a model organism (e.g., yeast and fruit flies), it is possible to assign potential or known functions to most of the genes in the genome. Only when this information is available is it possible to catalog the complete coding capacity of the genome and compare its contents with those of other genomes.

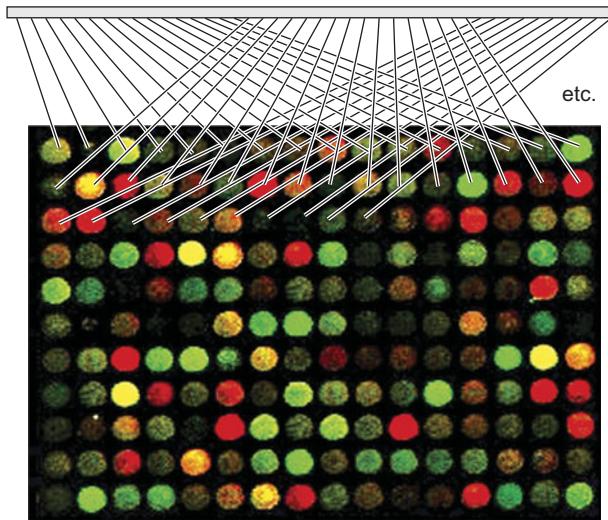
For the genomes of bacteria and simple eukaryotes, genome annotation is relatively straightforward, amounting essentially to the identification of open reading frames (ORFs). Although not all ORFs—especially small ones—are real protein-coding genes, this process is fairly effective, and the key challenge is in correctly assigning the functions of these genes.

For animal genomes with complex intron–exon structures, the challenge is far greater. In this case, a variety of bioinformatics tools are required to identify genes and determine the genetic composition of complex genomes. Computer programs have been developed that identify potential protein-coding genes through a variety of sequence criteria (Fig. 7-21), including the occurrence of extended ORFs that are flanked by appropriate 5' and 3' splice sites. As discussed in Chapter 14, splice donor and acceptor sites are short and somewhat degenerate sequences, but they nevertheless help identify exon–intron boundaries when considered in the context of additional information, such as expressed sequence tag (EST) sequence data, which we shall consider later. Nonetheless, computational methods have not yet been refined to the point of complete accuracy. Something like three-fourths of all genes can be identified in this way, but many are missed, and even among the predicted genes that are identified, small exons—particularly non-coding exons—are often overlooked.

### Whole-Genome Tiling Arrays Are Used to Visualize the Transcriptome

Once a whole-genome sequence is assembled for an organism, it can be used to comprehensively reveal all protein-coding and non-coding (e.g., introns and miRNA genes) sequences that are expressed in specific cells or tissues.

**FIGURE 7-22 Whole-genome tiling microarray.** The image represents a portion of a tiling array that has been hybridized with fluorescently labeled probes. The grid includes a high number of uniformly spaced DNA probes across a region of interest (e.g., an entire genome).



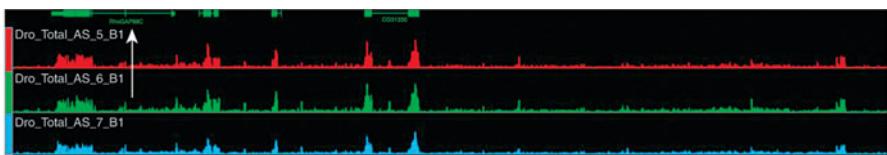
The portion of an organism's genome that acts as a template for RNA synthesis is known as the **transcriptome**. To identify this portion of the genome, synthetic, single-stranded DNAs of 50 nucleotides in length are spotted on a glass or silicon slide. Typically, one oligonucleotide is produced for every 100–150 bp of DNA sequence in a sequential manner across the genome, resulting in a “tiling array” of DNA sequences. The technology for genome-wide tiling is advancing rapidly, and it is now feasible to produce complete arrays on a single glass slide or silicon chip that is just 1 cm<sup>2</sup> in size. For example, 1 million 50-mers encompass the entire *Drosophila* genome, and all of these oligonucleotides can be spotted on a single DNA chip. Each spot on the chip (i.e., each oligonucleotide sequence) is so small that hybridization signals are detected by microsensors attached to a microscope, as we shall describe later.

To visualize the transcriptome, the tiling arrays are hybridized with fluorescently labeled RNA (or cDNA) probes (see Fig. 7-22). These probes might be derived from a specific cell type, such as the tail muscles of the sea squirt tadpole or yeast cells grown in a particular medium. The end result is a series of hybridization signals superimposed on all of the predicted protein-coding sequences across the genome (Fig. 7-23). An alternative strategy for transcriptome profiling is the high-throughput sequencing of cDNAs prepared from cultured cells or isolated tissues.

Whole-genome tiling arrays provide immediate information regarding the intron–exon structure of individual transcription units (Fig. 7-23). This is due to the unstable nature of intronic transcripts. Although total RNA is typically used for these experiments, the exonic sequences are more stable than the introns, which decay rapidly after their removal from primary transcripts (see Chapter 14). After labeling and hybridization to the tiling array chip, exonic sequences display more intense signals than introns.

Another useful feature of whole-genome tiling arrays is that they detect non-coding genes, such as those specifying miRNAs. These RNAs are usually processed from larger precursor RNAs (pri-RNAs) derived from

**FIGURE 7-23 Whole-genome tiling array reveals details of the intron–exon structure of a gene.** A 50-kb interval on *Drosophila* chromosome 3 that contains four different genes. The intron–exon structure of each transcription unit is shown at the top of the figure. (White arrow) The large intronic region that might contain a small (“micro-”) exon. Total RNA was extracted from progressively older embryos (red, young; green, older; and blue, still older) and hybridized to the tiling array, which contains 25-nucleotide sequences every 35 bp throughout the entire genome. Strong hybridization signals coincide with the exons, whereas there are weaker signals in the intronic regions. Based on the similar signals in all three colors this gene is expressed at similar levels at all three ages of embryos tested. (Reprinted, with permission, from Manak et al. 2006. *Nat. Genet.* 38: 1151–1158, Fig. 5. © Macmillan.)



transcription units that are 1–10 kb in length (see Chapter 20). The pri-RNA transcription units are easily detected by hybridization to tiling arrays. In some cases, miRNA genes contain introns that must be processed before the final production of the mature miRNA. Other types of non-coding transcripts are also detected, including “antisense” RNAs within the introns of protein-coding genes. It is possible that such RNAs function in a regulatory capacity to control the expression or function of protein-coding genes.

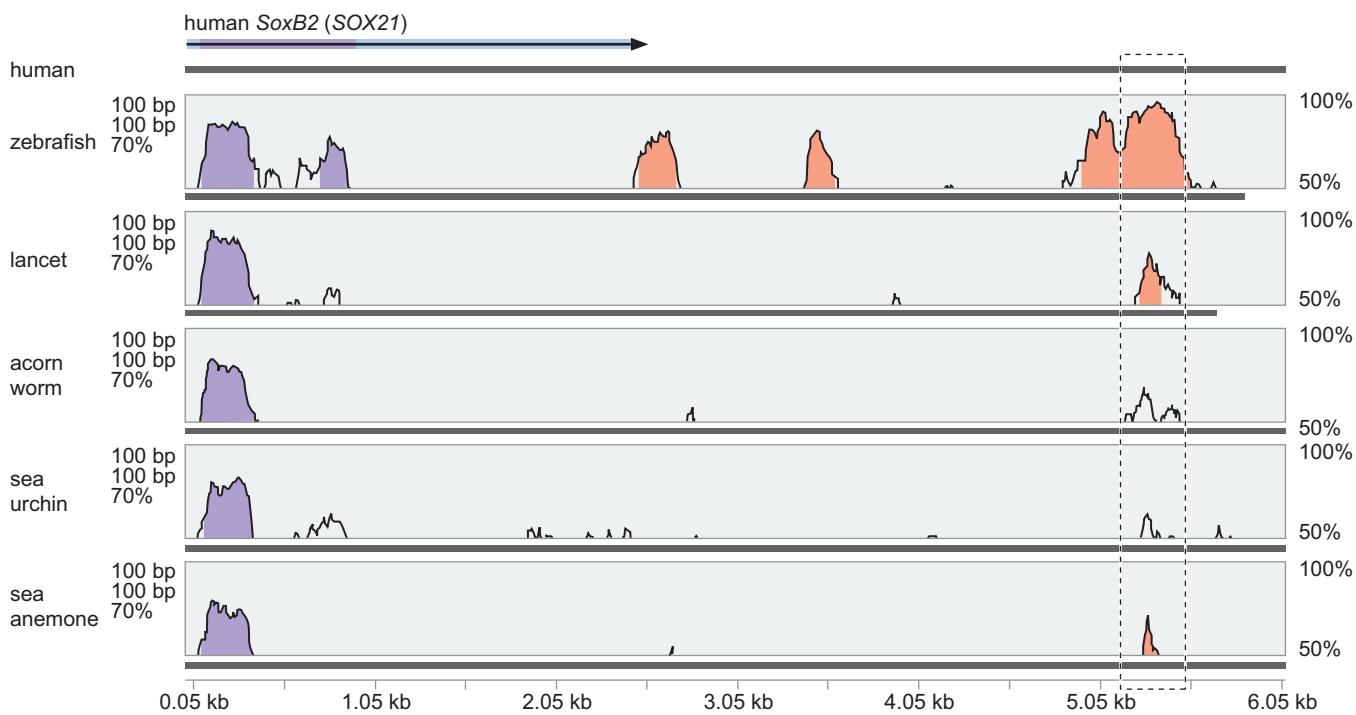
Tiling arrays have led to a rather startling observation: about one-third of a typical genome is transcribed, even though just a fraction of this transcription corresponds to protein-coding sequences (just 5% in the case of the human genome). It appears that most of the additional transcription is due to vast tracts of intronic DNA sequences. Many genes have remote, 5'-non-coding exons that reside far (sometimes a megabase or more) from the main body of the coding sequence. In some cases, these intronic regions produce miRNAs and additional types of non-coding RNAs. Extended 3'-UTRs represent another source of non-coding transcription.

### Regulatory DNA Sequences Can Be Identified by Using Specialized Alignment Tools

Genome technologies are effective at identifying genes and determining the structures of their transcription units. Once identified, a host of bioinformatics methods permits the determination of potential protein structure and function, for example, whether the protein contains any known domains or motifs or shares other features with known proteins. In particular, the Basic Local Alignment Search Tool, or BLAST, algorithm provides a powerful approach for searching, comparing, and aligning either protein or nucleic acid sequences. BLAST searches permit the rapid comparison of a given exon sequence with a vast database of protein-coding information. Significant sequence alignments with protein-coding sequences of known function (e.g., DNA-binding protein, replication factor, or membrane receptor) provide immediate insights into the potential activities of the gene and its putative protein products. Simple BLAST searches can also reveal the identities of non-coding transcripts that produce miRNAs (see Chapter 20).

In contrast to protein-coding sequences, the identification and characterization of regulatory sequences—those stretches of DNA controlling where and when the associated genes are ON and OFF in an organism—are extremely challenging, as we shall see in Chapter 19. In fact, some refer to the regulatory sequences as the “dark matter” of the genome. Genome-wide methods are only now becoming available for the identification of this important class of DNA sequence information.

A subset of vertebrate regulatory sequences can be identified using variations in the BLAST searches developed for characterizing protein-coding sequences. Cell-specific enhancers contain clustered binding sites for one or more sequence-specific DNA-binding proteins (see Chapter 19). In some cases, this clustering is sufficient for the identification of short stretches of DNA sequence alignment. A computer program called VISTA aligns the sequences contained in genomes of different related organisms over short windows, on the order of 10–20 bp, and thereby identifies imperfectly conserved non-coding sequences over stretches of just 50–75 bp (Fig. 7-24). Pufferfish and mice share approximately 10,000 short non-coding sequences. It is conceivable that many of these correspond to tissue-specific enhancers. However, it is likely that both animals, particularly mice, have at least 100,000 enhancers. Thus, these simple sequence alignments fail to capture the vast majority of regulatory sequences.

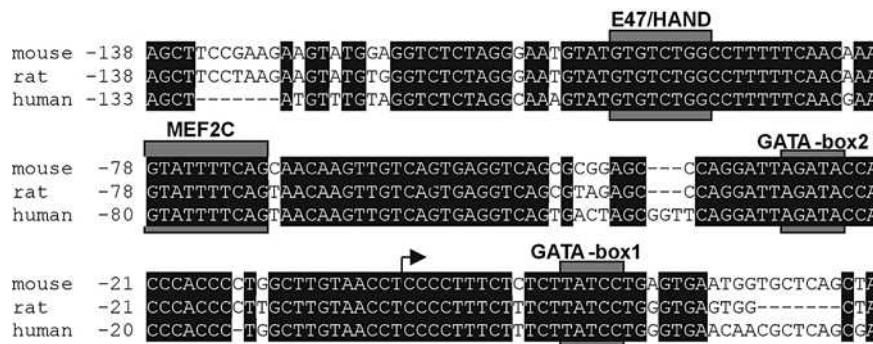


**FIGURE 7-24** Comparison of the *SoxB2* gene in divergent animals. The lavender signals correspond to conserved sequences in the 3'-UTR of the *SoxB2* transcription unit. The pink signals indicate conserved sequences that map downstream of the gene. The dashed rectangle identifies enhancers that mediate expression in the nervous system. (Adapted, with permission, from Royo J.L. et al. 2011. *Proc. Natl. Acad. Sci.* **108**: 14186–14191, Fig. 1A, p. 14187.)

Tissue-specific enhancers can also be identified by scanning genomic DNA sequences for potential binding sites of known regulatory proteins. Consider the case of the  $\alpha$ -catenin gene, which encodes a cell adhesion molecule. The gene is expressed in several different tissues, but it shows particularly strong expression in heart precursor cells called cardiomyocytes. It was possible to identify a heart-specific enhancer by surveying the flanking and intronic sequences of  $\alpha$ -catenin for matches to the binding sites of known heart cell regulatory proteins, including MEF2C, GATA-4, and E47/HAND (Fig. 7-25). Each of these proteins recognizes a spectrum of short sequence motifs of 6–10 bp. The spectrum of binding sites for each factor is described by a position-weighted matrix (PWM), which can be determined using a variety of computational and experimental methods such as SELEX (in vitro selection) assays (which we shall discuss in detail later in this chapter). When these PWMs were used to survey the  $\alpha$ -catenin locus, a single cluster of putative MEF2C-, GATA-4-, and E47/HAND-binding sites was identified. Experimental studies confirmed that this cluster of binding sites, located in the 5'-flanking region of the gene, function as a bona fide enhancer.

### Genome Editing Is Used to Precisely Alter Complex Genomes

The preceding methods, genome assemblies and annotation, are descriptive. They provide detailed atlases of whole-genome maps but do not provide the type of functional information that molecular biologists crave. However, a recently developed method, genome editing, permits the removal or modification of specific DNA segments within an otherwise



**FIGURE 7-25** *In silico* identification of a heart enhancer. An ~140-bp sequence in the 5'-flanking region of the  $\alpha$ -catenin gene is conserved in the mouse, rat, and human genomes. The conserved sequence contains binding sites for three critical regulators of heart differentiation: E47/HAND, MEF2C, and GATA. The mouse sequence has been shown to function as an authentic heart-specific enhancer. In principle, it could be identified by either VISTA alignments (see Chapter 20, Fig. 20-4) or the clustering of heart regulatory proteins. (Portion reprinted, with permission, from Vanpoucke G. et al. 2004. *Nucleic Acids Res.* **32**: 4155–4165, Fig. 1. © Oxford University Press.)

intact genome. This approach involves inducing a double-strand break (DSB) in a specific target DNA sequence that stimulates homologous recombination to repair the break using introduced modified DNA. During the break repair event, desired changes are introduced specifically to modify the genomic sequence. The targeted cleavage is performed by specially tailored nucleases, typically zinc-finger nucleases and “meganucleases,” engineered to cleave at a designated target site in the genome. However, a new class of “designer” nucleases—the transcriptional activator-like effector nucleases (TALENs)—has been shown to have increased efficiency. TALENs are emerging as an important tool for targeted genome editing in different model organisms as well as human stem cells.

## PROTEINS

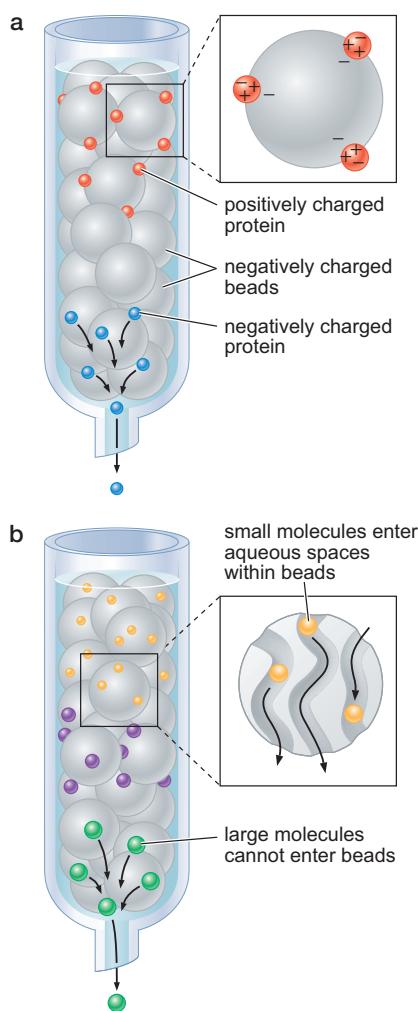
### Specific Proteins Can Be Purified from Cell Extracts

The purification of individual proteins is critical to understanding their function. Although in some instances, the function of a protein can be studied in a complex mixture, these studies can often lead to ambiguities. For example, if you are studying the activity of one specific DNA polymerase in a crude mixture of proteins (such as a cell lysate), other DNA polymerases and accessory proteins may be partly or completely responsible for any DNA synthesis activity that you observe. For this reason, the purification of proteins is a major part of understanding their function.

Each protein has unique properties that make its purification somewhat different. This is in contrast to different DNAs, which all share the same helical structure and are only distinguished by their precise sequence. The purification of a protein is designed to exploit its unique characteristics, including size, charge, shape, and, in many instances, function.

### Purification of a Protein Requires a Specific Assay

To purify a protein requires an assay that is unique to that protein. For the purification of a DNA, the same assay is almost always used, hybridization to its complement. As we shall see in the discussion of immunoblotting, an antibody can be used to detect specific proteins in the same way. In many instances, it is more convenient to use a more direct measure for the function of the protein. For example, a specific DNA-binding protein can be assayed by determining its interaction with the appropriate DNA (e.g., using an electrophoretic mobility shift assay, described in the section Nucleic Acid–Protein Interactions). Similarly, a DNA or RNA polymerase can be detected by



**FIGURE 7-26** Ion-exchange and gel-filtration chromatography. As described in the text, these two commonly used forms of chromatography separate proteins on the basis of their charge and size, respectively. Thus, in each case, a glass tube is packed with beads, and the protein mixture is passed through this matrix. The nature of the beads dictates the basis of protein separation. (a) Ion-exchange chromatography. In this example, the beads are negatively charged. Thus, positively charged proteins bind to them and are retained on the column, whereas negatively charged proteins pass through. Increasing the concentration of salt in the surrounding buffer can elute bound proteins by competing for the negative charges on the column. (b) Gel-filtration chromatography. The beads contain aqueous spaces into which small proteins can pass, slowing down their progress through the column. Larger proteins cannot enter the beads, allowing them to pass more rapidly through the column.

incorporation assays by adding the appropriate template and radioactive nucleotide precursor to a crude extract in a manner similar to the methods used to label DNA described above. As discussed in Chapter 9, Box 9-1, incorporation assays are useful for monitoring the purification and function of many different enzymes catalyzing the synthesis of polymers such as DNA, RNA, or proteins.

### Preparation of Cell Extracts Containing Active Proteins

The starting material for almost all protein purifications are extracts derived from cells. Unlike DNA, which is very resilient to temperature, even moderate temperatures readily denature proteins once they are released from a cell. For this reason, most extract preparation and protein purification is performed at 4°C. Cell extracts are prepared in several different ways. Cells can be lysed by detergent, shearing forces, treatment with low ionic salt (which causes cells to absorb water osmotically and “pop” easily), or rapid changes in pressure. In each case, the goal is to weaken and break the membrane surrounding the cell to allow proteins to escape. In some instances, this process of treating the membrane is performed at very low temperatures by freezing the cells before applying shearing forces (often using a coffee grinder or blender similar to the one in many kitchens).

### Proteins Can Be Separated from One Another Using Column Chromatography

The most common method for protein purification is **column chromatography**. In this approach to protein purification, protein fractions are passed through glass or plastic columns filled with appropriately modified small acrylamide or agarose beads. There are various ways columns can be used to separate proteins. Each separation technique exploits different properties of the proteins. Three basic approaches are described here. The first two, in this section, separate proteins on the basis of their charge or size, respectively. These methods are summarized in Figure 7-26.

***Ion-Exchange Chromatography*** In this technique, the proteins are separated by their surface ionic charge using beads that are modified with either positively charged or negatively charged chemical groups. Proteins that interact weakly with the beads (such as a weak, positively charged protein passed over beads modified with a negatively charged group) are released from the beads (or eluted) in a low-salt buffer. Proteins that interact more strongly require more salt to be eluted. In either case, the salt masks the charged regions, allowing the protein to be released from the beads. Because each protein has a different charge on its surface, they will each be eluted from the column at a characteristic salt concentration. By gradually increasing the concentration of salt in the eluting buffer, even proteins with similar charge characteristics can be separated into different fractions as they elute from the column.

***Gel-Filtration Chromatography*** This technique separates proteins on the basis of size and shape. The beads used for this type of chromatography do not have charged chemical groups attached. Instead, each bead has a variety of different-sized pores penetrating its surface (similar to the pores that DNA passes through in agarose or acrylamide gels). Small proteins can enter all of the pores and therefore can access more of the column and take longer to elute (in other words, they have more space to explore). Large proteins can access less of the column and elute more rapidly.

For each type of column, chromatography fractions are collected at different salt concentrations or elution times and assayed for the protein of interest. The fractions with the most activity are pooled and subjected to additional purification.

By passing proteins through several different columns, a protein is increasingly purified. Although it is rare that an individual column will purify a protein to homogeneity by repeatedly separating fractions that contain the protein of interest (as determined by the assay for the protein), a series of chromatographic steps can result in a fraction that contains many molecules of a specific protein and few molecules of any other protein. For example, although there are many proteins that elute in high salt from a positively charged column (indicating a high negative charge) or slowly from a gel-filtration column (indicating a relatively small size), there will be far fewer that satisfy both of these criteria.

**Affinity Chromatography** This method can facilitate more rapid protein purification. Specific knowledge of a protein can frequently be exploited to purify that protein more rapidly. For example, if a protein binds ATP during its function, the protein can be applied to a column of beads that are coupled to ATP. Only proteins that bind to ATP will bind to the column, allowing the large majority of proteins that do not bind ATP to pass through the column. The ATP-binding proteins can be further separated by sequentially adding solutions with increasing concentrations of ATP, which will elute proteins according to their affinity for ATP (the more ATP required to elute, the higher the affinity). This approach to purification is called **affinity chromatography**. Other reagents can be attached to columns to allow the rapid purification of proteins; these include specific DNA sequences (to purify DNA-binding proteins) or even specific proteins that are suspected to interact with the protein to be purified. Thus, before beginning a purification, it is important to think about what information is known regarding the target protein and to try to exploit this knowledge.

One very common form of protein affinity chromatography is **immunoaffinity chromatography**. In this approach, an antibody that is specific for the target protein is attached to beads. Ideally, this antibody will interact only with the intended target protein and allow all other proteins to pass through the beads. The bound protein can then be eluted from the column using salt, a pH gradient, or, in some cases, mild detergent. The primary difficulty with this approach is that frequently the antibody binds the target protein so tightly that the protein must be denatured before it can be eluted. Because protein denaturation is often irreversible, the target protein obtained in this manner may be inactive and therefore less useful.

Proteins can be modified to facilitate their purification. This modification usually involves adding short additional amino acid sequences to the beginning (amino terminus) or the end (carboxyl terminus) of a target protein. These additions, or “tags,” can be generated using the molecular cloning methods described above. The peptide tags add known properties to the modified proteins that assist in their purification. For example, adding six histidine residues in a row to the beginning or end of a protein will make the modified protein bind tightly to a column with  $\text{Ni}^{2+}$  ions attached to beads—a property that is uncommon among proteins in general. In addition, specific peptide **epitopes** (a sequence of 7–10 amino acids recognized by an antibody) have been defined that can be attached to any protein. This procedure allows the modified protein to be purified using immunoaffinity purification and a heterologous antibody that is specific for the added epitope. Importantly, such antibodies and epitopes can be chosen such that they

bind with high affinity under one condition (e.g., in the presence of  $\text{Ca}^{2+}$ ) but readily elute under a second condition (e.g., in the absence of  $\text{Ca}^{2+}$ ). This avoids the need to use denaturing conditions for elution.

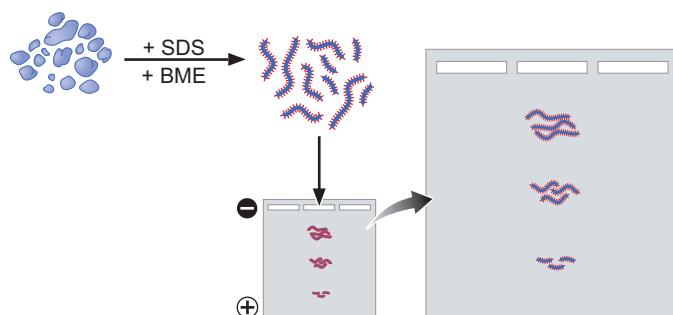
Immunoaffinity chromatography can also be used to rapidly precipitate a specific protein (and any proteins tightly associated with it) from a crude extract. In this case, precipitation is achieved by attaching the antibody to the same type of bead used in column chromatography. Because these beads are relatively large, they rapidly sink to the bottom of a test tube along with the antibody and any proteins bound to the antibody. This process, called **immunoprecipitation**, is used to rapidly purify proteins or protein complexes from crude extracts. Although the protein is rarely completely pure at this point, this is often a useful method to determine what proteins or other molecules (e.g., DNA; see the section on Chromatin Immunoprecipitation later in this chapter) are associated with the target protein.

### Separation of Proteins on Polyacrylamide Gels

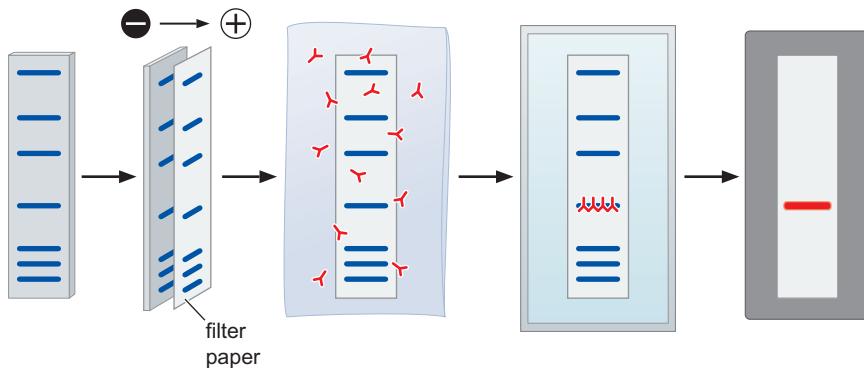
Proteins have neither a uniform negative charge nor a uniform structure. Rather, they are constructed from 20 distinct amino acids, some of which are uncharged, some are positively charged, and still others are negatively charged (Chapter 6, Fig. 6-2). In addition, as we discussed in Chapter 6, proteins have extensive secondary and tertiary structures and are often in multimeric complexes (quaternary structure). If, however, a protein is treated with the strong ionic detergent **sodium dodecyl sulfate (SDS)** and a reducing agent, such as mercaptoethanol, the secondary, tertiary, and quaternary structure is usually eliminated. Once coated with SDS, the protein behaves as an unstructured polymer. SDS ions coat the polypeptide chain, giving it a uniform negative charge. Mercaptoethanol reduces disulfide bonds, disrupting intramolecular and intermolecular disulfide bridges formed between cysteine residues. Under these conditions, as is the case with mixtures of DNA and RNA, electrophoresis can be used to resolve mixtures of proteins according to the length of individual polypeptide chains (Fig. 7-27). After electrophoresis, the proteins can be visualized with a stain, such as **Coomassie Brilliant Blue**, that binds to protein nonspecifically. When the SDS is omitted, electrophoresis can be used to separate proteins according to properties other than molecular weight, such as net charge and isoelectric point (see the later discussion).

### Antibodies Are Used to Visualize Electrophoretically Separated Proteins

Proteins are, of course, quite different from DNA and RNA, but the procedure known as **immunoblotting**, by which an individual protein is visu-



**FIGURE 7-27** SDS gel electrophoresis. A mixture of three proteins of different size are illustrated (much more complex mixtures are usually analyzed). Addition of SDS (shown in red) and  $\beta$ -mercaptoethanol denatures the proteins and provides each with a uniform negative charge. Separation on the basis of size is achieved by electrophoresis.



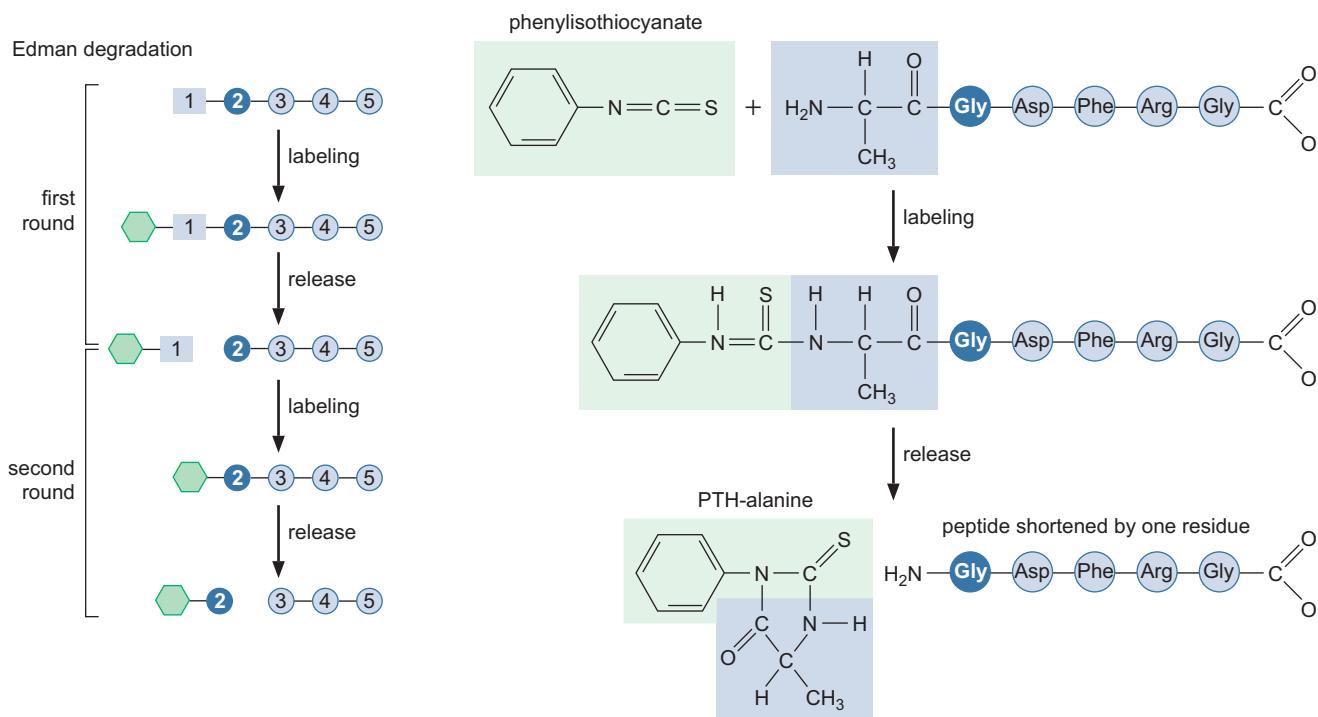
**FIGURE 7-28 Immunoblotting.** After proteins are separated by electrophoresis, they are transferred to filter paper (again using an electric field) in a manner that retains the same relative position of the proteins. After blocking nonspecific protein-binding sites, antibody to the protein of interest is added to the filter paper. The site of antibody binding is then detected using an attached enzyme that creates light when it acts on its substrate.

alized amid thousands of other proteins, is analogous in concept to Southern and northern blot hybridization (Fig. 7-28). Indeed, another name for immunoblotting is “western blotting” in homage to its similarity to these earlier techniques. In immunoblotting, electrophoretically separated proteins are transferred to a filter that nonspecifically binds proteins. As for Southern blotting, proteins are transferred to the membrane such that their position on the membrane mirrors their position in the original gel. Once the proteins are attached to the membrane, all of the remaining nonspecific binding sites are blocked by incubating with a solution of proteins unrelated to those being studied (often this is powdered milk, which primarily contains albumin proteins). The filter is then incubated in a solution of an antibody that specifically recognizes the protein of interest. The antibody can only bind to the filter if it finds its target protein on the filter. Finally, a chromogenic enzyme that is artificially attached to the antibody (or to a second antibody that binds the first antibody) is used to visualize the filter-bound antibody. Southern blotting, northern blotting, and immunoblotting have in common the use of **selective reagents** to visualize particular molecules in complex mixtures.

### Protein Molecules Can Be Directly Sequenced

Although more complex than the sequencing of nucleic acids, protein molecules can also be sequenced: that is, the linear order of amino acids in a protein chain can be directly determined. Two widely used methods for determining protein sequence are Edman degradation using an automated protein sequencer and tandem mass spectrometry. The ability to determine a protein’s sequence is very valuable for protein identification. Furthermore, because of the vast resource of complete or nearly complete genome sequences, the determination of even a small stretch of protein sequence is often sufficient to identify the gene that encoded that protein by finding a matching protein coding sequence.

**Edman Degradation** Edman degradation is a chemical reaction in which the amino acid’s residues are sequentially released from the amino terminus of a polypeptide chain (Fig. 7-29). One key feature of this method is that the amino-terminal-most amino acid in a polypeptide chain can be specifically modified by a chemical reagent called **phenylisothiocyanate (PITC)**, which modifies the free  $\alpha$ -amino group. This derivatized amino acid is then cleaved off the polypeptide by treatment with acid under conditions that do not destroy the remaining peptide bonds. The identity of the released amino acid derivative can be determined by its elution profile using a column chromatography method called high-performance liquid



**FIGURE 7-29** Protein sequencing by Edman degradation. The amino-terminal residue is labeled and can be removed without hydrolyzing the rest of the peptide. Thus, in each round, one residue is identified, and that residue represents the next one in the sequence of the peptide.

chromatography (HPLC) (each of the amino acids has a characteristic retention time). Each round of peptide cleavage regenerates a normal amino terminus with a free  $\alpha$ -amino group. Thus, Edman degradation can be repeated for numerous cycles, and thereby reveal the sequence of the amino-terminal segment of the protein. In practice, eight to 15 cycles of degradation are commonly performed for protein identification. This number of cycles is nearly always sufficient to identify an individual protein uniquely.

Amino-terminal sequencing by automated Edman degradation is a robust technique. Problems arise, however, when the amino terminus of a protein is chemically modified (e.g., by formyl or acetyl groups). Such blockage may occur *in vivo* or during the process of protein isolation. When a protein is amino-terminally blocked, it can usually be sequenced after digestion with a protease to reveal an internal region for sequencing.

**Tandem Mass Spectrometry (MS/MS)** Tandem mass spectrometry (MS/MS) can also be used to determine protein sequence and is the most common method in use today. Mass spectrometry is a method in which the mass of very small samples of a material can be determined with great accuracy. Very briefly, the principle is that material travels through the instrument (in a vacuum) in a manner that is sensitive to its mass/charge ratio. For small biological macromolecules such as peptides and small proteins, the mass of a molecule can be determined with the accuracy of a single dalton.

To use MS/MS to determine protein sequence, the protein of interest is usually digested into short peptides (often less than 20 amino acids) by digestion with a specific protease such as trypsin. This mixture of peptides is subjected to mass spectrometry, and each individual peptide will be separated from the others in the mixture by its mass/charge ratio. The individual

peptides are then captured and fragmented into all of the component peptides, and the mass of each of these component fragments is then determined (Fig. 7-30). Deconvolution of these data reveals an unambiguous sequence of the initial peptide. As with Edman degradation, the sequence of a single, approximately 15-amino-acid peptide from a protein is nearly always sufficient to identify the protein by comparison of the sequence to those predicted by DNA sequences. In contrast to Edman degradation, MS/MS will frequently determine the sequence of many peptides derived from an individual protein.

MS/MS has revolutionized protein sequencing and identification. Only very small amounts of material are needed, and complex mixtures of proteins can be analyzed simultaneously.

## PROTEOMICS

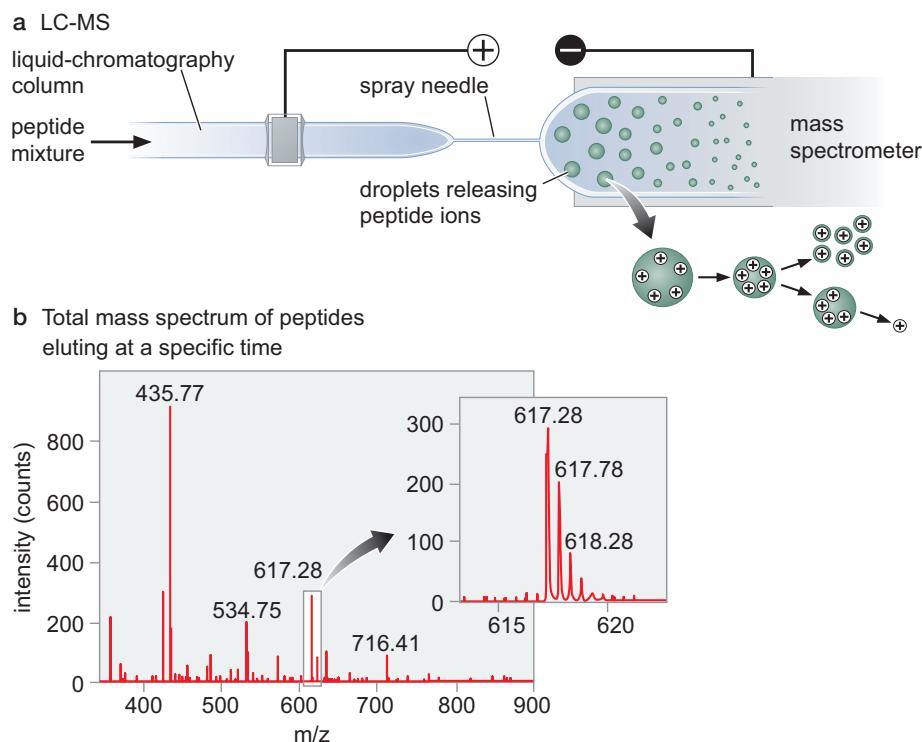
---

Determining the global levels of gene expression provides a rapid snapshot of the activity of a cell; however, there are important additional levels of regulation that cannot be monitored in this manner. Indeed, the level of transcription of a gene gives only a rough estimate of the level of expression of the encoded protein. If the mRNA is short-lived or poorly translated, then even an abundant mRNA will produce relatively little protein. In addition, many proteins are post-translationally modified in ways that profoundly affect their activities, and transcription profiling gives no data regarding this level of regulation.

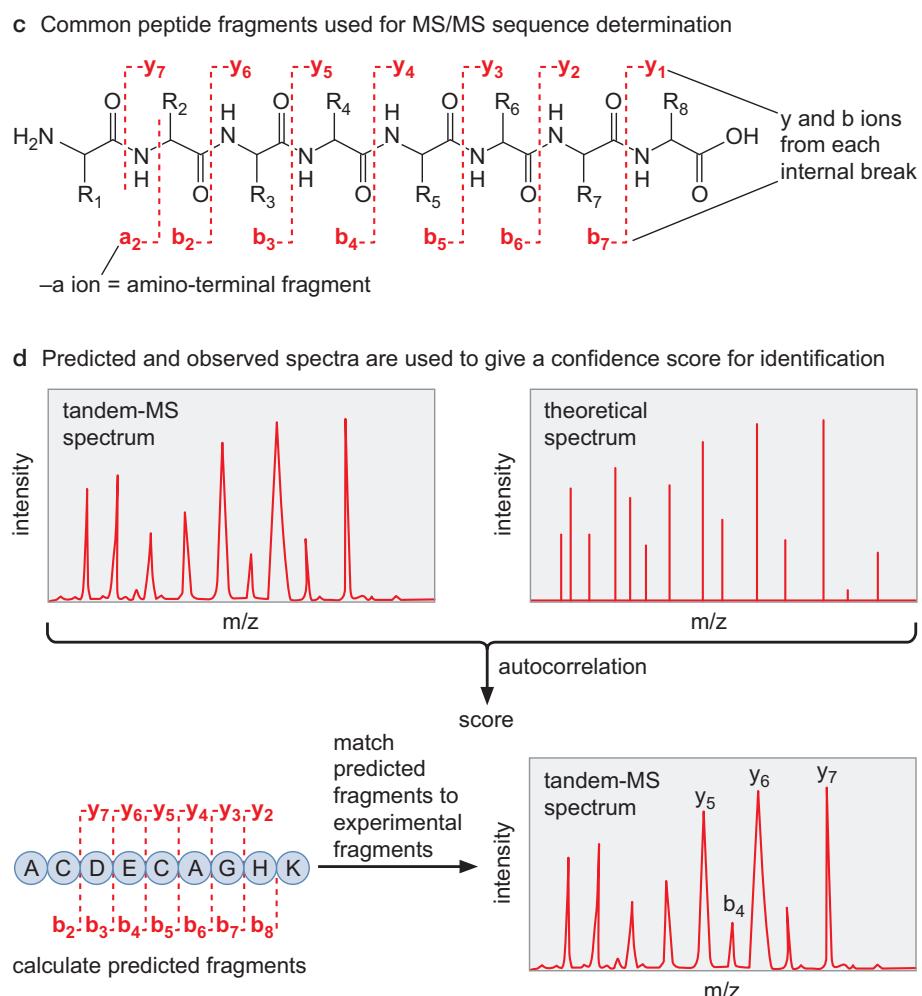
The availability of whole-genome sequences in combination with high-throughput analytic methods for protein separation and identification has ushered in the field of proteomics. The goal of proteomics is the identification of the full set of proteins produced by a cell or tissue under a particular set of conditions (called a proteome), their relative abundance, their modifications, and their interacting partner proteins. Whereas microarray analysis makes it possible to profile gene expression or DNA content on a genome-wide basis, the tools of proteomics seek to capture a similar snapshot of the cell's entire repertoire of proteins and their modifications (e.g., phosphorylation sites).

### Combining Liquid Chromatography with Mass Spectrometry Identifies Individual Proteins within a Complex Extract

A powerful method to identify all of the proteins in a complex mixture such as a crude cell extract uses a combination of liquid chromatography and mass spectrometry (described in the preceding section of this chapter). Although ideally one would simply analyze all of the proteins in a cell extract directly by mass spectrometry, in practice, the very high number of proteins present in such a mixture results in more peptides than can be resolved. Instead, researchers have developed powerful methods in which peptides are separated by two types of liquid chromatography before mass spectrometric analysis (LC-MS) (Fig. 7-31). In this approach, a crude cell extract is first digested with a sequence-specific protease (e.g., trypsin, which cleaves proteins after Arg and Lys residues) to generate peptides. The resulting mixture of peptides is fractionated by ion exchange chromatography (peptides are separated based on ionic interactions with the charged column material) and reverse phase chromatography (peptides are separated based on hydrophobic interactions with the column material). This procedure separates the highly complex, initial collection of peptides into many lower-complexity mixtures of peptides that can be distinguished from one another and sequenced more readily. Each subset of peptides is subjected to tandem mass spectrometry



**FIGURE 7-30** Using liquid chromatography-MS/MS to analyze the content of a protein mixture. (a) A peptide mixture is subjected to liquid chromatography followed by mass spectrometry. (b) As sets of peptides elute from the chromatography column, they are separated by mass and the results are displayed according to their mass/charge ratio ( $m/z$ ). Selected sets of related peptides (the differences between these closely related peaks are due to the presence of different atomic isotopes in the peptide) are fragmented, and the resulting peptide fragments are analyzed in a second round of mass spectroscopy. (c) Fragmentation of the peptide commonly breaks the peptide in the sites shown in the figure. The possible subpeptides that are generated are called b peptides (amino-terminal fragments), y peptides (carboxy-terminal fragments), and the a<sub>2</sub> peptide (the shortest amino-terminal fragment). (d) The observed spectra are compared with all of the possible theoretical spectra that are generated from the amino acid sequences of the proteins encoded by the organism from which the proteins were isolated. Typically, only a subset of peptides can be unambiguously identified. For example, Ile and Leu have identical masses. Nevertheless, clear identification of as few as three or four peptide fragments from a parental peptide is usually sufficient to identify the protein.



(MS/MS, discussed above) to sequence as many peptides in the population as possible. Finally, given a complete genome sequence for the organism under study and the peptide sequences from the mass spectrometric analysis, the tools of bioinformatics make it possible to assign each peptide to a particular protein-coding sequence (gene) in the genome.

In practice, this method detects only a subset of the proteins in a complex mixture of proteins such as that derived from an entire cell. A typical analysis can detect approximately 1000 different proteins. Nevertheless, additional fractionation methods and enhanced sensitivity of mass spectrometry can increase the completeness of these protein profiles in the future. Although LC-MS analysis is very good at identifying which proteins are present in a cell extract, currently it is more difficult to determine the relative abundance of proteins by this approach. To address this weakness, new technologies that quantify the abundance are being developed and have been used in some cases.

### Proteome Comparisons Identify Important Differences between Cells

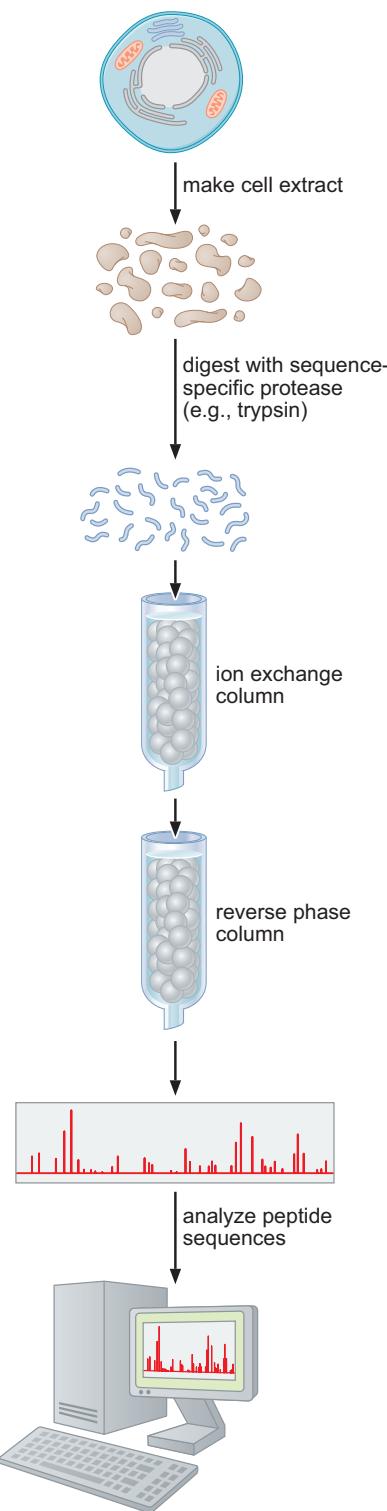
Although knowing the full complement of proteins in a cell has intrinsic value, in most cases, it is the differences between two cell types or between cells exposed to two different growth conditions that are most valuable. By determining the proteome in each situation, the differences in the proteins present can be determined. In turn, this analysis can identify proteins that are likely to be responsible for cellular differences and, therefore, represent good candidates for further study.

The value of comparative proteomics can be seen in an analysis of different cancer cells. It is frequently found that different individuals with apparently the same type of cancer respond very differently to the same chemotherapeutic treatment. By comparing the proteomes of different tumor samples, the apparently similar cells are found to have important differences in the proteins that they express. These differences can become valuable markers to distinguish between the different tumor types. More importantly, these markers can be used to select the most effective chemotherapies for each patient.

### Mass Spectrometry Can Also Monitor Protein Modification States

Because the modification state of a protein can profoundly affect its function, efforts are also underway to comprehensively identify the modification state of proteins in the cell. Specific modifications are commonly used to alter the activity or stability of a protein. For example, phosphorylation of proteins is used extensively to control their activity. Phosphorylation can cause a protein to alter its conformation in a functionally important manner (e.g., many protein kinases are only active after they are phosphorylated). Alternatively, the attachment of a phosphate can create a new binding site for another protein on the surface of the protein, leading to the assembly of new protein complexes. Other protein modifications include methylation, acetylation, and ubiquitylation. The last of these involves the attachment of the 76-amino-acid protein ubiquitin to a lysine residue via a pseudopeptide bond. Modification of a protein with multiple ubiquitin typically targets the protein for degradation.

Each type of modification causes a discrete change in the molecular mass of the protein. This can be monitored by mass spectrometry, and methods



**FIGURE 7-31** Separation of proteins by liquid chromatography followed by mass spectrometric analysis. The steps of the method are illustrated in the figure and described in the text.

have been developed to identify proteomes that include only those proteins with a particular modification. For example, the complete set of phosphorylated proteins in the cell is called the “phosphoproteome.” Methods to identify the subset of proteins that include a particular modification have been developed and generally exploit affinity resins that will specifically bind the modification of interest. For example, resins that include immobilized Fe<sup>3+</sup> (also called immobilized metal affinity chromatography [IMAC]) specifically bind phosphorylated peptides. Mixtures of peptides derived from crude cell extracts can be incubated with such a resin, and the small proportion of peptides that bind are enriched for phosphopeptides. These peptides can then be analyzed using LC-MS to identify the proteins that are modified and the sites of modification. This information is a valuable tool to identify the kinase that modified the protein and to test the importance of the modification by generating mutant proteins that cannot be modified.

### Protein–Protein Interactions Can Yield Information regarding Protein Function

Proteomics is also concerned with identifying all of the proteins that associate with another protein in a cell to generate what are called **interactomes**. A complete interactome for a cell would indicate all interactions between proteins in the cell. In what can be considered guilt by association, such interactions can be used to determine which processes a protein may be involved in. Proteins that are part of the same protein complex will frequently be involved in the same cellular process.

One method for determining protein–protein interactions is the yeast two-hybrid assay (see Chapter 19, Box 19-1), whereby the protein of interest serves as “bait” and a library of proteins can be tested as potential “prey.” A second approach is to use affinity resins or immunoprecipitation to rapidly purify a protein of interest along with any associated proteins. The resulting mixture of proteins can then be analyzed by LC-MS to identify the associated proteins. By repeating this procedure with all of the proteins in a cell, it is possible to obtain a comprehensive interaction diagram of protein–protein interactions within a cell.

The latter approach has been applied to the yeast *Saccharomyces cerevisiae*. More than 6000 *S. cerevisiae* proteins were purified by affinity chromatography (the gene for each protein was genetically modified or “tagged” to append a short carboxy-terminal extension that is known to bind two affinity resins), and mass spectrometry was used to identify any additional proteins that copurified with the tagged protein. Comparison of these data identified hundreds of protein complexes present in the cell—many of which were already known, but some of which were novel. The effectiveness of this study can be seen by the detection of a large number of well-documented protein complexes (e.g., RNA polymerase II) (Fig. 7-32).

## NUCLEIC ACID–PROTEIN INTERACTIONS

We now turn our attention to the various methods that can be used to detect the interactions between nucleic acids and proteins. These interactions are critical to determining the specificity and precision of the events described in this book. Be it transcription, recombination, DNA replication, DNA repair, mRNA splicing, or translation, the proteins that mediate these events must recognize particular nucleic acid structures or sequences to ensure that these events occur at the right place and time in the cell.



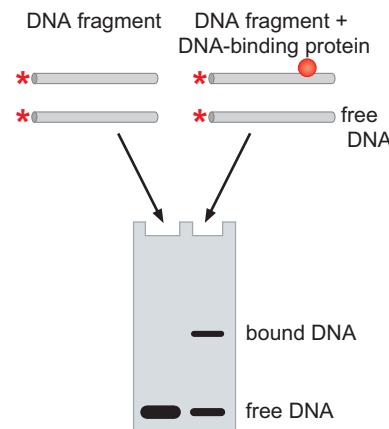
**FIGURE 7-32** The physical interactome map of *S. cerevisiae*. Shown here are the results of affinity purification/mass spectrometry studies of all of the proteins in *S. cerevisiae*. The figure is actually composed of a series of columns of boxes indicating which proteins coprecipitated with a given protein. If a protein is coprecipitated with the “tagged” protein, the box is yellow. If not, the box is black. In this view, proteins that are in the same complex have been clustered together on both the vertical and horizontal axes; thus, complexes are observed on the diagonal. A subset of all of the complexes (many of which are discussed elsewhere in the text) are labeled and shown in the image presented here. (Reprinted, with permission, from Collins S.R. et al. 2007. *Mol. Cell. Proteom.* 6: 439–450, Fig. 3b. © American Society for Biochemistry and Molecular Biology.)

Consistent with the importance of understanding nucleic acid–protein interaction, there are numerous robust assays that can be used to measure these events both *in vivo* and *in vitro*. In the following sections, we consider several of these assays, comparing their strengths and weaknesses.

### The Electrophoretic Mobility of DNA Is Altered by Protein Binding

Just as electrophoretic mobility can be used to determine the relative sizes of DNA, RNA, or protein molecules, it can also be used to detect protein–DNA interactions. If a given DNA molecule has a protein bound to it, migration of that DNA–protein complex through the gel is retarded compared with migration of the unbound DNA molecule. This forms the basis of an assay to detect specific DNA-binding activities. The general approach is as follows: a short double-stranded DNA (dsDNA) fragment containing the binding site of interest is radioactively labeled so that it can be detected in small quantities by polyacrylamide gel electrophoresis and autoradiography. A fluorescent label can also be used, but it is important that the fluorophor (see Chapter 9, Box 9-1 Fig. 1b) is not in a position that interferes with DNA binding. The resulting DNA “probe” is then mixed with the protein of interest, and the mixture is separated on a nondenaturing gel. If the protein binds to the probe DNA, the protein–DNA complex migrates more slowly, resulting in a shift in the location of the labeled DNA (Fig. 7-33). For this reason, this assay is referred to as an **electrophoretic mobility-shift assay** (EMSA) or, more colloquially, a band or gel shift assay. We have described this assay for a dsDNA fragment; however, the same approach can also be used to detect binding to single-stranded DNA (ssDNA) or to RNA.

EMSA can also be used to monitor the association of multiple proteins with the same DNA. These interactions can each be due to sequence-specific DNA binding. Alternatively, after an initial sequence-specific interaction, the subsequent protein can bind to the first DNA-bound protein. In either case, as an additional protein binds, it will further reduce the mobility of the DNA fragment. Using the EMSA in this way can be a very powerful method to identify how a series of proteins interacts with DNA.



**FIGURE 7-33** Electrophoretic mobility-shift assay. The principle of the mobility-shift assay is shown schematically. A protein is mixed with radiolabeled probe DNA containing a binding site for that protein. The mixture is resolved by acrylamide gel electrophoresis and visualized using autoradiography. DNA not mixed with protein runs as a single band corresponding to the size of the DNA fragment (left lane). In the mixture with the protein, a proportion of the DNA molecules (but not all of them at the concentrations used) binds the DNA molecule. Thus, in the right-hand lane, there is a band corresponding to free DNA and another corresponding to the DNA fragment in complex with the protein.

interdependently. Different proteins binding to the same DNA probe also can be distinguished because proteins of different size will affect the mobility of the DNA to different extents—the larger the protein, the slower the migration. If two proteins cause a shift to the same extent, then a second method can be used to distinguish which one is bound. The addition of an antibody directed against a protein will cause a “supershift” if that protein is associated with the DNA. Thus, by adding an antibody to a potential binding protein, the presence of the protein in the protein–DNA complex can be assessed.

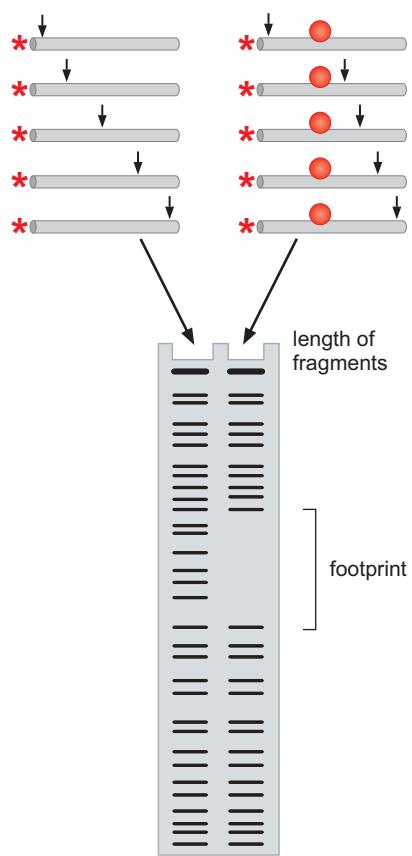
One weakness of EMSAs is that they do not reveal intrinsically what sequence in the DNA the protein binds. Two types of additional experiments can be performed to identify the protein-binding site within the DNA probe. One approach is to add an excess of short dsDNA oligomers to the protein before incubation with the DNA probe. If the protein-binding site is contained within the oligomer, then the protein will bind the dsDNA oligomer instead of a specific, DNA probe. Alternatively, mutations can be made in the DNA probe to assess their effect on protein binding. Although these approaches can be taken without knowledge of potential binding sites, in most instances, prior experiments or the conservation of certain DNA sequences within the DNA probe help to simplify the choice of sequences to test.

### DNA-Bound Protein Protects the DNA from Nucleases and Chemical Modification

How can a protein-binding site in DNA be identified more readily? A series of powerful approaches allows identification of the DNA site bound by the protein and of the chemical groups in the DNA (methyl, amino, or phosphate) that the protein contacts. The basic principle that underlies these methods is as follows: if a DNA fragment is labeled with a radioactive atom only at one end of one strand, then the location of any break in this strand can be deduced from the size of the labeled fragment that results. The size, in turn, can be determined by high-resolution denaturing electrophoresis in a polyacrylamide gel (similar to the gels used to analyze DNA sequencing products) followed by detection of the labeled ssDNA fragments. For reasons that will become clear, these methods are generally called **DNA footprinting**.

The most common of these approaches is **nuclease protection footprinting**. After incubating the DNA and the end-labeled DNA together, the resulting complexes are briefly exposed to a DNA nuclease (most often DNase I, which cuts one strand of the target dsDNA). DNA sites bound by protein are protected from nuclease cleavage, creating a region of the DNA without cut sites (Fig. 7-34). The resulting “footprint” is revealed by the absence of bands of sizes that correspond to the site of protein binding. The related **chemical protection footprinting** relies on the ability of a bound protein to protect bases in the binding site from base-specific chemical reagents that (after a further reaction) give rise to backbone cuts. In both methods, it is important that the number of nuclease cut sites or chemical modifications is titrated to be approximately one per DNA probe. This is because only the cut site that is nearest the labeled DNA end will be detected after gel electrophoresis and labeled DNA detection.

By changing the order of the first two steps, a third method, **chemical interference footprinting**, determines which features of the DNA structure are *necessary* for the protein to bind. Before protein is added to the DNA, an average of one chemical change per DNA is made. The modified DNA is incubated with the DNA-binding protein, and protein–DNA complexes



**FIGURE 7-34** Nuclease protection footprinting. (Stars) The radioactive labels at the ends of the DNA fragments; (arrows) sites where DNase cuts; (red circles) *Lac* repressor bound to operator. On the left, DNA molecules cut at random by DNase are separated by size using gel electrophoresis. On the right, DNA molecules are first bound to repressor and then subjected to DNase treatment. The “footprint” is indicated on the right. This corresponds to the collection of fragments generated by DNase cutting at sites in free DNA but not in DNA with repressor bound to it. In the latter case, these sites are inaccessible because they are within the operator sequence and hence covered by repressor.

are isolated. One popular method to separate the protein-bound DNA from unbound DNA is to use the EMSA. After detecting the labeled DNA in the EMSA gel, the shifted (protein-bound) and unshifted (unbound) DNA can easily be separated. If a modification at a particular site does not prevent binding of the protein, DNA isolated from the complex will contain that modification. If, on the other hand, a modification prevents the protein from recognizing the DNA, then no DNA modified at the site will be found in the protein-bound DNA sample. As with the chemical protection assay, the sites of chemical modification are detected by treating the DNA with reagents that cleave the DNA at sites of chemical modification.

The reagents used for chemical modification can probe very specific aspects of the DNA. For example, the chemical ethylnitrosourea (ENU) specifically modifies the phosphate residues in the backbone of DNA. Other chemicals specifically modify certain bases in the major or the minor groove. Using a variety of chemicals can provide a precise understanding of the contacts a particular protein makes with the bases and with the phosphates in the sugar–phosphate backbone of DNA.

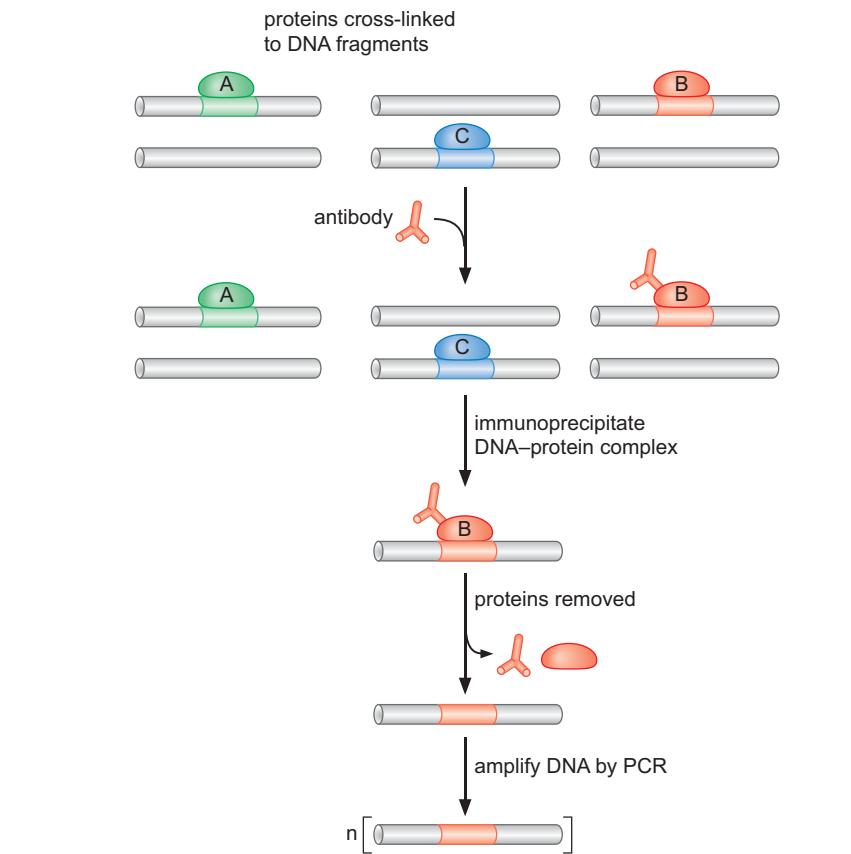
Footprinting is a powerful approach that immediately identifies the site on the DNA to which a protein binds; however, as a group, these methods require more robust DNA binding by a protein. In general, for a DNA footprinting assay to be effective, >90% of the DNA probe must be bound by protein. This level of binding is required because the footprinting assay detects the *lack* of a signal (because of protection from cleavage or modification of the DNA) rather than the appearance of a new band. This situation is in contrast to the more sensitive EMSA, in which protein binding to the labeled DNA results in the formation of a new band in a region of the gel that, in the absence of protein binding, lacks any DNA molecules.

### Chromatin Immunoprecipitation Can Detect Protein Association with DNA in the Cell

Although *in vitro* assays for protein–DNA binding can be informative, it is often important to determine whether a protein binds a particular DNA site in a living cell. For any particular DNA-binding protein, there are many potential binding sites in the entire genome of a cell. Despite this, in many instances, only a subset of these sites will be occupied. In some instances, binding of other proteins may inhibit association of the protein with a potential DNA-binding site. In other instances, binding of adjacent proteins may be required for robust binding. In either case, knowing whether a protein (e.g., a transcriptional regulator; see Chapter 19) is bound to a particular site (e.g., at a particular promoter region) in the cell can be a powerful piece of evidence that it acts to regulate an event occurring at that site (e.g., transcriptional activation).

Chromatin immunoprecipitation, often just called ChIP, is a powerful technique to monitor protein–nucleic acid interactions in the cell. In outline, the technique is performed as follows: formaldehyde is added to living cells, cross-linking DNA to any bound proteins and proteins bound tightly to other proteins. The cross-linked cells are lysed, and the DNA is broken into small fragments (200–300 bp each) by sonication. Using an antibody specific for the protein of interest (e.g., a transcription regulator), the fragments of DNA attached to that protein can be separated from the majority of the DNA in the cell by immunoprecipitation (or IP). Once the immunoprecipitation is complete, the cross-linking between protein and DNA is reversed, allowing analysis of the DNA sequences that are present in the IP (Fig. 7-35).

The most important step of a ChIP experiment is to determine whether a particular region of DNA is bound by the protein and therefore present in the IP.



**FIGURE 7-35** Chromatin immunoprecipitation (ChIP).

This can be accomplished by one of two basic approaches. To determine if a particular region of DNA (e.g., a promoter) is bound by the protein of interest, PCR can be performed using primers that are targeted to that region. If the protein was bound to that DNA at the time of cross-linking, the sequence will be present in the IP and will be amplified. There are two important controls that are generally included in this assay. First, PCR primers targeting another region of DNA (one to which the protein is known or expected not to bind) are used; in that case, no DNA should be amplified (Fig. 7-35). Second, before performing the IP, a small amount of the total DNA is set aside, and both the test and the control primers are used to amplify the DNA in this unfractionated sample. If the PCR primers amplify with the same efficiency, both sequences should be amplified equally from this starting population of DNA. This control ensures that any differences in the extent of PCR amplification of the ChIP DNA using the two different primer sets are due to a difference in abundance and not different efficiency of the PCR.

A second approach to identify the DNA sequences associated with a particular protein in the cell is to use tiling DNA microarrays. In this approach, the DNA that is cross-linked to the protein and the total DNA isolated from the cell are labeled with two different fluorophores (for this example, we refer to them as red and green, respectively). The two populations are mixed together and hybridized to the microarray. Regions with a high red:green ratio will represent binding sites of the protein. Those with a low red:green ratio are regions that are not bound. This approach is particularly powerful because whole genomes can be examined simultaneously, and no prior knowledge of the potential binding site is required. Because this approach

analyzes DNA samples fractionated by ChIP using tiling DNA chips, this approach is commonly referred to as ChIP-Chip.

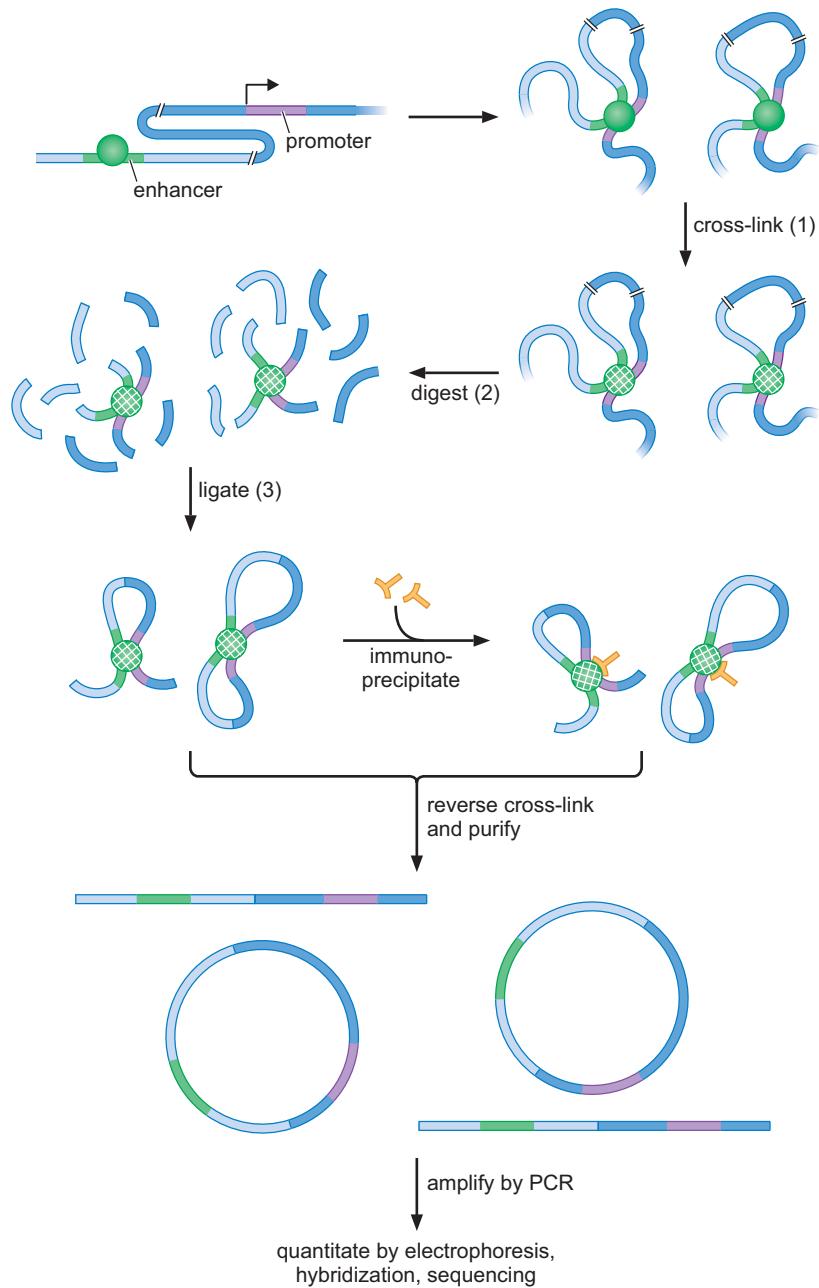
Although ChIP-Chip is very powerful and therefore routinely used, it does have limitations of which the investigator needs to be aware. First, similar to the EMSA, the resolution of ChIP is limited. It is not possible to show that a protein is bound to a specific short DNA sequence, merely that it is bound to a site within a given 200–300-bp fragment. Thus, ChIP is adequate to show that a regulatory protein is bound upstream of one rather than another gene, but it does not show exactly where upstream of the gene the protein is bound. As for the EMSA, mutations in the DNA would be necessary to test whether a protein binds to a specific site. Second, only proteins for which antibodies are available can be studied using ChIP. Even more important, proteins can be identified only if the relevant epitope (the specific region of a protein recognized by an antibody) is exposed when the protein in question is cross-linked to the DNA (and perhaps to other proteins with which it interacts at the gene). In an extension of this complication, if a given protein is not detected under one environmental or physiological condition, but then is detected under another, the obvious interpretation is that the protein binds to that region of DNA only in response to the change in environmental conditions. But an alternative explanation might be that the protein in question is bound all of the time, yet its epitope is concealed by another protein, present under one set of conditions but not the other.

Whole-genome technologies are evolving at a rapid pace, and there are a number of emerging variations in the basic ChIP assays described here. For example, in an approach called ChIP-Seq—immunoprecipitated DNA derived from cross-linked and sheared chromatin is subjected to direct DNA sequencing, and the DNA sequence reads are then aligned on the corresponding genome assembly. The frequent identification of sequences in a particular genomic site is evidence for protein binding to this site. ChIP-Seq is similar to the ChIP-ChIP method, but is sometimes easier because it skips the need for creating whole-genome tiling arrays. One only needs to know the genome sequence of the organism/cell being studied to map the sites of DNA binding. In Chapter 19, we consider these specialized methods in more detail in the context of their applications to identifying enhancers.

### Chromosome Conformation Capture Assays Are Used to Analyze Long-Range Interactions

Chromosomes fold up in various three-dimensional (3D) forms, and these structures influence genome stability (Chapter 10), chromosome segregation (Chapter 8), and gene regulation and activity (Chapter 19). Long-range interactions are known to occur between widely spaced genes and their corresponding regulatory elements, some of which can be found up to several megabases away. In one example, described in more detail in Chapter 21, the enhancer that controls the expression of the Sonic hedgehog gene in the developing limbs of mammalian embryos is located ~1 Mb away from the transcription start site of the gene. Expression depends on the ability of the remote enhancer to loop over long distances to the Sonic hedgehog promoter. Chromosome conformation capture assays (3C) can be used to detect such interactions. The method, illustrated in Figure 7-36, works as follows: the treatment of intact cells with formaldehyde serves to link interacting genomic regions by cross-linking proteins to DNA and proteins to other proteins. The chromatin is then broken up by digestion with restriction endonucleases or by physical disruption, such as sonication. The resulting DNA is subjected to ligation under conditions that favor intramolecular ligation of the associated DNA fragments. At this point, the cross-linking is reversed and the ligation mixture is purified. Alternatively, after ligation, the mixture can be immunoprecipitated with a

**FIGURE 7-36** Chromosome Conformation Capture schematic. 3C assays involve three basic steps: (1) interacting chromosome segments are cross-linked with formaldehyde, (2) the DNA is digested, and (3) cross-linked DNA fragments are ligated to produce products that are amplified and can be further analyzed.



specific antibody that recognizes the protein of interest, as discussed in the preceding section.

4C and 5C assays are variants of the 3C method that permit detection of all chromosomal interactions with a fixed anchor point within the genome. In one striking example, these approaches were used to study “the archipelago of enhancers” regulating the *HoxD* locus. The *HoxD* gene cluster is involved in organizing growth patterns, in particular, of developing limbs. Transcription of these genes is coordinated in waves, activated by regulatory sequences that lie several hundred kilobases from the gene cluster. The first wave of expression occurs during early limb development, controlled by one set of enhancers, whereas a second set of genes is expressed in a late wave of transcription that occurs with digit formation, controlled by yet another set of enhancers. Using 3C-related techniques, investigators determined that

several enhancers, shown to be distributed over an 800-kb interval, interact with the *Hoxd13* promoter.

### In Vitro Selection Can Be Used to Identify a Protein's DNA- or RNA-Binding Site

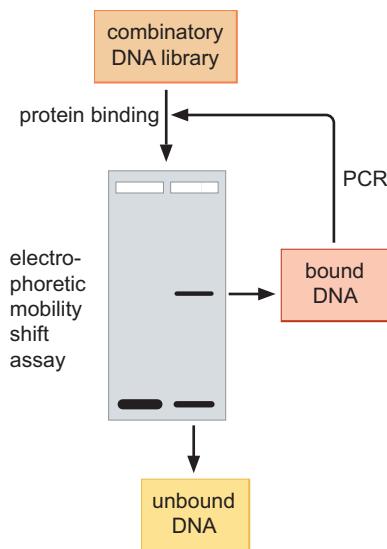
As more and more DNA-binding proteins are identified and understood, the amino acid motifs associated with sequence-specific DNA binding have become relatively easy to identify (e.g., helix-turn-helix motifs; see Chapter 6). Despite these findings, our understanding of these nucleic acid–binding protein domains has not evolved to the point that the primary amino acid sequence of a protein is sufficient to reveal the DNA sequence to which it binds. And yet, this information is often very important for identifying potential regulatory regions that can be targeted for subsequent analysis.

How can the DNA sequence recognized by a particular protein be identified? One powerful approach, called **in vitro selection** or **SELEX** (for Systematic Evolution of Ligands by Exponential Enrichment), involves the use of the sequence specificity of the protein to probe a diverse library of oligonucleotides. By characterizing the enriched DNA, the sequences that bind tightly to the protein can be identified.

The first step in this method is to produce a large library of ssDNA oligonucleotides using chemical DNA synthesis (which we describe in the first part of this chapter). Importantly, the middle 10–12 bases of these oligonucleotides are randomized (by adding a mixture of all four nucleotide precursors to these steps in oligonucleotide synthesis). The randomized region of each nucleotide is flanked on either side by defined sequences. After the oligonucleotide library is synthesized, a short primer is annealed to the defined 3' end of the oligonucleotides and extended to convert the randomized ssDNA library to a randomized dsDNA library.

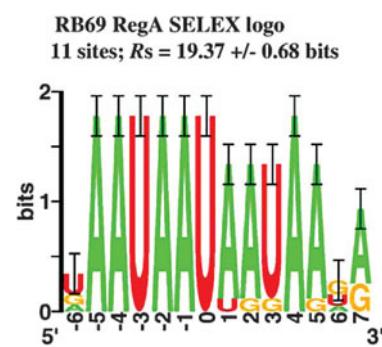
Enriching for oligonucleotides that bind the protein of interest can be accomplished using methods similar to those we have already discussed. After incubation of the protein with the library of oligonucleotides, the entire reaction can be separated in an EMSA. The DNA in the shifted complex will be strongly enriched for DNA sequences that are tightly bound by the protein. Alternatively, if an antibody is available that recognizes the protein of interest, an immunoprecipitation, similar to the ChIP assay, can be used to separate protein and bound DNA from unbound DNA. Regardless of the mechanism of enrichment, PCR is then used to amplify the bound DNA (using short oligonucleotides that hybridize to the nonrandomized end regions of the dsDNA library). This amplification step is necessary because only a small percentage of the starting oligonucleotides will bind to the protein. Repeating the binding, enrichment, and amplification steps will greatly enrich for the sequences that are most tightly bound by the protein of interest (Fig. 7-37). Typically, three to five rounds of enrichment are performed to identify the DNA sequences that are most tightly associated with the protein of interest.

The DNA sequence specificity of the protein can be determined by sequencing a subset of the enriched DNAs. Typically only a subset of the sequences within the randomized region will be conserved, because most DNA-binding proteins do not recognize more than six or seven nucleotides. Computational analysis is generally used to assist in identifying the most conserved sequences. The final sequence of bases can be represented by a sequence logo, in which the size of the G, A, T, or C characters represents the frequency of appearance of each nucleotide in the library of enriched oligonucleotides (Fig. 7-38).



**FIGURE 7-37** In vitro selection scheme.

A combinatorial DNA library in which the middle 10–12 bases are randomized is bound to the protein of interest. Protein-bound DNA is separated from unbound DNA using an EMSA. Bound DNA is eluted from the gel and subjected to PCR using primers directed against constant regions flanking the random regions of the DNA. These sequences are subjected to two to five more cycles of binding and enrichment to identify the highest affinity.



**FIGURE 7-38** SELEX sequence logo. In vitro selection was used to isolate RNAs that bind the translational repressor protein RB69 RegA. The image shows the logo of selected sequences. The letter height is proportional to the frequency of each base at that position, with the most frequently occurring base at the top. (Reprinted, with permission, from Dean T.R. et al. 2005. *Virology* 336: 26–36, Fig. 4a. © Elsevier.)

## BIBLIOGRAPHY

---

### Books

- Green M. and Sambrook J. 2012. *Molecular cloning: A laboratory manual*, 4th ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
- Griffiths A.J.F., Gelbart W.M., Lewontin R.C., and Miller J.H. 2002. *Modern genetic analysis*, 2nd ed. W.H. Freeman, New York.
- Hartwell L., Hood L., Goldberg M.L., Reynolds A.E., Silver L.M., and Veres R.C. 2003. *Genetics: From genes to genomes*, 2nd ed. McGraw-Hill, New York.
- Snustad D.P. and Simmons M.J. 2002. *Principles of genetics*, 3rd ed. Wiley, New York.

### Genomic Analysis

- Frazer K.A., Pachter L., Poliakov A., Rubin E.M., and Dubchak I. 2004. VISTA: Computational tools for comparative genomics. *Nucleic Acids Res.* **32**: W273–W279.

- Human Genome. 2001. *Nature* **409**: 813–960.
- Human Genome. 2001. *Science* **291**: 1145–1434.
- International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431**: 931–945.
- Mouse Genome. 2002. *Nature* **420**: 509–590.
- Osoegawa K., Mammoser A.G., Wu C., Frengen E., Zeng C., Catanese J.J., and de Jong P.J. 2001. A bacterial artificial chromosome library for sequencing the complete human genome. *Genome Res.* **11**: 483–496.
- The Human Genome at Ten. 2011. *Nature* **464**: 649–671.

### Proteomic Analysis

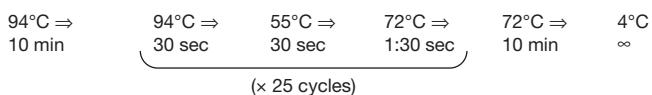
- Yates J.R. III, Gilchrist A., Howell K.E., and Bergeron J.J. 2005. Proteomics of organelles and large cellular structures. *Nat. Rev. Mol. Cell. Biol.* **6**: 702–714.

## QUESTIONS

### MasteringBiology®

*For instructor-assigned tutorials and problems, go to MasteringBiology.*

For answers to even-numbered questions, see Appendix 2: Answers.



**Question 1.** How does DNA migrate through a gel when an electrical field is applied for electrophoresis? Explain your choice and how the DNA is visualized after electrophoresis.

**Question 2.** The restriction endonuclease, XbaI, recognizes the sequence 5'-CTCGAG-3' and cleaves between the C and T on each strand.

- A. What is the calculated frequency of this sequence occurring in a genome?
- B. The restriction endonuclease, SalI, recognizes the sequence 5'-GTCGAC-3' and cleaves between the G and T on each strand. Do you think the sticky ends produced after XbaI and SalI cleavage could adhere to each other? Explain your choice.

**Question 3.** Generally describe two methods for labeling a DNA probe.

**Question 4.** Compare and contrast Southern blot and northern blot.

**Question 5.** Plasmid cloning vectors are specially designed to possess several features that are useful for cloning and expression. In a sentence or two, describe the role of each of the following features: origin of replication, restriction enzyme recognition sites, selectable marker, and promoter.

**Question 6.** Explain how a genomic DNA library differs from a cDNA library. What is the advantage of using a cDNA library?

**Question 7.** The following times and temperatures are an example of the steps for PCR. You can use Figure 7-12 to help you answer the following questions.

- A. Why is the first step is carried out at 94°C?
- B. What happens in the reaction when the temperature shifts to 55°C during cycling?
- C. During cycling, what occurs when the temperature is at 72°C?

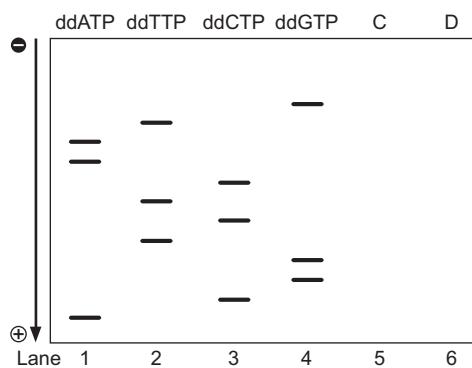
**Question 8.** Describe the basis for separation of proteins for ion-exchange, gel-filtration, and affinity column chromatography.

**Question 9.** Explain the purpose of adding SDS to protein samples for polyacrylamide gel electrophoresis.

**Question 10.** Three assays for testing interactions between protein and DNA are the electromobility shift assay (EMSA), DNA footprinting, and chromatin immunoprecipitation (ChIP).

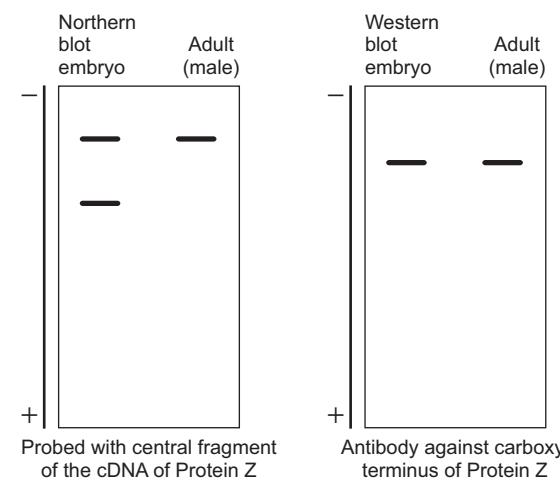
- A. In a DNA footprinting assay, explain why only one strand of the DNA can be end-labeled for the experiment to work.
- B. Following the immunoprecipitation step in a chromatin immunoprecipitation (ChIP), explain how to identify the DNA sequences that remain bound to the protein of interest.

**Question 11.** You decide to perform dideoxy sequencing on a PCR product. You add the appropriate <sup>32</sup>P-labeled primer, DNA polymerase, DNA template (the PCR product), buffer, dNTP mix, and a small amount of one of the four ddNTPs to four reaction tubes. You run the reactions in the thermal cycler, load each reaction into a separate lane of a polyacrylamide gel, and separate the products by gel electrophoresis. In the figure below, the lanes are labeled according to the ddNTP added.



- What is the sequence of the **template** strand? Be sure to label the 5' and 3' ends.
- Suppose that you accidentally added 10-fold more ddGTP to the reaction in lane 4. What effect would that have on the banding pattern in lane 4?
- In lane 5, draw what you would expect to see if you prepared a reaction using a nucleotide mix containing only dATP, dTTP, dCTP, and dGTP.
- In lane 6, draw what you would expect to see if you prepared a reaction using a nucleotide mix containing only ddATP, dTTP, dCTP, dGTP.

**Question 12.** You want to characterize the developmental expression of the gene in *Drosophila melanogaster*. You isolate mRNA from embryos and adult flies and perform a northern blot using a labeled DNA probe specific to Gene Z mRNA, a gene required for development. The results are depicted below.



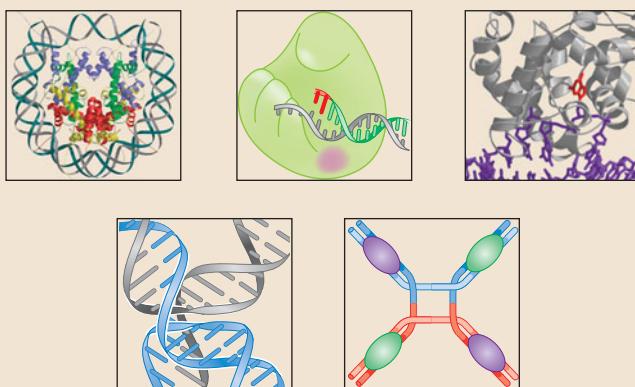
Intrigued, you isolate protein Z from embryos and adult flies and perform a western blot using an antibody against the carboxyl terminus of the protein. The results are depicted below. You are surprised to find a single band of the same molecular weight in both embryos and adult flies.

- Propose a hypothesis to explain these results.
- Propose a modification to the western blot experimental strategy that would allow you to test your hypothesis. Assume you have access to any necessary reagents.

*This page intentionally left blank*

P A R T      3

# MAINTENANCE OF THE GENOME



## O U T L I N E

---

- **CHAPTER 8**  
Genome Structure, Chromatin,  
and the Nucleosome, 199
- **CHAPTER 9**  
The Replication of DNA, 257
- **CHAPTER 10**  
The Mutability and Repair  
of DNA, 313
- **CHAPTER 11**  
Homologous Recombination  
at the Molecular Level, 341
- **CHAPTER 12**  
Site-Specific Recombination and  
Transposition of DNA, 377

PART 3 IS DEDICATED TO THE PROCESSES THAT propagate, maintain, and alter the genome from one cell generation to the next. In Chapters 8–12, we shall examine the following.

- How are the very large DNA molecules that make up the chromosomes of eukaryotic organisms packaged within the nucleus?
- How is DNA replicated completely during the cell cycle, and how is this achieved with high fidelity?
- How is DNA protected from spontaneous and environmental damage, and how is damage, once inflicted, reversed?
- How are DNA sequences exchanged between chromosomes in processes known as recombination and transposition, and what are the biological roles of these processes?

In answering these questions, we shall see that the DNA molecule is subject both to conservative processes that act to maintain it unaltered from generation to generation and to other processes that bring about profound changes in the genetic material that help drive organism diversity and evolution.

We start, in Chapter 8, by describing how the very large DNA molecules that make up chromosomes vary in their organization and size between different organisms. The large size of chromosomal DNA requires that it is not naked but packaged into a more compact form to fit inside the cell. The packaged form of DNA is called chromatin. Being packaged in this way not only reduces the length of the chromosomes but also alters the accessibility and behavior of the DNA. In addition, chromatin can be modified to increase or decrease that accessibility. These changes contribute to ensuring it is replicated, recombined, and transcribed at the right time and in the right place. Chapter 8 introduces us to the histone and nonhistone components of chromatin, to the structure of chromatin, and to the enzymes that modulate the accessibility of the chromosomal DNA.

The structure of DNA offered a likely mechanism for how genetic material is duplicated. Chapter 9 describes this copying mechanism in detail. We describe the enzymes that synthesize DNA and the complex molecular machines that allow both strands of the DNA to be replicated simultaneously. We also discuss how the process of DNA replication is initiated and how this event is carefully regulated by cells to ensure the appropriate chromosome number is maintained.

But the replication machinery is not infallible. Each round of replication results in errors, which, if left uncorrected, would become mutations in daughter DNA molecules. In addition, DNA is a fragile molecule that undergoes damage spontaneously and from chemicals and radiation. Such damage must be detected and mended if the genetic material is to avoid rapidly accumulating an unacceptable load of mutations. Chapter 10 is devoted to the mechanisms that detect and repair damage in DNA. Organisms from bacteria to humans rely on similar, and often highly conserved, mechanisms for preserving the integrity of their DNA. Failure of these systems has catastrophic consequences, such as cancer.

The final two chapters of Part 3 reveal a complementary aspect of DNA metabolism. In contrast to the conservative processes of replication and repair, which seek to preserve the genetic material with minimal alteration, the processes considered in these chapters are designed to bring about new arrangements of DNA sequences. Chapter 11 covers the topic of homologous recombination—the process of breakage and reunion by which very similar chromosomes (homologs) exchange equivalent segments of DNA.

Homologous recombination, which allows both the generation of genetic diversity and the replacement of missing or damaged sequences, is a major mechanism for repairing broken DNA molecules. Models for pathways of homologous recombination are described, as well as the fascinating set of molecular “machines” that search for homologous sequences between DNA molecules and then create and resolve the intermediates predicted by the pathway models.

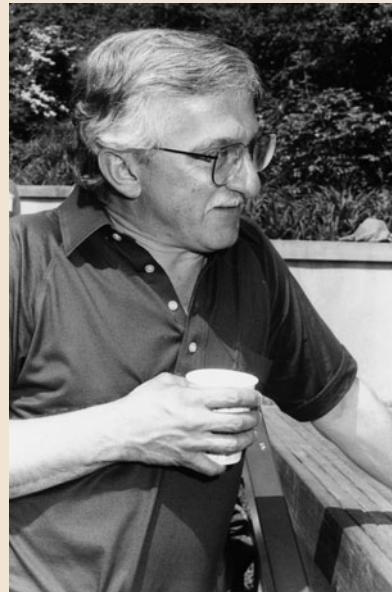
Finally, Chapter 12 brings us to two specialized kinds of recombination known as site-specific recombination and transposition. These processes lead to the vast accumulation of some sequences within the genomes of many organisms, including humans. We will discuss the molecular mechanisms and biological consequences of these forms of genetic exchange.

## PHOTOS FROM THE COLD SPRING HARBOR LABORATORY ARCHIVES

---



**Reiji Okazaki, 1968 Symposium on Replication of DNA in Microorganisms.** Okazaki had at this time just shown how, during DNA replication, one of the new strands is synthesized in short fragments that are only later joined together. The existence of these “Okazaki fragments” explained how an enzyme that synthesizes DNA in only one direction can nevertheless make two strands of opposite polarity simultaneously (Chapter 9).



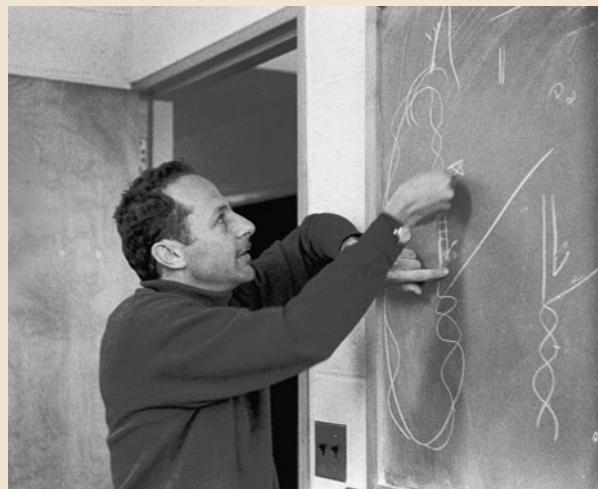
**Paul Modrich, 1993 Symposium on DNA and Chromosomes.** A pioneer in the DNA repair field (Chapter 10), Modrich worked out much of the mechanistic basis of mismatch repair.



**Carol Greider, Titia de Lange, and Elizabeth Blackburn, 2001 Telomeres Meetings.** Blackburn discovered the repeated sequences characteristic of telomeres at the ends of chromosomes. Later, while a graduate student in Blackburn's lab, Greider discovered telomerase, the enzyme that maintains the telomeres (Chapter 9). Shown between them here is de Lange, whose work focuses on proteins that bind to and protect telomeres within the cell. Blackburn and Greider, together with Jack Szostak, won the 2009 Nobel Prize in Physiology or Medicine.



**Arthur Kornberg, 1978 Symposium on DNA: Replication and Recombination.** Kornberg's extensive contributions to the study of DNA replication (Chapter 9) began with purifying the first enzyme that could synthesize DNA, a DNA polymerase from *Escherichia coli*. His experiments showed that a DNA template was required for new DNA synthesis, confirming a prediction of the model for DNA replication proposed by James Watson and Francis Crick. For this work Kornberg shared, with Severo Ochoa, the 1959 Nobel Prize in Physiology or Medicine.



**Matthew Meselson, 1968 Symposium on Replication of DNA in Microorganisms.** Meselson was Stahl's partner in the experiment showing that DNA replication is semiconservative (see photo on next page, and Chapter 2). Meselson later made major contributions to a number of fields, including purification of the first restriction enzyme, published the year this photo was taken. Furthermore, he is widely known for his work toward preventing the production and use of chemical and biological weapons.



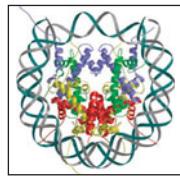
**Barbara McClintock and Robin Holliday, 1984 Symposium on Recombination at the DNA Level.** McClintock proposed the existence of transposons to account for the results of her genetic studies with maize, carried out in the 1940s (Chapter 12); the Nobel Prize in Physiology or Medicine in recognition of this work came more than 30 years later, in 1983. Holliday proposed the fundamental model of homologous recombination that bears his name (Chapter 11).



**Franklin Stahl and Max Delbrück, 1958 Symposium on Exchange of Genetic Material: Mechanism and Consequences.** Stahl, together with Matt Meselson (see photo on previous page), demonstrated that DNA is replicated by a semiconservative mechanism. This was once famously called “the most beautiful experiment in biology” (Chapter 2). Stahl subsequently contributed much to our understanding of homologous recombination (Chapter 11). Delbrück was the influential cofounder of the so-called “Phage Group”—a group of scientists who spent their summers at Cold Spring Harbor Laboratory and developed bacteriophage as the first model system of molecular biology (Appendix 1).

*This page intentionally left blank*

CHAPTER 8



# Genome Structure, Chromatin, and the Nucleosome

## OUTLINE

In Chapter 4, we considered the structure of DNA in isolation. Within the cell, however, DNA is associated with proteins, and each DNA molecule and its associated protein is called a **chromosome**. This organization holds true for prokaryotic and eukaryotic cells and even for viruses. Packaging of the DNA into chromosomes serves several important functions. First, the chromosome is a compact form of the DNA that readily fits inside the cell. Second, packaging the DNA into chromosomes serves to protect the DNA from damage. Completely naked DNA molecules are relatively unstable in cells. In contrast, chromosomal DNA is extremely stable. Third, only DNA packaged into a chromosome can be transmitted efficiently to both daughter cells when a cell divides. Finally, the chromosome confers an overall organization to each molecule of DNA. This organization regulates the accessibility of the DNA and, therefore, all of the events in the cell that involve DNA.

Half of the molecular mass of a eukaryotic chromosome is protein. In eukaryotic cells, a given region of DNA with its associated proteins is called **chromatin**, and the majority of the associated proteins are small, basic proteins called **histones**. Although not nearly as abundant, other proteins, referred to as the **nonhistone proteins**, are also associated with eukaryotic chromosomes. These proteins include the numerous DNA-binding proteins that regulate the replication, repair, recombination, and transcription of cellular DNA. Each of these topics is discussed in more detail in the next five chapters.

The protein component of chromatin performs another essential function: compacting the DNA. The following calculation makes the importance of this function clear. A human cell contains  $3 \times 10^9$  bp per **haploid set** of chromosomes. As we learned in Chapter 4, the average thickness of each base pair (the “rise”) is 3.4 Å. Therefore, if the DNA molecules in a haploid set of chromosomes were laid out end to end, the total length of DNA would be  $\sim 10^{10}$  Å, or 1 m! For a diploid cell (as human cells typically are), this length is doubled to 2 m. Because the diameter of a typical human cell nucleus is only 10–15 µm, it is obvious that the DNA must be compacted by many orders of magnitude to fit in such a small space. How is this achieved?

- Genome Sequence and Chromosome Diversity, 200
  - Chromosome Duplication and Segregation, 208
  - The Nucleosome, 220
  - Higher-Order Chromatin Structure, 229
  - Regulation of Chromatin Structure, 236
  - Nucleosome Assembly, 249
- Visit Web Content for Structural Tutorials and Interactive Animations

Most compaction in human cells (and all other eukaryotic cells) is the result of the regular association of DNA with histones to form structures called **nucleosomes**. The formation of nucleosomes is the first step in a process that allows the eukaryotic DNA to be folded into much more compact structures that reduce the linear length by as much as 10,000-fold. But compacting the DNA does not come without cost. Association of the DNA with histones and other packaging proteins limits the accessibility of the DNA. This reduced accessibility can interfere with proteins that direct the replication, repair, recombination, and—perhaps most significantly—transcription of the DNA. Eukaryotic cells exploit the inhibitory properties of chromatin to regulate gene expression and many other events involving DNA. Alterations to individual nucleosomes allow specific regions of the chromosomal DNA to interact with other proteins. These alterations are mediated by enzymes that modify and move nucleosomes. These processes are both dynamic and local, allowing enzymes and regulatory proteins access to different regions of the chromosome at different times. Therefore, understanding the structure of nucleosomes and the regulation of their association with DNA is critical to understanding the regulation of most events involving DNA in eukaryotic cells.

Although prokaryotic cells typically have smaller genomes, the need to compact their DNA is still substantial. *Escherichia coli* must pack its ~1-mm chromosome into a cell that is only 1  $\mu\text{m}$  in length. It is less clear how prokaryotic DNA is compacted. Bacteria have no histones or nucleosomes, for example, but they do have other small basic proteins that may serve similar functions. In this chapter, we focus on the better-understood chromosomes and chromatin of eukaryotic cells. We first consider the underlying DNA sequences of chromosomes from different organisms, focusing in particular on the change in protein-coding content. We then discuss the overall mechanisms that ensure that chromosomes are accurately transmitted as cells divide. The remainder of the chapter focuses on the structure and regulation of eukaryotic chromatin and its fundamental building block, the nucleosome.

## GENOME SEQUENCE AND CHROMOSOME DIVERSITY

---

Before we discuss the structure of chromosomes in detail, it is important to understand the features of the DNA molecules that are their foundation. The sequencing of the genomes of thousands of organisms has provided a wealth of information concerning the makeup of chromosomal DNAs and how their characteristics have changed as organisms have increased in complexity.

### Chromosomes Can Be Circular or Linear

The traditional view is that prokaryotic cells have a single circular chromosome and eukaryotic cells have multiple linear chromosomes (Table 8-1). As more prokaryotic organisms have been studied, however, this view has been challenged. Although the most studied prokaryotes (such as *E. coli* and *Bacillus subtilis*) do, indeed, have one circular chromosome, there are now numerous examples of prokaryotic cells that have multiple chromosomes, linear chromosomes, or even both. In contrast, all eukaryotic cells have multiple linear chromosomes. Depending on the eukaryotic organism, the number of chromosomes typically varies from two to less than 50, but in rare instances can reach thousands (e.g., in the macronucleus of the protozoa *Tetrahymena*) (see Table 8-1).

**TABLE 8-1** Variation in Chromosome Makeup in Different Organisms

Species	Number of Chromosomes	Chromosome Copy Number	Form of Chromosome(s)	Genome Size (Mb)
<b>Prokaryotes</b>				
<i>Mycoplasma genitalium</i>	1	1	Circular	0.58
<i>Escherichia coli</i> K-12	1	1	Circular	4.6
<i>Agrobacterium tumefaciens</i>	4	1	3 circular, 1 linear	5.67
<i>Sinorhizobium meliloti</i>	3	1	Circular	6.7
<b>Eukaryotes</b>				
<i>Saccharomyces cerevisiae</i> (budding yeast)	16	1 or 2	Linear	12.1
<i>Schizosaccharomyces pombe</i> (fission yeast)	3	1 or 2	Linear	12.5
<i>Caenorhabditis elegans</i> (roundworm)	6	2	Linear	97
<i>Arabidopsis thaliana</i> (weed)	5	2	Linear	125
<i>Drosophila melanogaster</i> (fruit fly)	4	2	Linear	180
<i>Tetrahymena thermophilus</i> (protozoa)				
Micronucleus	5	2	Linear	125
Macronucleus	225	10–10,000	Linear	
<i>Fugu rubripes</i> (fish)	22	2	Linear	393
<i>Mus musculus</i> (mouse)	19+X and Y	2	Linear	2600
<i>Homo sapiens</i>	22+X and Y	2	Linear	3200

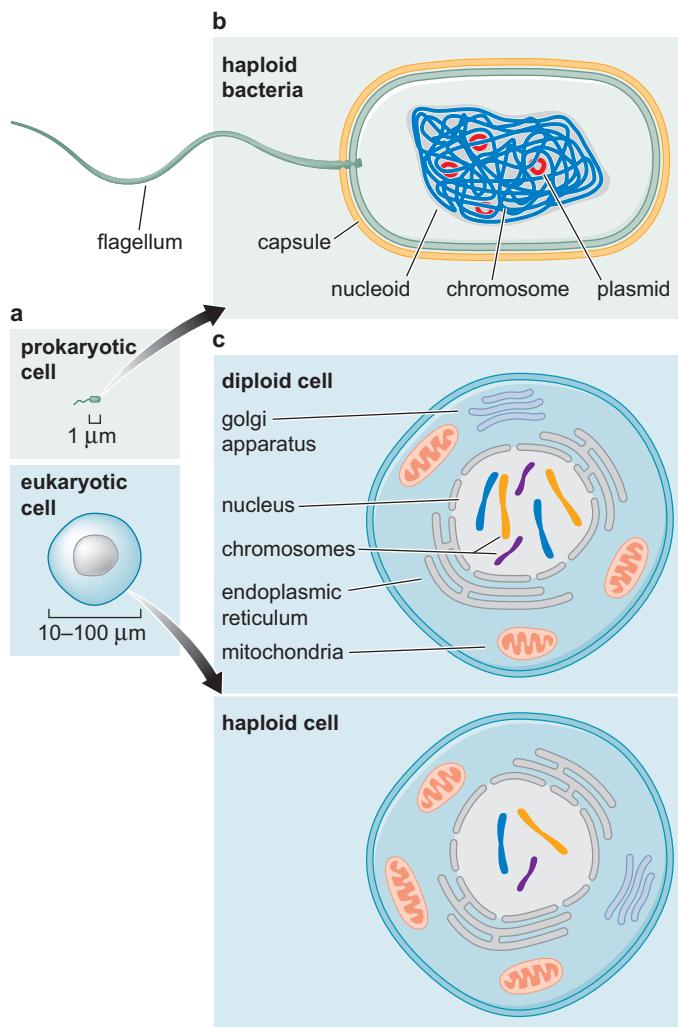
Circular and linear chromosomes each pose specific challenges that must be overcome for maintenance and replication of the genome. Circular chromosomes require topoisomerases to separate the daughter molecules after they are replicated. Without these enzymes, the two daughter molecules would remain interlocked, or catenated, with each other after replication (see Chapter 4, Fig. 4-23). In contrast, the ends of the linear eukaryotic chromosomes have to be protected from enzymes that normally degrade DNA ends and present a different set of difficulties during DNA replication, as we shall see in Chapter 9.

### Every Cell Maintains a Characteristic Number of Chromosomes

Prokaryotic cells typically have only one *complete* copy of their chromosome(s) that is packaged into a structure called the **nucleoid** (Fig. 8-1b). When prokaryotic cells are dividing rapidly, however, portions of the chromosome in the process of replicating are present in two and sometimes even four copies. Prokaryotes also frequently carry one or more smaller independent circular DNAs, called **plasmids**. Unlike the larger chromosomal DNA, plasmids typically are not essential for bacterial growth. Instead, they carry genes that confer desirable traits to the bacteria, such as antibiotic resistance. In addition, unlike chromosomal DNA, plasmids are often present in many complete copies per cell.

The majority of eukaryotic cells are **diploid**; that is, they contain two copies of each chromosome (see Fig. 8-1c). The two copies of a given chromosome are called **homologs**—one being derived from each parent. But not all cells in a eukaryotic organism are diploid; a subset of eukaryotic cells are either haploid or polyploid. **Haploid** cells contain a single copy of each chromosome and are involved in sexual reproduction (e.g., sperm and eggs are haploid cells). **Polyploid** cells have more than two copies of each chromosome. Indeed, some organisms maintain the majority of their adult cells in a polyploid state. In extreme cases, there can be hundreds or even thousands of

**FIGURE 8-1** Comparison of typical prokaryotic and eukaryotic cells. (a) The diameter of eukaryotic cells can vary between 10 and 100  $\mu\text{m}$ . The typical prokaryotic cell is  $\sim 1 \mu\text{m}$  long. (b) Prokaryotic chromosomal DNA is located in the nucleoid and occupies a substantial portion of the internal region of the cell. Unlike the eukaryotic nucleus, the nucleoid is not separated from the remainder of the cell by a membrane. Plasmid DNAs are shown in red. (c) Eukaryotic chromosomes are located in the membrane-bound nucleus. Haploid (one copy) and diploid (two copies) cells are distinguished by the number of copies of each chromosome present in the nucleus. (Adapted, with permission, from Brown T.A. 2002. *Genomes*, 2nd ed., p. 32, Fig. 2.1. © BIOS Scientific Publishers by permission of Taylor & Francis.)



copies of each chromosome. This type of global genome amplification allows a cell to generate larger amounts of RNA and, in turn, protein. For example, megakaryocytes are specialized polyploid cells (about 28 copies of each chromosome) that produce thousands of platelets, which lack chromosomes but are an essential component of human blood (there are about 200,000 platelets per milliliter of blood). By becoming polyploid, megakaryocytes can maintain the very high levels of metabolism necessary to produce large numbers of platelets. The segregation of such a large number of chromosomes is difficult; therefore, polyploid cells have almost always stopped dividing. No matter the number, eukaryotic chromosomes are always contained within a membrane-bound organelle called the **nucleus** (see Fig. 8-1c).

### Genome Size Is Related to the Complexity of the Organism

Genome size (the length of DNA associated with one haploid complement of chromosomes) varies substantially between different organisms (Table 8-2). Because more genes are required to direct the formation of more complex organisms (at least when comparing bacteria, single-cell eukaryotes, and multicellular eukaryotes; see Chapter 21), it is not surprising that genome size is roughly correlated with an organism's apparent complexity. Thus,

**TABLE 8-2** Comparison of the Gene Density in Different Organisms' Genomes

Species	Genome Size (Mb)	Approximate Number of Genes	Gene Density (genes/Mb)
<b>Prokaryotes (bacteria)</b>			
<i>Mycoplasma genitalium</i>	0.58	500	860
<i>Streptococcus pneumoniae</i>	2.2	2300	1060
<i>Escherichia coli K-12</i>	4.6	4400	950
<i>Agrobacterium tumefaciens</i>	5.7	5400	960
<i>Sinorhizobium meliloti</i>	6.7	6200	930
<b>Eukaryotes (animals)</b>			
Fungi			
<i>Saccharomyces cerevisiae</i>	12	5800	480
<i>Schizosaccharomyces pombe</i>	12	4900	410
Protozoa			
<i>Tetrahymena thermophila</i>	125	27,000	220
Invertebrates			
<i>Caenorhabditis elegans</i>	103	20,000	190
<i>Drosophila melanogaster</i>	180	14,700	82
<i>Ciona intestinalis</i>	160	16,000	100
<i>Locusta migratoria</i>	5000	nd	nd
Vertebrates			
<i>Fugu rubripes</i> (pufferfish)	393	22,000	56
<i>Homo sapiens</i>	3200	20,000	6.25
<i>Mus musculus</i> (mouse)	2600	22,000	8.5
Plants			
<i>Arabidopsis thaliana</i>	120	26,500	220
<i>Oryza sativa</i> (rice)	430	~45,000	~100
<i>Zea mays</i> (corn)	2200	>45,000	>20
<i>Triticum aestivum</i> (wheat)	16,000	nd	nd
<i>Fritillaria assyriaca</i> (tulip)	~120,000	nd	nd

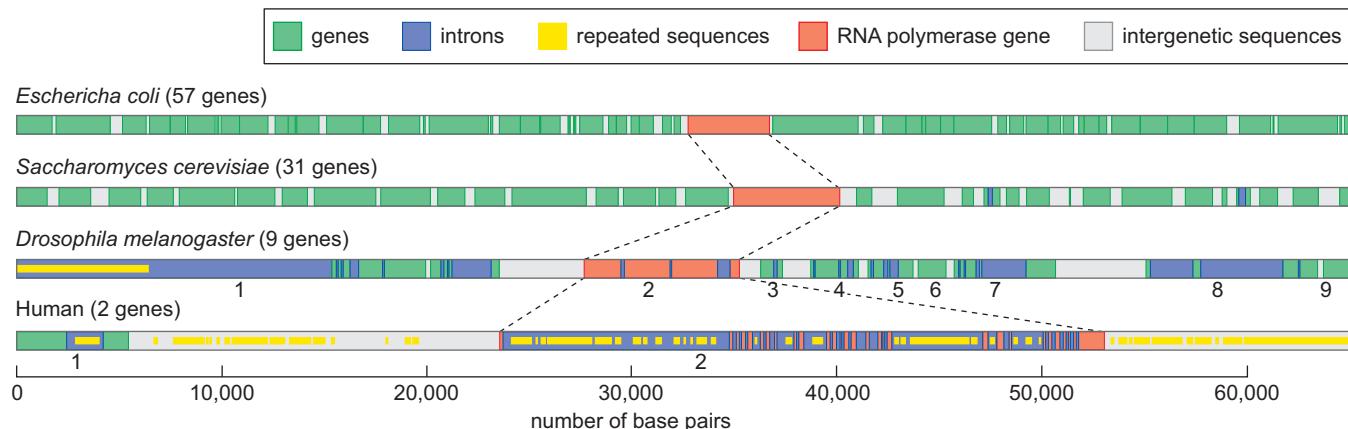
nd, Not determined.

prokaryotic cells typically have genomes of <10 Mb. The genomes of single-cell eukaryotes are typically <50 Mb, although some complex protozoans can have genomes >200 Mb. Multicellular organisms have even larger genomes that can reach sizes >100,000 Mb.

Although there is a rough correlation between genome size and organism complexity, this relationship is far from perfect. Many organisms of apparently similar complexities have very different genome sizes: a fruit fly has a genome about 25 times smaller than that of a locust, and the rice genome is about 40 times smaller than that of wheat (see Table 8-2). These examples point out that the number of genes, rather than genome size, is more closely related to organism complexity. This becomes clear when we examine the relative gene densities of different genomes.

### The *E. coli* Genome Is Composed Almost Entirely of Genes

The great majority of the single chromosome of the bacteria *E. coli* encodes proteins or structural RNAs (Fig. 8-2). The majority of the non-coding sequences are dedicated to regulating gene transcription (as we shall see in Chapter 18). Because a single site of transcription initiation is often used to control the expression of several genes, even these regulatory regions are kept to a minimum in the genome. One critical element of the *E. coli*



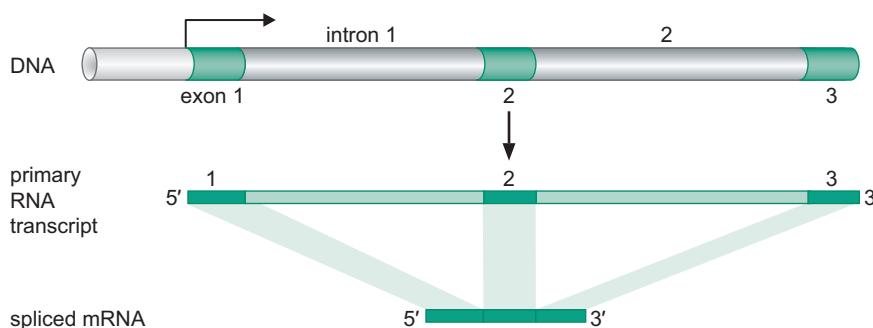
**FIGURE 8-2** Comparison of chromosomal gene density for different organisms. A 65-kb region of DNA including the gene for the largest subunit of RNA polymerase (RNA polymerase II for the eukaryotic cells) is illustrated for each organism. In each case, the RNA polymerase encoding DNA is indicated in red. Coding DNA for other genes is indicated in green, intron DNA in purple, repeated DNA in yellow, and unique intergenic DNA in gray. Note how the number of genes included in the 65-kb region decreases as organism complexity increases.

genome is not a gene or a sequence that regulates gene expression. Instead, the *E. coli* origin of replication is dedicated to directing the assembly of the replication machinery (as we shall discuss in Chapter 9). Despite its important role, this region is still very small, occupying only a few hundred base pairs of the 4.6-Mb *E. coli* genome.

### More Complex Organisms Have Decreased Gene Density

What explains the dramatically different genome sizes of organisms of apparently similar complexity (such as the fruit fly and locust)? The differences are largely related to gene density. One simple measure of gene density is the average number of genes per megabase of genomic DNA. For example, if an organism has 5000 genes and a genome size of 50 Mb, then the gene density for that organism is 100 genes/Mb. When the gene densities of different organisms are compared, it becomes clear that different organisms use the gene-encoding potential of DNA with varying efficiencies. There is a roughly inverse correlation between organism complexity and gene density—the less complex the organism, the higher the gene density. For example, the highest gene densities are found for viruses that, in some instances, use both strands of the DNA to encode overlapping genes. Although overlapping genes are rare, bacterial gene density is consistently near 1000 genes/Mb.

Gene density in eukaryotic organisms is consistently lower and more variable than in their prokaryotic counterparts (see Table 8-2). Among eukaryotes, there is a general trend for gene density to decrease with increasing organism complexity. The simple unicellular eukaryote *Saccharomyces cerevisiae* has a gene density close to that of prokaryotes (~500 genes/Mb). In contrast, the human genome is estimated to have a 50-fold lower gene density. In Figure 8-2, the amount of DNA sequence devoted to the expression of a related gene conserved across all organisms (the large subunit of RNA polymerase) is compared. As you can see, there is a vast difference in the amount of DNA devoted to the expression of one gene despite the very similar size of the protein encoded. What is responsible for this reduction in gene density?



**FIGURE 8-3 Schematic of RNA splicing.** Transcription of pre-mRNA is initiated at the arrow shown above exon 1. This primary transcript is then processed (by splicing) to remove non-coding introns to produce messenger RNA (mRNA).

### Genes Make Up Only a Small Proportion of the Eukaryotic Chromosomal DNA

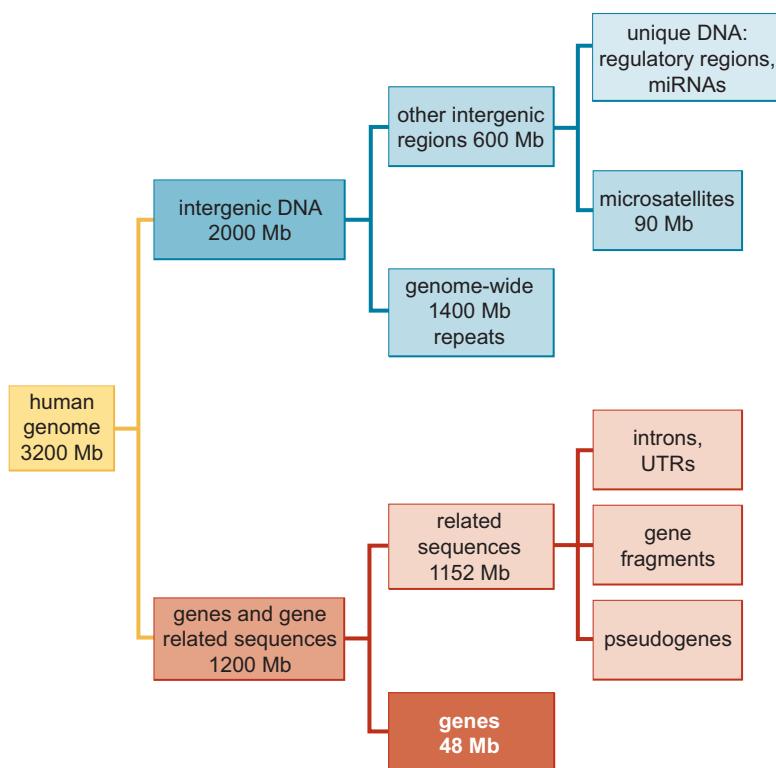
Two factors contribute to the decreased gene density observed in eukaryotic cells: increases in gene size and increases in the DNA between genes, called **intergenic sequences**. The major reason that gene size is larger in more complex organisms is not that the average protein is bigger or that more DNA is required to encode the same protein. Instead, protein-encoding genes in eukaryotes frequently have discontinuous protein-coding regions. These interspersed non-protein-coding regions, called **introns**, are removed from the RNA after transcription in a process called **RNA splicing** (Fig. 8-3); we shall consider RNA splicing in detail in Chapter 14. The presence of introns can dramatically increase the length of DNA required to encode a gene (Table 8-3). For example, the average transcribed region of a human gene is ~27 kb (this should not be confused with the gene density), whereas the average protein-coding region of a human gene is 1.3 kb. A simple calculation reveals that only 5% of the average human protein-encoding gene directly encodes the desired protein. The remaining 95% is made up of introns. Consistent with their higher gene density, simpler eukaryotes have far fewer introns. For example, in the yeast *S. cerevisiae*, only 3.5% of genes have introns, none of which is >1 kb (see Table 8-3).

**TABLE 8-3 Contribution of Introns and Repeated Sequences to Different Genomes**

Species	Gene Density (genes/Mb)	Average Number of Introns per Gene	% of Repetitive DNA
<b>Prokaryotes (bacteria)</b>			
<i>Escherichia coli</i> K-12	950	0	<1
<b>Eukaryotes (animals)</b>			
Fungi			
<i>Saccharomyces cerevisiae</i>	480	0.04	3.4
Invertebrates			
<i>Caenorhabditis elegans</i>	190	5	6.3
<i>Drosophila melanogaster</i>	82	3	12
Vertebrates			
<i>Fugu rubripes</i>	56	5	2.7
<i>Homo sapiens</i>	6.25	6	46
Plants			
<i>Arabidopsis thaliana</i>	220	3	nd
<i>Oryza sativa</i> (rice)	~100	nd	42

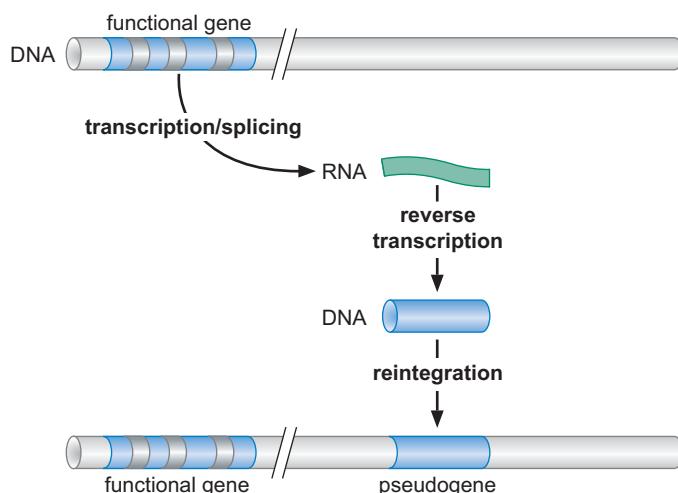
nd, Not determined.

**FIGURE 8-4** Organization and content of the human genome. The human genome is composed of many different types of DNA sequences, the majority of which do not encode proteins. Shown are the distribution and amount of each of the various types of sequences. (Adapted from Brown T.A. 2002. *Genomes*, 2nd ed., p. 23, Box 1.4. © BIOS Scientific Publishers by permission of Taylor & Francis.)



An explosion in the amount of intergenic sequences in more complex organisms is responsible for the remaining decreases in gene density. Intergenic DNA is the portion of a genome that does not encode proteins or structural RNAs. More than 60% of the human genome is composed of intergenic sequences, and much of this DNA has no known function (Fig. 8-4). There are two kinds of intergenic DNAs: unique and repeated. About one-quarter of the intergenic DNA is unique. One contributor to an increase in unique intergenic sequences is an increase in regions of the DNA that are required to direct and regulate transcription, called **regulatory sequences**. As organisms become more complex and encode for more genes, the regulatory sequences required to coordinate gene expression also grow in complexity and size. The unique regions of the human intergenic DNA also include many apparently nonfunctional relicts, including nonfunctional mutant genes, gene fragments, and pseudogenes. The mutant genes and gene fragments arise from simple random mutagenesis or mistakes in DNA recombination. Pseudogenes arise from the action of an enzyme called **reverse transcriptase** (Fig. 8-5; see Chapter 12). This enzyme copies RNA into double-stranded DNA (referred to as **copy DNA** or **cDNA**). Reverse transcriptase is only expressed by certain types of viruses that require this enzyme to reproduce. But, as a side effect of infection by such a virus, cellular mRNAs can be copied into DNA, and the resulting DNA fragments reintegrate into the genome at a low rate. These copies are not expressed, however, because they lack the correct regulatory sequences to direct their expression (such sequences are generally not part of a gene's RNA product; see Chapter 13).

Finally, it is clear that there are likely to be functions of the unique intergenic regions in eukaryotic cells that are not yet understood. One example is the recent identification of microRNAs, commonly referred to as **miRNAs**. These small structural RNAs act to regulate the expression of other genes



**FIGURE 8-5** Processed pseudogenes arise from integration of reverse-transcribed messenger RNAs. When reverse transcriptase is present in a cell, messenger RNA (mRNA) molecules can be copied into double-stranded DNA. In rare instances, these DNA molecules can integrate into the genome creating pseudogenes. Because introns are rapidly removed from newly transcribed RNAs, these pseudogenes have the common characteristic of lacking introns. This distinguishes the pseudogene from the copy of the gene from which it was derived. In addition, pseudogenes lack the appropriate promoter sequences to direct their transcription because these are not part of the mRNA from which they are derived.

by altering either the stability of the product mRNA or its ability to be translated (we consider gene regulation by small RNAs in Chapter 20). Because these sequences are still being discovered, they are not included in Table 8-2; however, it has been estimated that human cells may have more than 500 miRNA genes. Similarly, thousands of long-intervening non-coding RNAs (lincRNAs) have also been identified. Although these RNAs do not encode for any proteins of significant length, recent studies suggest that they act to regulate gene expression both positively and negatively in a manner that has yet to be fully understood. Another function likely to be encoded in the unique intergenic regions is origins of replication, which have yet to be identified in most eukaryotic organisms.

### The Majority of Human Intergenic Sequences Are Composed of Repetitive DNA

Almost half of the human genome is composed of DNA sequences that are repeated many times in the genome. There are two general classes of repeated DNA: microsatellite DNA and genome-wide repeats. **Microsatellite DNA** is composed of very short (<13 bp), tandemly repeated sequences. The most common microsatellite sequences are dinucleotide repeats (e.g., CACACACACACACACA). These repeats arise from difficulties in accurately duplicating the DNA and represent nearly 3% of the human genome.

**Genome-wide repeats** are much larger than their microsatellite counterparts. Each genome-wide repeat unit is >100 bp in length and many are >1 kb. These sequences can be found either as single copies dispersed throughout the genome or as closely spaced clusters. Although there are numerous classes of such repeats, their common feature is that all are forms of **transposable elements**.

Transposable elements are sequences that can “move” from one place in the genome to another. During **transposition**, as this movement is called, the element moves to a new position in the genome, often leaving the original copy behind. Thus, these sequences can multiply and accumulate throughout the genome. Movement of transposable elements is a relatively rare event in human cells. Nevertheless, over long periods of evolutionary time, these elements have been so successful at propagating copies of themselves that they now comprise ~45% of the human genome. In Chapter 12,

we consider the mechanism by which transposable elements move around the genome and how their movement is controlled to prevent integration into genes.

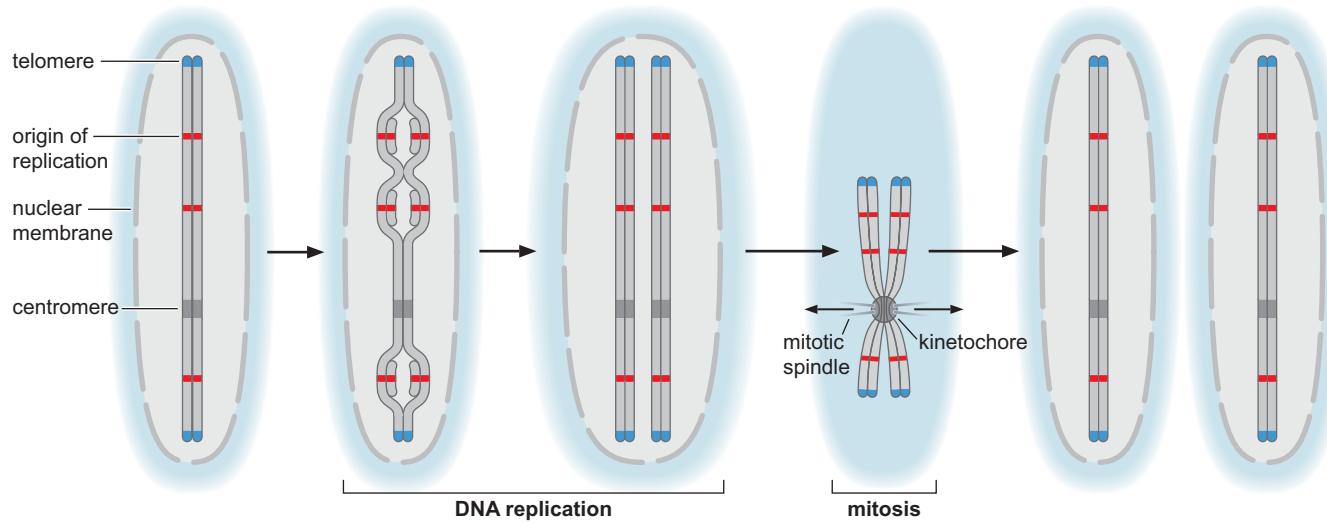
Although we have discussed the nature of the intergenic sequence in the context of the human genome, many of the same features are found in other organisms. For example, comparison of the sequences of several plants with very large genomes (such as maize) indicates that transposable elements are likely to comprise an even larger percentage of these genomes. Similarly, even in the compact genomes of *E. coli* and *S. cerevisiae*, there are examples of transposable elements and microsatellite repeats (see Fig. 8-2). The difference is that these elements have been less successful at occupying the genomes of these simpler organisms. This lack of success is likely due to a combination of inefficient duplication and more efficient elimination (either by repair events or through selection against organisms in which duplication has occurred).

Although it is tempting to refer to repeated DNA as “junk DNA,” the stable maintenance of these sequences over thousands of generations suggests that intergenic DNA confers a positive value (or selective advantage) to the host organism.

## CHROMOSOME DUPLICATION AND SEGREGATION

### Eukaryotic Chromosomes Require Centromeres, Telomeres, and Origins of Replication to Be Maintained during Cell Division

There are several important DNA elements in eukaryotic chromosomes that are not genes and that are not involved in regulating the expression of genes (Fig. 8-6). These elements include origins of replication that direct the



**FIGURE 8-6** Centromeres, origins of the replication, and telomeres are required for eukaryotic chromosome maintenance. Each eukaryotic chromosome includes two telomeres, one centromere, and many origins of replication. Telomeres are located at both ends of each chromosome. Unlike telomeres, the single centromere found on each chromosome is not in a defined position. Some centromeres are near the middle of the chromosome, and others are closer to a telomere. Origins of replication are located throughout the length of each chromosome (e.g., approximately every 30 kb in the budding yeast *S. cerevisiae*).

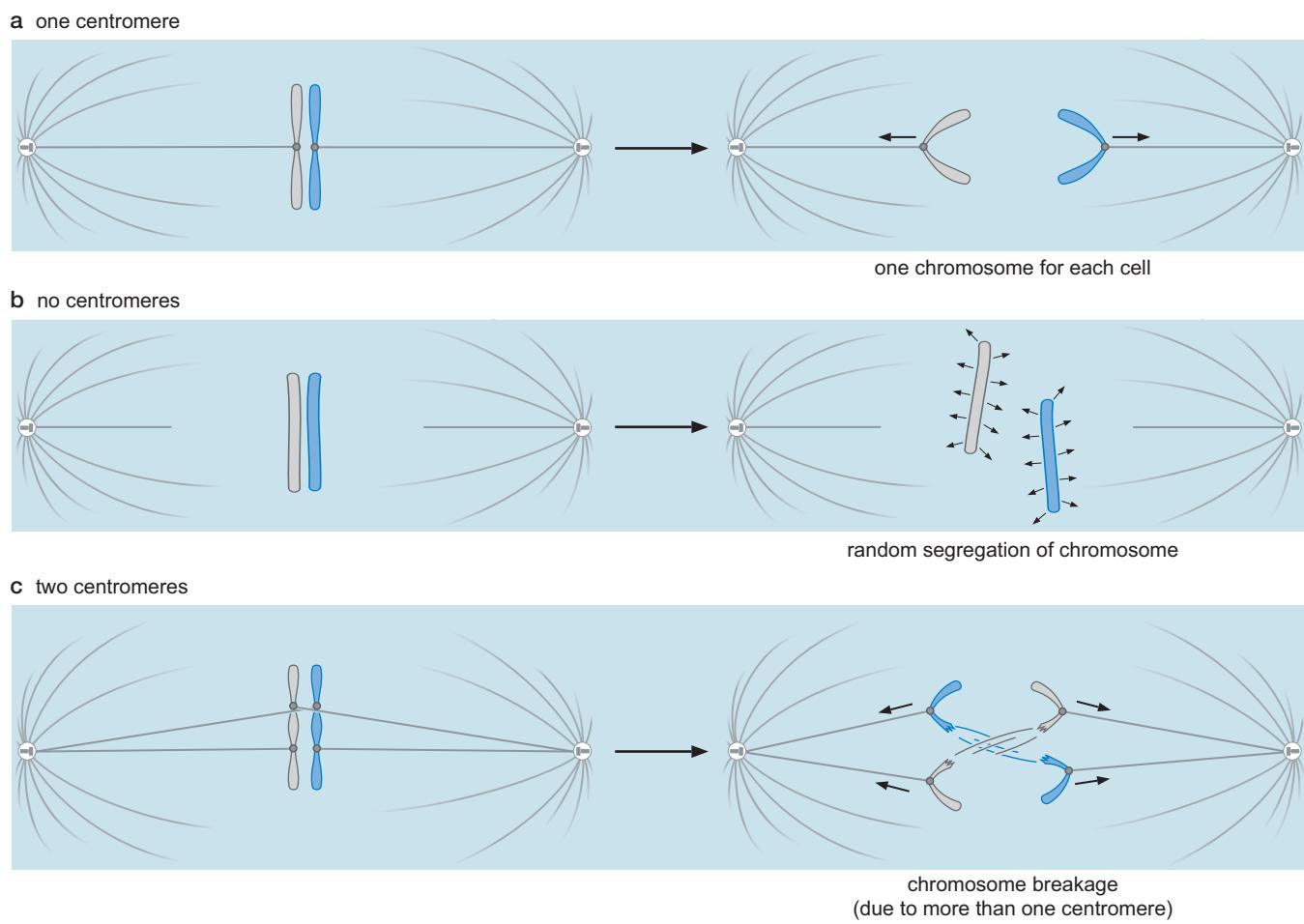
duplication of the chromosomal DNA, centromeres that act as “handles” for the movement of replicated chromosomes into daughter cells, and telomeres that protect and replicate the ends of linear chromosomes. All of these features are critical for the proper duplication and segregation of the chromosomes during cell division. We now look at each of these elements in more detail.

**Origins of replication** are the sites at which the DNA replication machinery assembles and replication is initiated. They are typically found some 30–40 kb apart throughout the length of each eukaryotic chromosome. Prokaryotic chromosomes also require origins of replication. Unlike their eukaryotic counterparts, prokaryotic chromosomes typically have only a single site of replication initiation. In general, origins of replication are found in non-coding regions. The DNA sequences that are recognized as origins of replication are discussed in detail in Chapter 9.

**Centromeres** are required for the correct segregation of the chromosomes after DNA replication. The two copies of each replicated chromosome are called **sister chromosomes**, and during cell division they must be separated with one copy going to each of the two daughter cells. Like origins of replication, centromeres direct the formation of an elaborate protein complex called a **kinetochore**. The kinetochore assembles at each centromere DNA, and before chromosome segregation, the kinetochore binds to protein filaments called **microtubules** that eventually pull the sister chromosomes away from each other and into the two daughter cells. In contrast to the many origins of replication found on each eukaryotic chromosome, it is critical that each chromosome include *only one* centromere (Fig. 8-7a). In the absence of a centromere, the replicated chromosomes segregate randomly, resulting in daughter cells that either have lost a chromosome or have two copies of a chromosome (Fig. 8-7b). The presence of more than one centromere on each chromosome is equally disastrous. If the associated kinetochores are attached to filaments pulling in opposite directions, this can lead to chromosome breakage (Fig. 8-7c). Centromeres vary greatly in size. In the yeast *S. cerevisiae*, centromeres are composed of unique sequences that are <200 bp in length. In contrast, in the majority of eukaryotes, centromeres are >40 kb and are composed of largely repetitive DNA sequences (Fig. 8-8).

**Telomeres** are located at the two ends of a linear chromosome. Like origins of replication and centromeres, telomeres are bound by a number of proteins. In this case, the proteins perform two important functions. First, telomeric proteins distinguish the natural ends of the chromosome from sites of chromosome breakage and other DNA breaks in the cell. Ordinarily, DNA ends are sites of frequent recombination and DNA degradation. The proteins that assemble at telomeres form a structure that is resistant to both of these events. Second, telomeres act as specialized origins of replication that allow the cell to replicate the ends of the chromosomes. For reasons described in detail in Chapter 9, the standard DNA replication machinery cannot completely replicate the ends of a linear chromosome. Telomeres facilitate end replication through the recruitment of an unusual DNA polymerase called **telomerase**.

In contrast to most of the chromosome, a portion of the telomere is maintained in a single-stranded form (Fig. 8-9). Most telomeres have a simple repeating sequence that varies from organism to organism. This repeat is typically composed of a short TG-rich repeat. For example, human telomeres have the repeating sequence of 5'-TTAGGG-3'. As we shall see in Chapter 9, the repetitive nature of telomeres is a consequence of their unique method of replication.

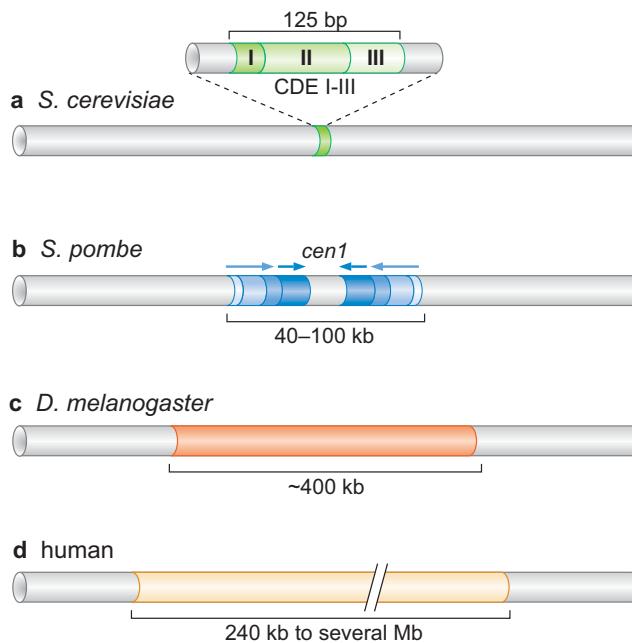


**FIGURE 8-7** More or less than one centromere per chromosome leads to chromosome loss or breakage. (a) Normal chromosomes have one centromere. After replication of a chromosome, each copy of the centromere directs the formation of a kinetochore. These two kinetochores then bind to opposite poles of the mitotic spindle and are pulled to the opposite sides of the cell before cell division. (b) Chromosomes lacking centromeres are rapidly lost from cells. In the absence of the centromere, the chromosomes do not attach to the spindle and are randomly distributed to the two daughter cells. This leads to frequent events in which one daughter gets two copies of a chromosome and the other daughter cell is missing the same chromosome. (c) Chromosomes with two or more centromeres are frequently broken during segregation. If a chromosome has more than one centromere, it can be bound simultaneously to both poles of the mitotic spindle. When segregation is initiated, the opposing forces of the mitotic spindle break chromosomes attached to both poles.

### Eukaryotic Chromosome Duplication and Segregation Occur in Separate Phases of the Cell Cycle

During cell division, the chromosomes must be duplicated and segregated into the daughter cells. In bacterial cells, these events occur simultaneously; that is, as the DNA is replicated, the resulting two copies are separated into opposite sides of the cell. Although it is clear that these events are tightly coordinated in bacteria, how this coordination is achieved is poorly understood. In contrast, eukaryotic cells duplicate and segregate their chromosomes at distinct times during cell division. We focus on these events for the remainder of our discussion of chromosomes.

The events required for a single round of cell division are collectively known as the **cell cycle** (see Interactive Animation 8-1). Most eukaryotic



**FIGURE 8-8** Centromere size and composition vary dramatically among different organisms. *Saccharomyces cerevisiae* centromeres are small and composed of nonrepetitive sequences. In contrast, the centromeres of other organisms such as the fruit fly, *Drosophila melanogaster*, and the fission yeast, *Schizosaccharomyces pombe*, are much larger and are mostly composed of repetitive sequences. Only the central 4–7 kb of the *S. pombe* centromere is nonrepetitive, and the large majority of the *Drosophila* and human centromeres are repetitive DNA.

cell divisions maintain the number of chromosomes in the daughter cells that were present in the parental cell. This type of division is called **mitotic cell division**.

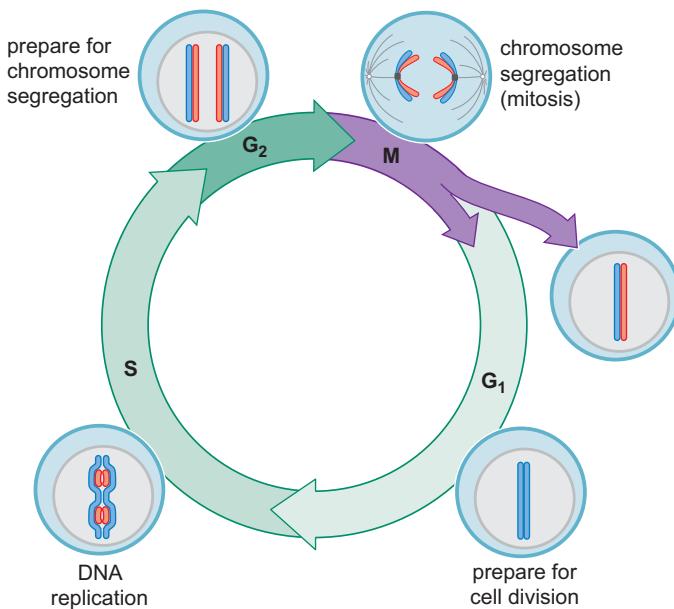
The mitotic cell cycle can be divided into four phases: G<sub>1</sub>, S, G<sub>2</sub>, and M (Fig. 8-10). Chromosome replication occurs during the **synthesis**, or **S phase**, of the cell cycle, resulting in the duplication of each chromosome (Fig. 8-11). Each chromosome of the duplicated pair is called a **chromatid**, and the two chromatids of a given pair are called **sister chromatids**. Sister chromatids are held together after duplication through a process called **sister-chromatid cohesion**, and this tethered state is maintained until the chromosomes segregate from one another. Sister-chromatid cohesion is mediated by a protein called **cohesin**, which we shall describe later.

Chromosome segregation occurs during **mitosis**, or the **M phase**, of the cell cycle. We consider the overall process of mitosis later, but first we focus on three key steps in the process (Fig. 8-12). First, each pair of sister chromatids is bound to a structure called the **mitotic spindle**. This structure is composed of long protein fibers called **microtubules** that are attached to one of the two **microtubule-organizing centers** (also called **centrosomes** in animal cells or **spindle pole bodies** in yeasts and other fungi). The microtubule-organizing centers are located on opposite sides of the cell, forming “poles” toward which the microtubules pull the chromatids. Attachment of the chromatids to the microtubules is mediated by the **kinetochore** assembled at each centromere (see Fig. 8-6). Second, the cohesion between the chromatids is dissolved by proteolysis of cohesin. Before



**FIGURE 8-9** Structure of a typical telomere. The repeated sequence (from human cells) is shown in a representative box. Note that the region of single-stranded DNA at the 3' end of the chromosome can be hundreds of bases long.

**FIGURE 8-10** Eukaryotic mitotic cell cycle. There are four stages of the eukaryotic cell cycle. Chromosomal replication occurs during S phase, and chromosome segregation occurs during M phase. The G<sub>1</sub> and G<sub>2</sub> gap phases allow the cell to prepare for the next event in the cell cycle. For example, many eukaryotic cells use the G<sub>1</sub> phase of the cell cycle to establish that the level of nutrients is sufficiently high to allow the completion of cell division.

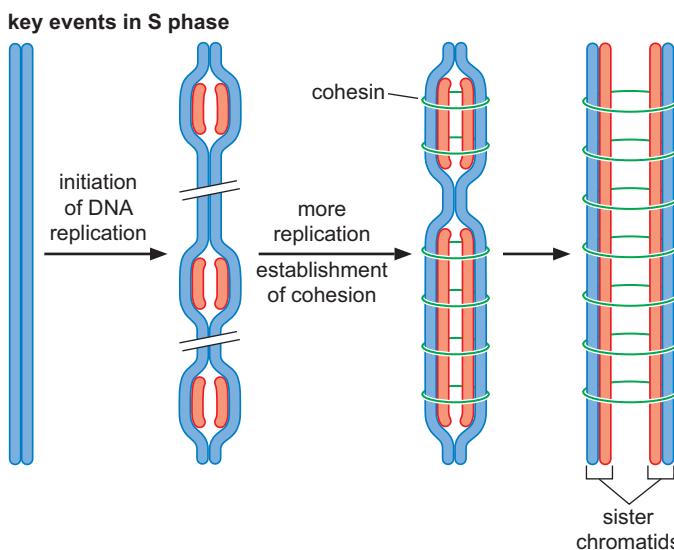


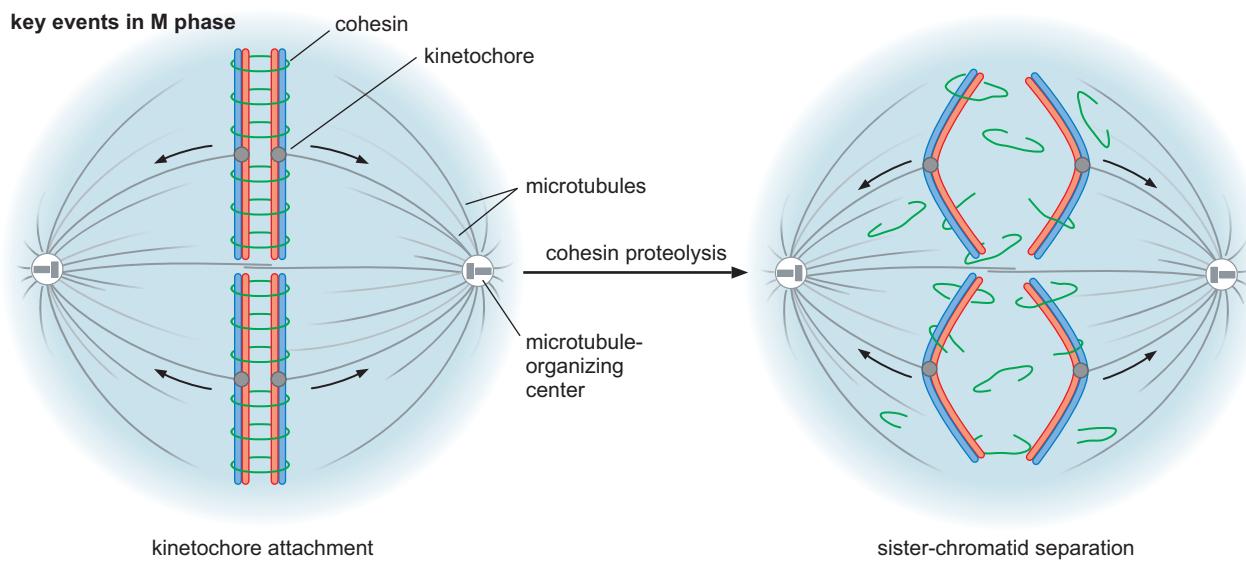
cohesion is dissolved, it resists the pulling forces of the mitotic spindle. After cohesion is dissolved, the third major event in mitosis can occur: **sister-chromatid separation**. In the absence of the counterbalancing force of chromatid cohesion, the chromatids are rapidly pulled toward opposite poles of the mitotic spindle. Thus, cohesion between the sister chromatids and attachment of sister-chromatid kinetochores to opposite poles of the mitotic spindle play opposing roles that must be carefully coordinated for chromosome segregation to occur properly.

### Chromosome Structure Changes as Eukaryotic Cells Divide

As chromosomes proceed through a round of cell division, their structure is altered numerous times; however, there are two main states for the chromosomes (Fig. 8-13). The chromosomes are in their most compact form as

**FIGURE 8-11** Events of S phase. Two major chromosomal events occur during S phase. DNA replication copies each chromosome completely, and shortly after replication has occurred, sister-chromatid cohesion is established by ring-shaped cohesin molecules, which are hypothesized to encircle the two copies of the recently replicated DNA. Each blue or red “tube” represents a single-stranded DNA molecule, with red DNA being newly synthesized.

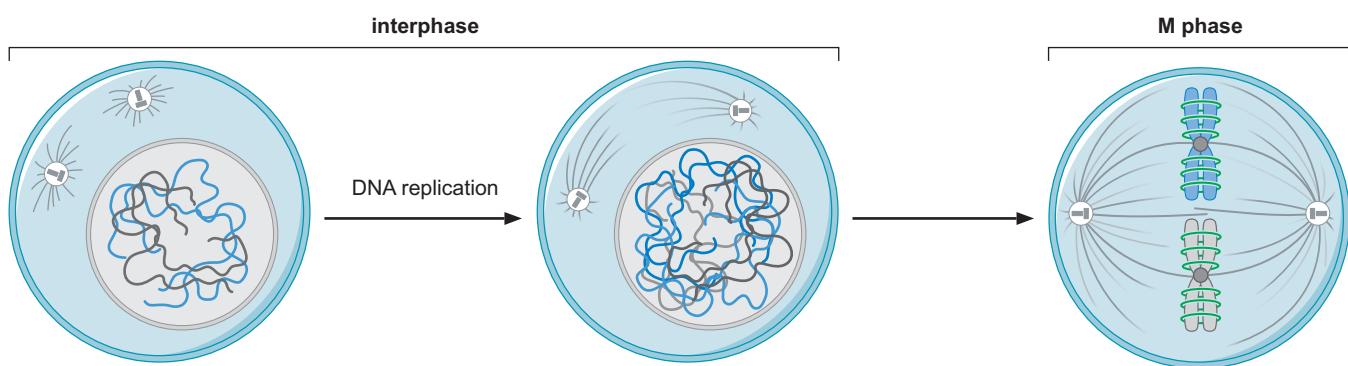




**FIGURE 8-12** Events of mitosis (M phase). Three major events occur during mitosis. First, the two kinetochores of each linked sister-chromatid pair attach to opposite poles of the mitotic spindle. Once all kinetochores are bound to opposite poles, sister-chromatid cohesion is eliminated by destroying the cohesin ring. Finally, after cohesion is eliminated, the sister chromatids are segregated to opposite poles of the mitotic spindle.

cells segregate their chromosomes. The process that results in this compact form is called **chromosome condensation**. In this condensed state, the chromosomes are disentangled from one another, greatly facilitating the segregation process.

During phases of the cell cycle when chromosome segregation is not occurring (collectively referred to as **interphase**), the chromosomes are significantly less compact. Indeed, at these stages of the cell cycle, the chromosomes are likely to be highly intertwined, resembling more of a plate (really a sphere) of spaghetti than the organized chromosomes seen during mitosis. Nevertheless, even during these stages, the structures of the chromosomes change. DNA replication requires the nearly complete disassembly and reassembly of the proteins associated with each chromosome. Sister-chromatid cohesion is established immediately after replication, linking the newly



**FIGURE 8-13** Changes in chromatin structure. Chromosomes are maximally condensed in M phase and decondensed throughout the rest of the cell cycle ( $G_1$ ,  $S$ , and  $G_2$  in mitotic cells). Together, these decondensed stages are referred to as interphase.

replicated chromatids to one another. As transcription of individual genes is turned on and off or up and down, there are associated changes in the structure of the chromosomes in those regions occurring throughout the cell cycle. Thus, the chromosome is a constantly changing structure that is more like an organelle than a simple string of DNA.

### Sister-Chromatid Cohesion and Chromosome Condensation Are Mediated by SMC Proteins

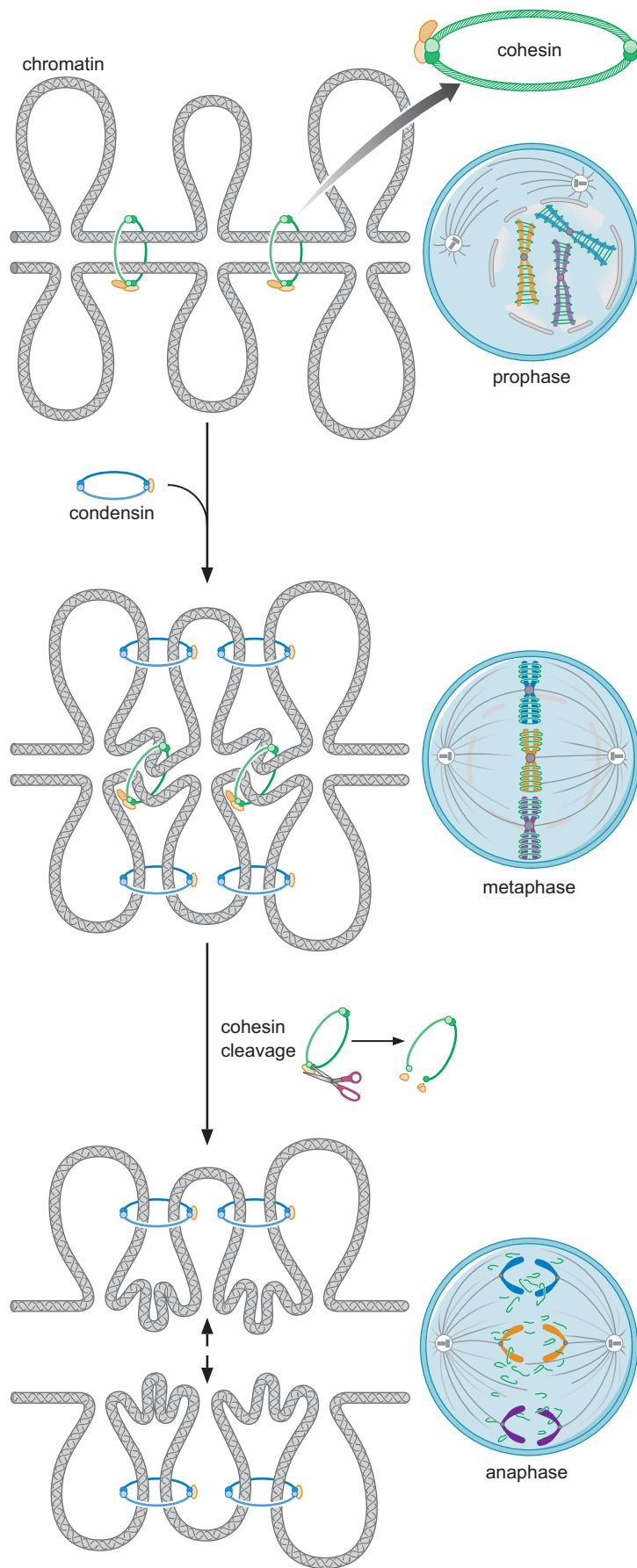
The key proteins that mediate sister-chromatid cohesion and chromosome condensation are related to one another. The structural maintenance of chromosome (SMC) proteins are extended proteins that form defined pairs by interacting through lengthy coiled-coil domains (see Chapter 6, Fig. 6-9). Together with non-SMC proteins, they form multiprotein complexes that act to link two DNA helices together. Cohesin is an SMC-protein-containing complex that, as we discussed above, is required to link the two daughter DNA duplexes (sister chromatids) together after DNA replication. It is this linkage that is the basis for sister-chromatid cohesion. The structure of cohesin is thought to be a large ring composed of two SMC proteins and two non-SMC proteins. Although the exact mechanism of sister-chromatid cohesion is still under investigation, a prominent model proposes that chromatid cohesion occurs as the result of both sister chromatids passing through the center of the cohesin protein ring (Fig. 8-14). In this model, proteolytic cleavage of the non-SMC subunit of cohesin results in the opening of the ring, loss of sister-chromatid cohesion, and the movement of the daughter chromosomes to opposite cell poles.

The chromosome condensation that accompanies chromosome segregation also requires a related SMC-containing complex called **condensin**. Condensin shares many of the features of the cohesin complex, suggesting that it too is a ring-shaped complex. If so, it may use its ring-like nature to induce chromosome condensation. For example, by linking different regions of the same chromosome together, condensin could reduce the overall linear length of the chromosome (Fig. 8-14).

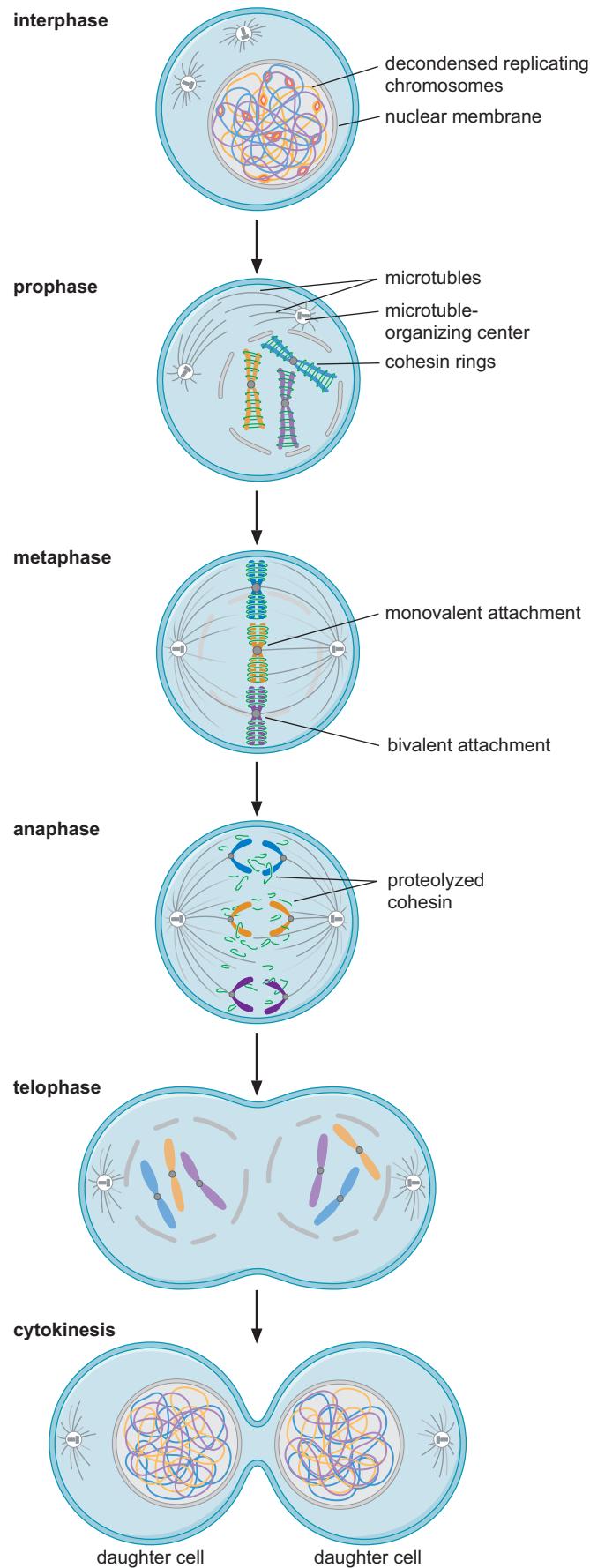
### Mitosis Maintains the Parental Chromosome Number

We now return to the overall process of mitosis. Mitosis occurs in several stages (Fig. 8-15). During **prophase**, the action of condensin and type II topoisomerases (to help to untangle the chromosomes) drives chromosomes to condense into the highly compact form required for segregation. At the end of prophase, in most cells the nuclear envelope breaks down and the cell enters metaphase.

During metaphase, the mitotic spindle forms and the kinetochores of sister chromatids attach to the microtubules. Proper chromatid attachment is only achieved when the two kinetochores of a sister-chromatid pair are attached to microtubules emanating from opposite microtubule-organizing centers. This type of attachment is called **bivalent attachment** (see Fig. 8-15) and results in the microtubules exerting tension on the chromatid pair by pulling the sisters in opposite directions. Attachment of both chromatids to microtubules emanating from the same microtubule-organizing center or attachment of only one chromatid of the pair, called **monovalent attachment**, does not result in tension. If bivalent attachment does not occur subsequently, monovalent attachment could lead to both copies of a chromosome moving into one daughter cell. The tension exerted by bivalent



**FIGURE 8-14** Model for the structure and function of cohesins and condensins. Cohesins and condensins are ring-shaped protein complexes that include two SMC proteins that play important roles in bringing distant or different regions of DNA together. The proposed ring-shaped structure of these proteins would allow a flexible but strong link between two regions of DNA. In this illustration, the SMC proteins are (green) cohesin or (blue) condensin. (Adapted, with permission, from Haering C.H. et al. 2002. *Mol. Cell* 9: 773–788, Fig. 8, p. 785. © Elsevier.)



**FIGURE 8-15** Mitosis in detail. Before mitosis, the chromosomes are in a decondensed state called interphase. During prophase, chromosomes are condensed and detangled in preparation for segregation, and the nuclear membrane surrounding the chromosomes breaks down in most eukaryotes. During metaphase, each sister-chromatid pair attaches to opposite poles of the mitotic spindle. Anaphase is initiated by the loss of sister-chromatid cohesion, resulting in the separation of sister chromatids. Telophase is distinguished by the loss of chromosome condensation and the re-formation of the nuclear membrane around the two populations of segregated chromosomes. Cytokinesis is the final event of the cell cycle during which the cellular membrane surrounding the two nuclei constricts and eventually completely separates into two daughter cells. All DNA molecules are double-stranded.

attachment is opposed by sister-chromatid cohesion and results in all of the chromosomes aligning in the middle of the cell between the two microtubule-organizing centers (this position is called the metaphase plate). Importantly, chromosome segregation starts only after all sister-chromatid pairs have achieved bivalent attachment.

Chromosome segregation is triggered by proteolytic destruction of the cohesin molecules, resulting in the loss of sister-chromatid cohesion. This loss occurs as cells enter **anaphase**, during which the sister chromatids separate and move to opposite sides of the cell. Once the two sisters are no longer held together, they cannot resist the outward pull of the microtubule spindle. Bivalent attachment ensures that the two members of a sister-chromatid pair are pulled toward opposite poles and each daughter cell receives one copy of each duplicated chromosome.

The final step of mitosis is **telophase**, during which the nuclear envelope re-forms around each set of segregated daughter chromosomes. At this point, cell division can be completed by physically separating the shared cytoplasm of the two presumptive cells in a process called **cytokinesis**.

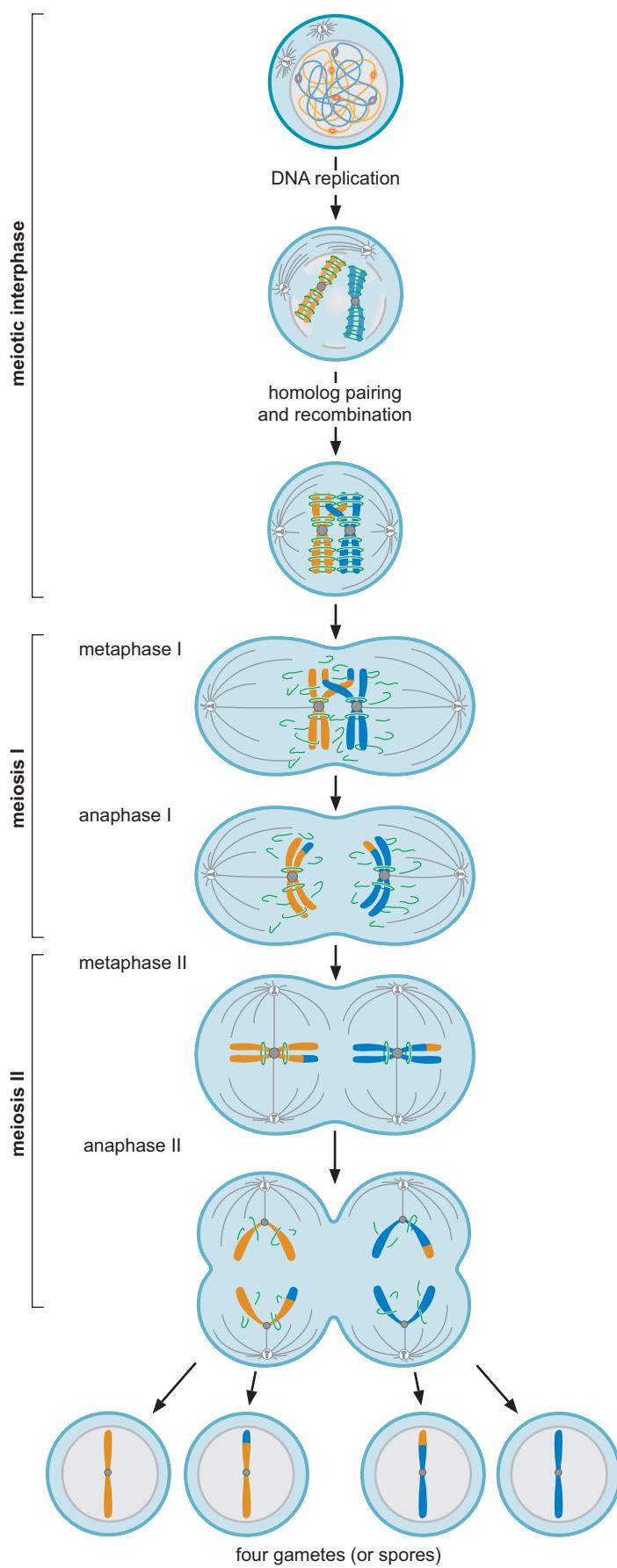
### **During Gap Phases, Cells Prepare for the Next Cell Cycle Stage and Check That the Previous Stage Is Completed Correctly**

The remaining two phases of the mitotic cell cycle are gap phases. G<sub>1</sub> occurs before DNA synthesis, and G<sub>2</sub> occurs between S phase and M phase. The gap phases of the cell cycle provide time for the cell to accomplish two goals: (1) to prepare for the next phase of the cell cycle and (2) to check that the previous phase of the cell cycle has been completed appropriately. For example, before entry into S phase, most cells must reach a certain size and level of protein synthesis to ensure that there are adequate proteins and nutrients to complete the next round of DNA synthesis. If there is a problem with a previous step in the cell cycle, **cell cycle checkpoints** stop the cell cycle to provide time for the cell to complete that step. For example, cells with damaged DNA arrest the cell cycle in G<sub>1</sub> before DNA synthesis or in G<sub>2</sub> before mitosis to prevent either event from occurring with damaged chromosomes. These delays allow time for the damage to be repaired before the cell cycle continues.

### **Meiosis Reduces the Parental Chromosome Number**

A second type of eukaryotic cell division is specialized to produce cells that have half the number of chromosomes as the parental cell. These cells go on to form egg and sperm cells involved in mating. This is accomplished by following DNA replication with two rounds of chromosome segregation. Like the mitotic cell cycle, the **meiotic cell cycle** includes a G<sub>1</sub>, S, and an elongated G<sub>2</sub> phase (Fig. 8-16). During the meiotic S phase, each chromosome is replicated, and the daughter chromatids remain associated as in the mitotic S phase. Cells that enter meiosis must be diploid and thus contain two copies of each chromosome before DNA replication, one derived from each parent. After DNA replication, these related sister-chromatid pairs, called **homologs**, pair with each other and recombine. Recombination between the homologs creates a physical linkage between the two homologs that is required to connect the two related sister-chromatid pairs during chromosome segregation. We discuss the details of meiotic recombination in Chapter 11.

The most significant difference between the mitotic and meiotic cell cycles occurs during chromosome segregation. Unlike mitosis, during which a single round of chromosome segregation follows DNA replication,



**FIGURE 8.16 Meiosis in detail.** Like mitosis, meiosis can be divided into discrete stages. After DNA replication, homologous sister chromatids pair with each other to form structures with four related chromosomes. For simplicity, only a single chromosome is shown segregating with the blue copies being from one parent and the yellow copies from the other. During pairing, chromatids from the different homologs recombine to form a link between the homologous chromosomes called a chiasma. During metaphase I, the two kinetochores of each sister-chromatid pair attach to one pole of the meiotic spindle. Homologous sister-chromatid kinetochores attach to opposite poles, creating tension that is resisted by the chiasma between the homologs and the cohesion between the sister chromatid arms. Entry into anaphase I is driven by the loss of sister-chromatid cohesion along the arms of the chromosomes. The loss of arm cohesion allows the recombined homologs to separate from one another. The sister chromatids remain attached through cohesion at the centromere. Meiosis II is very similar to mitosis. During meiotic metaphase II, two meiotic spindles are formed. As in mitotic metaphase, the kinetochores associated with each sister-chromatid pair attach to opposite poles of the meiotic spindles. During anaphase II, the remaining centromeric cohesion between the sisters is lost, and the sister chromatids separate from each other. The four separate sets of chromosomes are then packaged into nuclei and separated into four cells to create four spores or gametes. All DNA molecules are double-stranded. (Adapted, with permission, from Murray A. and Hunt T. 1993. *The cell cycle: An introduction*, Fig. 10.2. © Oxford University Press, Inc.)

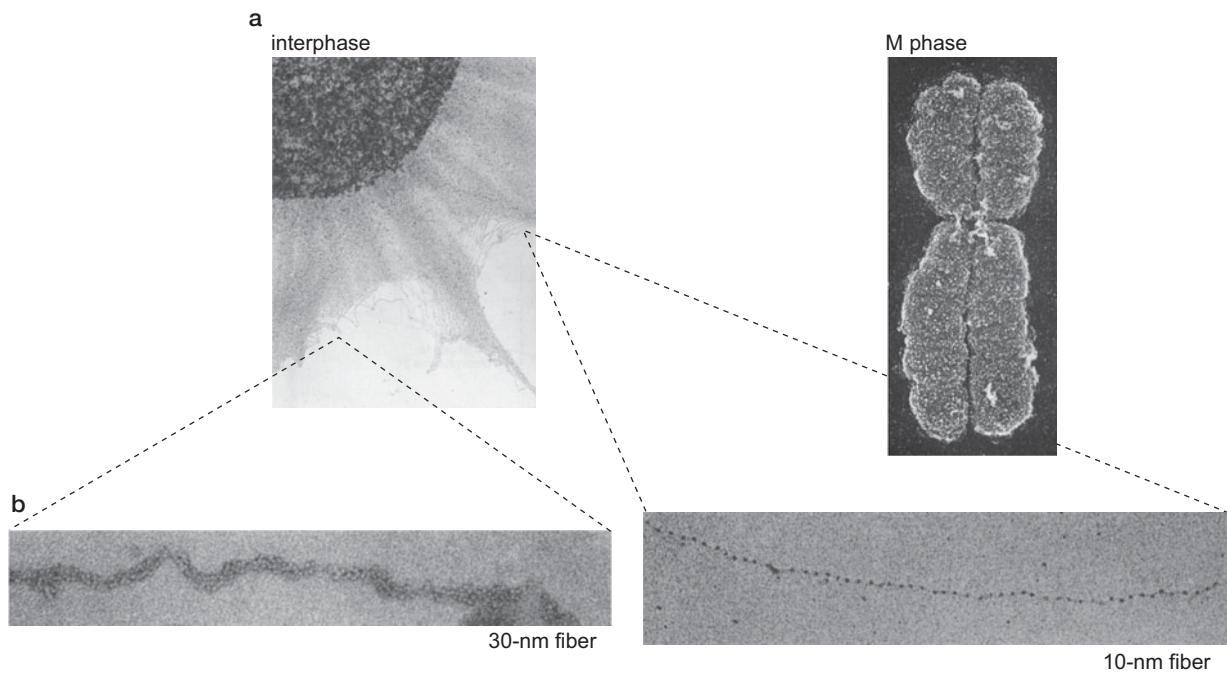
chromosomes participating in meiosis go through two rounds of segregation known as meiosis I and II. Like mitosis, each of these segregation events includes a prophase, metaphase, and anaphase stage. During the metaphase of **meiosis I**, also called metaphase I, the homologs attach to opposite poles of the microtubule-based spindle. This attachment is mediated by the kinetochore. Because both kinetochores of each sister-chromatid pair are attached to the same pole of the microtubule spindle, this interaction is referred to as **monovalent attachment** (in contrast to the bivalent attachment seen in mitosis, in which the kinetochores of each sister-chromatid pair bind to opposite poles of the spindle). As in mitosis, the paired homologs initially resist the tension of the spindle pulling them apart. In the case of meiosis I, this resistance is mediated through the physical connections between the homologs, called chiasma or crossovers, that are the result of recombination between the homologs. This resistance also requires sister-chromatid cohesion along the arms of the sister chromatids. When cohesion along the arms is eliminated during anaphase I, the recombined homologs are released from each other and segregate to opposite poles of the cell. Importantly, the cohesion between the sisters is maintained near the centromere, keeping the sister chromatids paired.

The second round of segregation during meiosis, **meiosis II**, is very similar to mitosis. The major difference is that a round of DNA replication does not precede this segregation event. Instead, a spindle is formed in association with each of the two newly separated sister-chromatid pairs. As in mitosis, during **metaphase II**, these spindles attach in a bivalent manner to the kinetochores of each sister-chromatid pair. The cohesion that remains at the centromeres after meiosis I is critical to oppose the pull of the spindle. The second round of chromosome segregation occurs in **anaphase II** and is initiated by the elimination of centromeric cohesion. At this point, there are four sets of chromosomes in the cell, each of which contains a single copy of each chromosome. A nucleus forms around each set of chromosomes, and then the cytoplasm is divided to form four haploid cells. These cells are now ready to mate to form new diploid cells.

### Different Levels of Chromosome Structure Can Be Observed by Microscopy

Microscopy has long been used to observe chromosome structure and function. Indeed, long before it was clear that chromosomes were the source of the genetic information in the cell, their movements and changes during cell division were well-understood. The compact nature of condensed mitotic or meiotic chromosomes makes them relatively easy to visualize even by simple light microscopy. Microscopic analysis of condensed chromosomes is used to determine the chromosomal makeup of human cells and detect such abnormalities as chromosomal deletions or individuals with too few or too many copies of a chromosome.

Outside of mitosis (i.e., in interphase), chromosomal DNA is less compact (Fig. 8-17a). In the electron microscope, two states of chromatin are observed: fibers with a diameter of either 30 nm or 10 nm (Fig. 8-17b). The 30-nm fiber is a more compact version of chromatin that is frequently folded into large loops reaching out from a protein core or scaffold. In contrast, the 10-nm fiber is a less compact form of chromatin that resembles a regular series of “beads on a string.” These beads are nucleosomes, and these protein–DNA structures play a critical role in regulating the structure and function of chromosomes. In the rest of the chapter, we first focus on the nature of the nucleosome, including how they are formed, and then describe



**FIGURE 8-17** Forms of chromatin structure seen in the electron microscope. (a) Electron micrographs of interphase and condensed M-phase DNA show the changes in the structure of chromatin. (b) Electron micrographs of different forms of chromatin in interphase cells show the 30-nm and 10-nm chromatin fibers (beads on a string). (a, Reprinted, with permission, from Alberts B. et al. 2002. *Molecular biology of the cell*, 4th ed., Figs. 4-21 and 4-23. Garland Science/Taylor & Francis LLC. © V. Foe.)

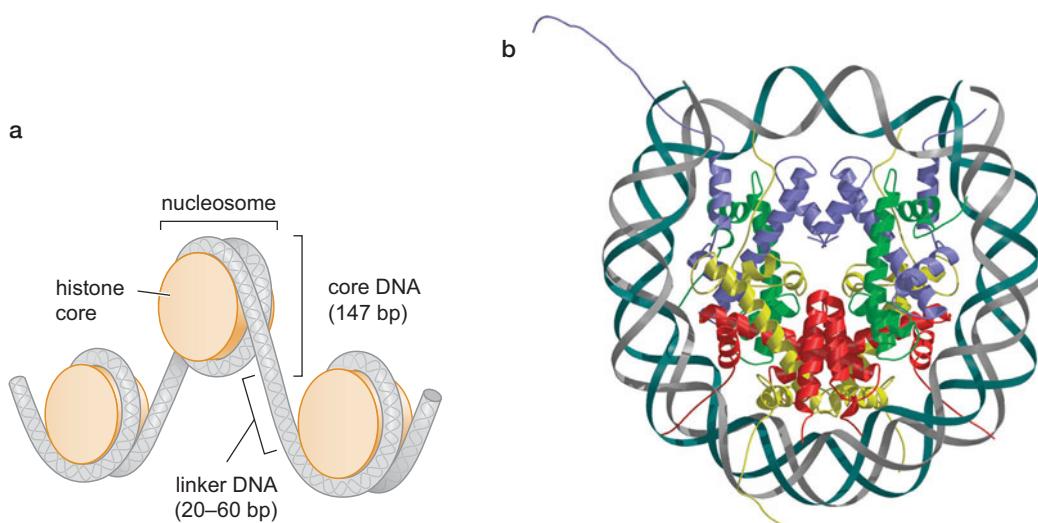
how nucleosome-dependent structures are regulated and how they control the accessibility of nuclear DNA.

## THE NUCLEOSOME

### Nucleosomes Are the Building Blocks of Chromosomes

The majority of the DNA in eukaryotic cells is packaged into nucleosomes. Each nucleosome is composed of a core of eight histone proteins and the DNA wrapped around them. The DNA between each nucleosome (the “string” in the “beads on a string” image in Fig. 8.17b) is called **linker DNA**. By assembling into nucleosomes, the DNA is compacted approximately sixfold. This is far short of the 1000–10,000-fold DNA compaction observed in eukaryotic cells. Nevertheless, this first stage of DNA packaging is essential for all of the remaining levels of DNA compaction.

The DNA most tightly associated with the nucleosome, called the **core DNA**, is wound about 1.65 times around the outside of the histone octamer like thread around a spool (Fig. 8-18). The length of DNA associated with each nucleosome can be determined using nuclease treatment (Box 8-1, Micrococcal Nuclease and the DNA Associated with the Nucleosome). The ~147-bp length of this DNA is an invariant feature of nucleosomes in all eukaryotic cells. In contrast, the length of the linker DNA between nucleosomes is variable. Typically, this distance is 20–60 bp, and each eukaryote has a characteristic average linker DNA length (Table 8-4). The difference



**FIGURE 8-18** DNA packaged into nucleosomes. (a) Schematic of the packaging and organization of nucleosomes. (b) Crystal structure of a nucleosome showing DNA wrapped around the histone protein core. (Red) H2A; (yellow) H2B; (purple) H3; (green) H4. Note that the colors of the different histone proteins here and in following structures are the same. (Luger K. et al. 1997. *Nature* **389**: 251–260.) Image prepared with MolScript, BobScript, and Raster3D.

in average linker DNA length is likely to reflect the differences in the larger structures formed by nucleosomal DNA in each organism, rather than differences in the nucleosomes themselves (see the next section on Higher-Order Chromatin Structure).

In any cell, there are stretches of DNA that are not packaged into nucleosomes. Typically, these are regions of DNA engaged in gene expression, replication, or recombination. Although not bound by nucleosomes, these sites are typically associated with nonhistone proteins that are either regulating or participating in these events. We discuss the mechanisms that remove nucleosomes from DNA and maintain such regions of DNA in a nucleosome-free state later and in Chapter 19.

### Histones Are Small, Positively Charged Proteins

Histones are by far the most abundant proteins associated with eukaryotic DNA. Eukaryotic cells commonly contain five abundant histones: H1, H2A, H2B, H3, and H4. Histones H2A, H2B, H3, and H4 are the **core histones**, and two copies of each of these histones form the protein core around which nucleosomal DNA is wrapped. Histone H1 is not part of the nucleosome core particle. Instead, it binds to the linker DNA and is referred to as a **linker**

**TABLE 8-4** Average Lengths of Linker DNA in Various Organisms

Species	Nucleosome Repeat Length (bp)	Average Linker DNA Length (bp)
<i>Saccharomyces cerevisiae</i>	160–165	13–18
Sea urchin (sperm)	~260	~110
<i>Drosophila melanogaster</i>	~180	~33
Human	185–200	38–53

**TABLE 8-5** General Properties of the Histones

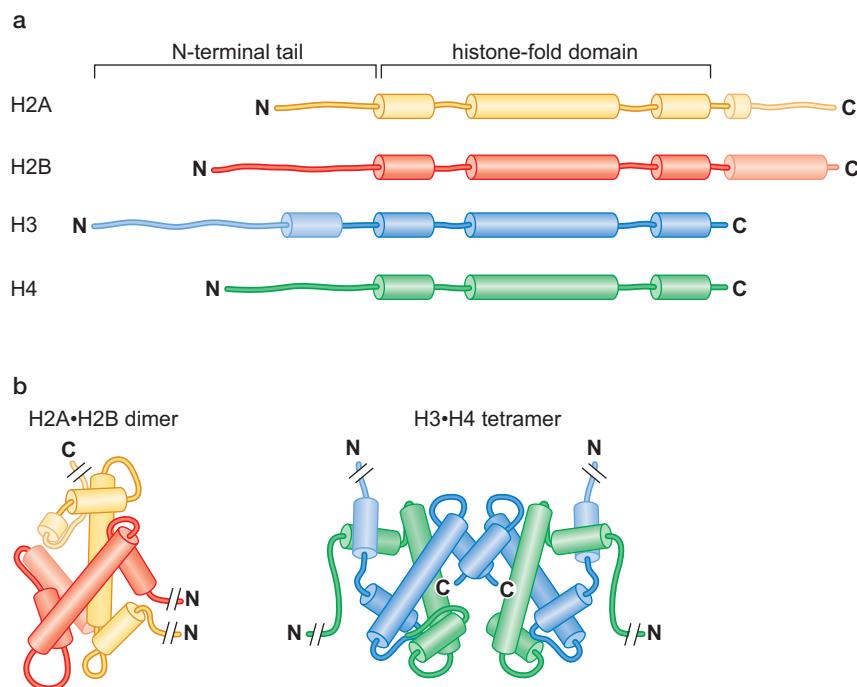
Histone Type	Histone	Molecular Weight ( $M_r$ )	Lysine and Arginine (%)
Core histones	H2A	14,000	20
	H2B	13,900	22
	H3	15,400	23
	H4	11,400	24
Linker histone	H1	20,800	32

**histone.** The four core histones are present in equal amounts in the cell. H1 is half as abundant as the other histones, which is consistent with the finding that only one molecule of H1 can associate with a nucleosome.

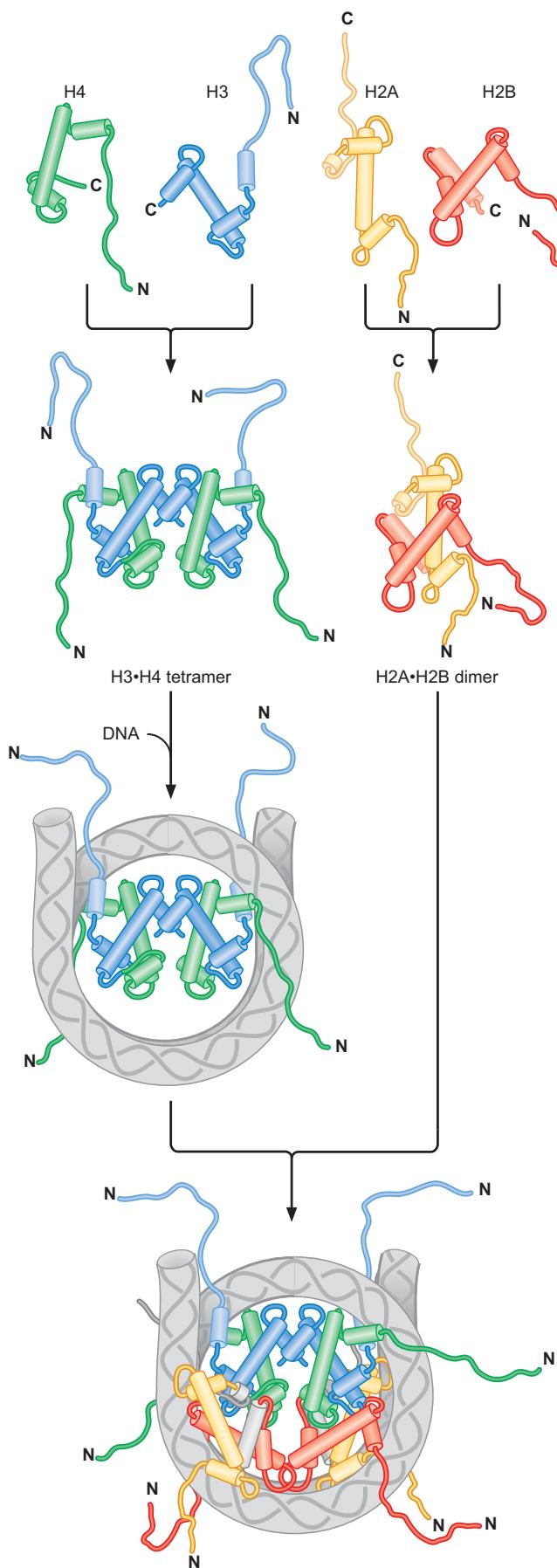
Consistent with their close association with the negatively charged DNA molecule, the histones have a high content of positively charged amino acids (Table 8-5). At least 20% of the residues in each histone are either lysine or arginine. The core histones are also relatively small proteins ranging in size from 11 to 15 kilodaltons (kDa). Histone H1 is slightly larger at ~21 kDa.

The protein core of the nucleosome is a disc-shaped structure that assembles in an ordered fashion only in the presence of DNA. Without DNA, the core histones form intermediate assemblies in solution. A conserved region found in every core histone, called the **histone-fold domain**, mediates the assembly of these histone-only intermediates (Fig. 8-19). The histone-fold domain is composed of three  $\alpha$ -helical regions separated by two short unstructured loops. This domain mediates the formation of head-to-tail heterodimers of specific pairs of histones. H3 and H4 histones first form heterodimers that then come together to form a tetramer with two molecules each of H3 and H4. In contrast, H2A and H2B form heterodimers in solution but not tetramers.

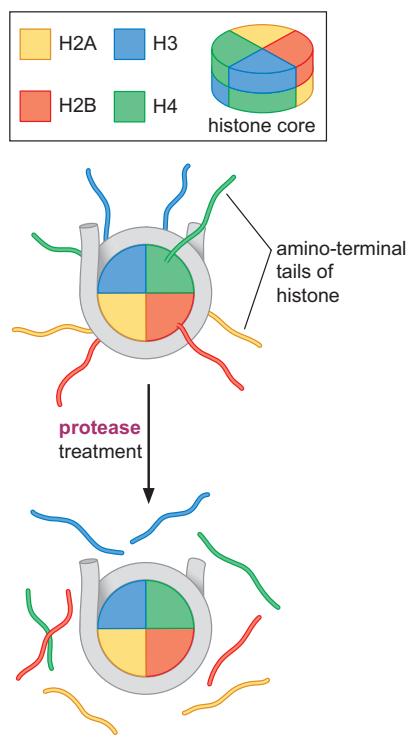
The assembly of a nucleosome involves the ordered association of these building blocks with DNA (Fig. 8-20). First, the H3·H4 tetramer binds to



**FIGURE 8-19** Core histones share a common structural fold. (a) The four histones are diagrammed as linear molecules. The regions of the histone-fold motif that form  $\alpha$  helices are indicated as cylinders. Note that there are adjacent regions of each histone that are structurally distinct including additional  $\alpha$ -helical regions. (b) The helical regions of two histones (here H2A and H2B) come together to form a dimer. H3 and H4 also use a similar interaction to form H3<sub>2</sub>·H4<sub>2</sub> tetramers. (Adapted, with permission, from Alberts B. et al. 2002. *Molecular biology of the cell*, 4th ed., p. 209, Fig. 4-26. © Garland Science/Taylor & Francis LLC.)



**FIGURE 8-20** Assembly of a nucleosome. The assembly of a nucleosome is initiated by the formation of a H3<sub>2</sub>H4<sub>2</sub> tetramer. The tetramer then binds to double-stranded DNA. The H3<sub>2</sub>H4<sub>2</sub> tetramer bound to DNA recruits two copies of the H2A-H2B dimer to complete the assembly of the nucleosome. (Adapted, with permission, from Alberts B. et al. 2002. *Molecular biology of the cell*, 4th ed., p. 210, Fig. 4-27. Garland Science/Taylor & Francis LLC, © J. Waterborg.)



**FIGURE 8-21** Amino-terminal tails of the core histones are accessible to proteases. Treatment of nucleosomes with limiting amounts of proteases that cleave after basic amino acids (e.g., trypsin) specifically removes the amino-terminal “tails” leaving the histone core intact.

DNA; then two H2A·H2B dimers join the H3·H4-DNA complex to form the final nucleosome. We discuss how and when this assembly process is accomplished in the cell later in this chapter.

The core histones each have an amino-terminal extension, called a **tail** because it lacks a defined structure and is accessible within the intact nucleosome. This accessibility can be detected by treatment of nucleosomes with the protease trypsin (which specifically cleaves proteins after positively charged amino acids). Treatment of nucleosomes with trypsin rapidly removes the accessible amino-terminal tails of the histones but cannot cleave the tightly packed histone-fold regions (Fig. 8-21). The exposed amino-terminal tails are not required for the association of DNA with the histone octamer, because DNA is still tightly associated with the nucleosome after protease treatment. Instead, the tails are the sites of extensive post-translational modifications that alter the function of individual nucleosomes. These modifications include phosphorylation, acetylation, and methylation on serine, lysine, and arginine residues. We shall return to the role of histone tail modification in nucleosome function later. We now turn to the detailed structure of the nucleosome.

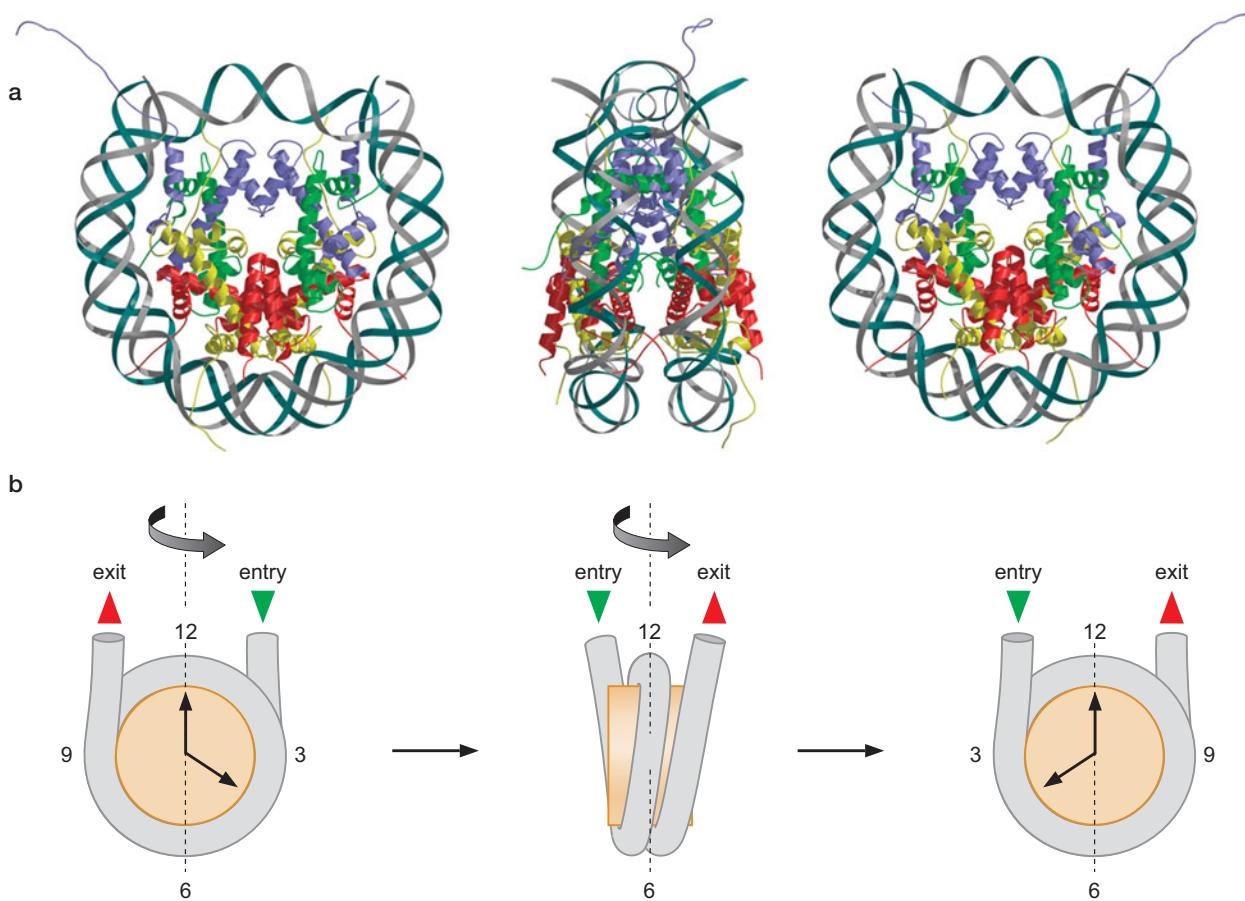
### The Atomic Structure of the Nucleosome

The high-resolution three-dimensional (3D) structure of the nucleosome core particle (see Fig. 8-18b) (147 bp of DNA plus an intact histone octamer) has provided many insights into nucleosome function. The high affinity of the nucleosome for DNA, the distortion of the DNA when bound to the nucleosome, and the lack of DNA sequence specificity can each be explained by the nature of the interactions between the histones and the DNA. The structure also sheds light on the function and location of the amino-terminal tails. Finally, the interaction between the DNA and the histone octamer provides insight into the dynamic nature of the nucleosome and the process of nucleosome assembly. We discuss each of these properties of the nucleosome in the following sections.

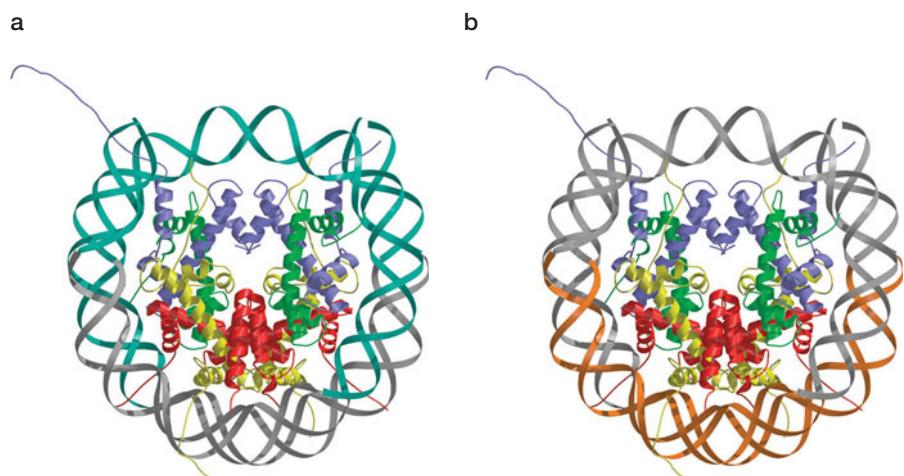
### Histones Bind Characteristic Regions of DNA within the Nucleosome

Although not perfectly symmetrical, the nucleosome has an approximate twofold axis of symmetry, called the **dyad axis**. This can be visualized by thinking of the face of the octamer disc as a clock with the midpoint of the 147 bp of DNA located at the 12 o'clock position (Fig. 8-22). This places the ends of the DNA just short of 11 and 1 o'clock. A line drawn from 12 o'clock to 6 o'clock through the middle of the disc defines the dyad axis. Rotation of the nucleosome around this axis by 180° reveals a view of the nucleosome nearly identical to that observed before rotation (see Structural Tutorial 8-1).

The H3·H4 tetramers and H2A·H2B dimers each interact with a particular region of the DNA within the nucleosome (Fig. 8-23). Of the 147 bp of DNA included in the structure, the histone-fold regions of the H3·H4 tetramer interact with the central 60 bp. The amino-terminal region of H3 most proximal to the histone-fold region forms a fourth  $\alpha$  helix that interacts with the final 13 bp at each end of the bound DNA (this  $\alpha$  helix is distinct from the unstructured H3 amino-terminal tail described above). If we picture the nucleosome with a clock face as described above, the H3·H4 tetramer forms the top half of the histone octamer. Histone H3·H4 tetramers occupy a key position in the nucleosome by binding the middle *and* both ends of the DNA (turquoise DNA in



**FIGURE 8-22** The nucleosome has an approximate twofold axis of symmetry. (a) Three-dimensional structure. (b) Cartoon illustrating “clock face” analogy to nucleosome. Three views of the nucleosome are shown in each representation. Each view shows a 90° rotation around the axis between 12 and 6 o’clock positions illustrated in the first panel of b. (a, Luger K. et al. 1997. *Nature* 389: 251–260.) Images prepared with MolScript, BobScript, and Raster3D.



**FIGURE 8-23** Interactions of the histones with nucleosomal DNA. (a) H3-H4 binds the middle and the ends of the DNA (turquoise). (b) H2A-H2B binds 30 bp of DNA on one side of the nucleosome (orange). (Luger K. et al. 1997. *Nature* 389: 251–260.) Images prepared with MolScript, BobScript, and Raster3D.

## ► KEY EXPERIMENTS

**Box 8-1** Micrococcal Nuclease and the DNA Associated with the Nucleosome

Nucleosomes were first purified by treating chromosomes with a sequence-nonspecific nuclease called **micrococcal nuclease (MNase)**. The ability of this enzyme to cleave DNA is primarily governed by the accessibility of the DNA. Thus, MNase cleaves protein-free DNA sequences rapidly and protein-associated DNA sequences poorly. Limited treatment of chromosomes with this enzyme results in a nuclease-resistant population of DNA molecules that are primarily associated with histones. These DNA molecules are between 160 and 220 bp in length and are associated with two copies each of histones H2A, H2B, H3, and H4. On average, these particles include the DNA tightly associated with the nucleosome as well as one unit of linker DNA. More extensive MNase treatment degrades all of the linker DNA. The remaining minimal nucleosome includes only 147 bp of DNA and is called the **nucleosome core particle**.

The average length of DNA associated with each nucleosome can be measured in a simple experiment (Box 8-1 Fig. 1). Chromatin is treated with the enzyme micrococcal nuclease but this time only gently. This results in single cuts in some but not all linker DNA. After nuclease treatment, the DNA is extracted from all proteins (including the histones) and subjected to gel electrophoresis to separate the DNA by size. Electrophoresis reveals a “ladder” of DNA fragments that are multiples of the average nucleosome-to-nucleosome distance. A ladder of fragments is observed because the MNase-treated chromatin is only partially digested. Thus, sometimes, multiple nucleosomes remain unseparated by digestion, leading to DNA fragments equivalent to all of the DNA bound by these nucleosomes. Further digestion would result in all linker DNA being cleaved and the formation of nucleosome core particles and a single ~147-bp fragment.

**BOX 8-1 FIGURE 1** Progressive digestion of nucleosomal DNA with MNase. (Courtesy of R.D. Kornberg.)

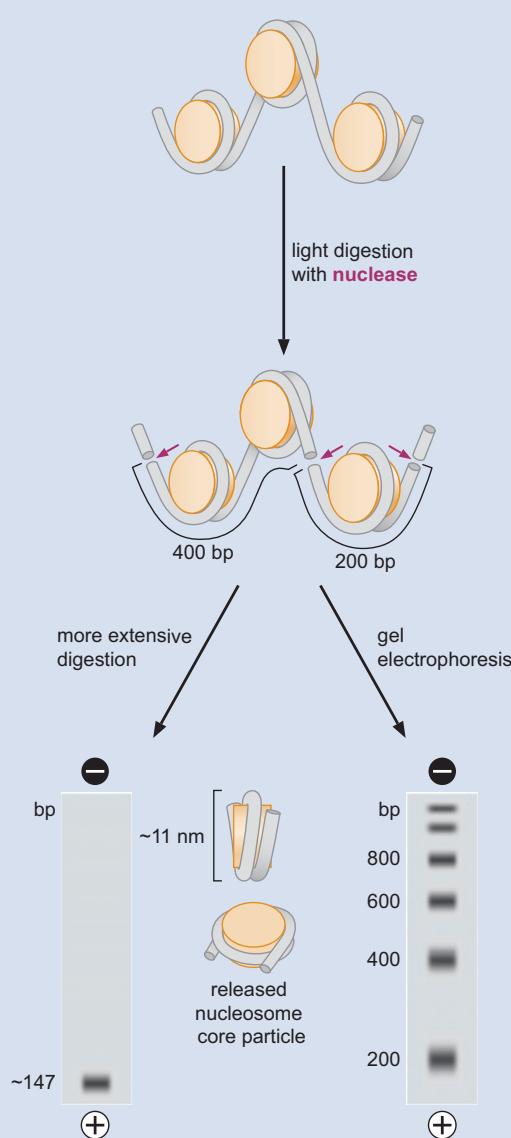


Fig. 8-23a). The two H2A-H2B dimers each associate with ~30 bp of DNA on either side of the central 60 bp of DNA bound by H3 and H4. Using the clock analogy again, the DNA associated with H2A-H2B is located from ~5 o'clock to 9 o'clock on either face of the nucleosome disc. Together, the two H2A-H2B dimers form the bottom part of the histone octamer located across the disc from the DNA ends (orange DNA in Fig. 8-23b).

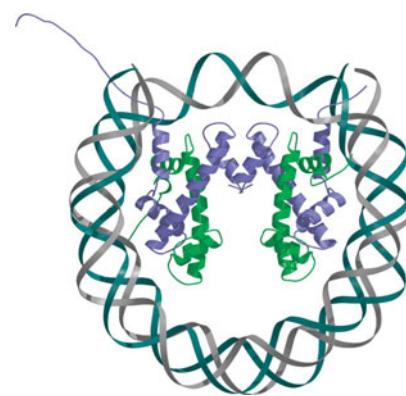
The extensive interactions between the H3-H4 tetramer and the DNA help to explain the ordered assembly of the nucleosome (Fig. 8-24). H3-H4 tetramer association with the middle and ends of the bound DNA would result in the DNA being extensively bent and constrained, making the association of H2A-H2B dimers relatively easy. In contrast, the relatively short length of DNA bound by H2A-H2B dimers is not sufficient to prepare the DNA for H3-H4 tetramer binding.

### Many DNA Sequence–Independent Contacts Mediate the Interaction between the Core Histones and DNA

A closer look at the interactions between the histones and the nucleosomal DNA reveals the structural basis for the binding and bending of the DNA within the nucleosome. Fourteen distinct sites of contact are observed, one for each time the minor groove of the DNA faces the histone octamer (Fig. 8-25). The association of DNA with the nucleosome is mediated by a large number (about 40) of hydrogen bonds between the histones and the DNA. The majority of these hydrogen bonds are between the proteins and the oxygen atoms in the phosphodiester backbone near the minor groove of the DNA. Only seven hydrogen bonds are made between the protein side chains and the bases, and all of these are made in the minor groove of the DNA.

The large number of these hydrogen bonds (a typical sequence-specific DNA-binding protein only has about 20 hydrogen bonds with DNA) provides the driving force to bend the DNA. The highly basic nature of the histones further facilitates DNA bending by masking the negative charge of the phosphates that ordinarily resists DNA bending. This is because when DNA is bent, the phosphates on the inside of the bend are brought into unfavorably close proximity. The positively charged nature of the histones also facilitates the close juxtaposition of the two adjacent DNA helices necessary to wrap the DNA more than once around the histone octamer.

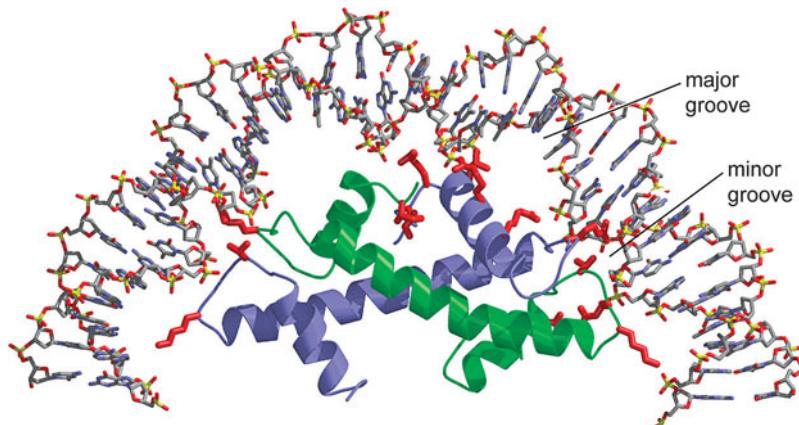
The finding that all of the sites of contact between the histones and the DNA involve either the minor groove or the phosphate backbone is consistent with the non-sequence-specific nature of the association of the histone octamer with DNA. Neither the phosphate backbone nor the minor groove is rich in base-specific information. Moreover, of the seven hydrogen bonds formed with the bases in the minor groove, *none* is with parts of the bases that distinguish between G:C and A:T base pairs (see Chapter 4, Fig. 4-10).



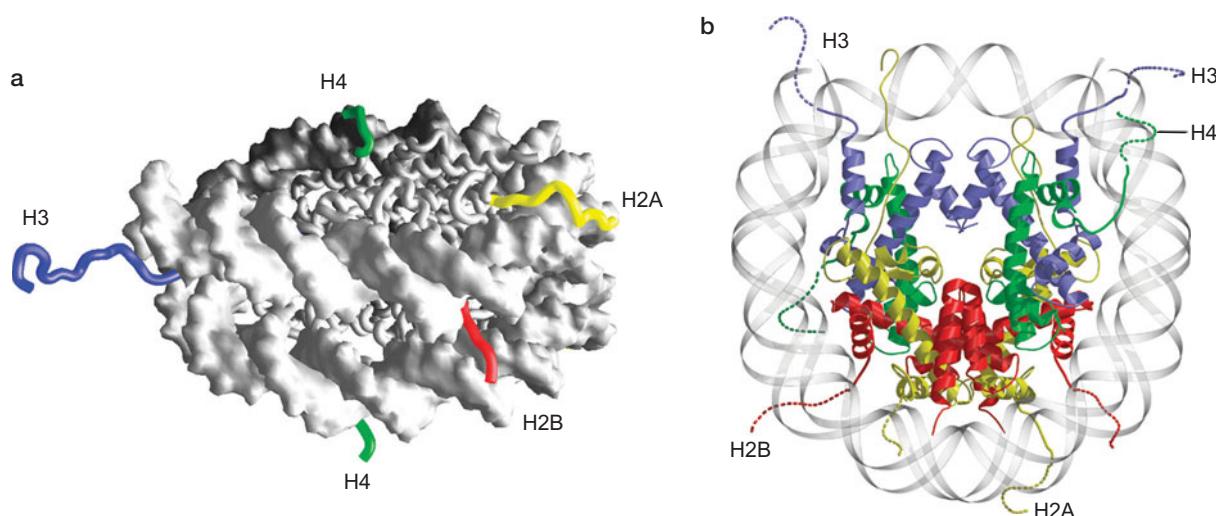
**FIGURE 8-24 Nucleosome lacking H2A and H2B.** The H2A and H2B histones have been artificially removed from this view of the nucleosome. This structure is likely to resemble the DNA·H3<sub>2</sub>·H4<sub>2</sub> tetramer intermediate in the assembly of a nucleosome (see Fig. 8-20). (Luger K. et al. 1997. *Nature* **389**: 251–260.) Image prepared with MolScript, BobScript, and Raster3D.

### The Histone Amino-Terminal Tails Stabilize DNA Wrapping around the Octamer

The structure of the nucleosome also tells us something regarding the histone amino-terminal tails. The four H2B and H3 tails emerge from between



**FIGURE 8-25 Sites of contact between the histones and the DNA.** For clarity, only the interactions between a single H3-H4 dimer are shown. A subset of the parts of the histones that interact with the DNA is highlighted in red. Note that these regions cluster around the minor groove of the DNA. (Luger K. et al. 1997. *Nature* **389**: 251–260.) Image prepared with MolScript, BobScript, and Raster3D.



**FIGURE 8-26** Histone tails emerge from the core of the nucleosome at specific positions. (a) The side view illustrates that the H3 and H2B tails emerge from between the two DNA helices. In contrast, the H4 and H2A tails emerge either above or below both DNA helices. (Luger K. et al. 1997. *Nature* **389**: 251–260.) Image prepared with GRASP. (b) The position of the tails relative to the entry and exit of the DNA. This view reveals that the histone tails emerge at numerous positions relative to the DNA. (Davey C.A. et al. 2002. *J. Mol. Biol.* **319**: 1097–1113.) Image prepared with MolScript, BobScript, and Raster3D.

the two DNA helices. In each case, their path of exit is formed by two adjacent minor grooves, making a “gap” between the two DNA helices just big enough for a polypeptide chain (Fig. 8-26a). Strikingly, the H2B and H3 tails emerge at approximately equal distances from each other around the octamer disc (at ~1 and 11 o’clock for the H3 tails and 4 and 8 o’clock for H2B). Instead of emerging between the two DNA helices, the H2A and H4 amino-terminal tails emerge from either “above” or “below” both DNA helices (Fig. 8-26a). These tails are also distributed around the face of the nucleosome with the H2A tails emerging at 5 and 7 o’clock and the H4 tails at 3 and 9 o’clock (Fig. 8-26b). By emerging both between and on either side of the DNA helices, the histone tails can be thought of as the grooves of a screw, directing the DNA to wrap around the histone octamer disc in a left-handed manner. As we discussed in Chapter 4, the left-handed nature of the DNA wrapping introduces negative supercoils in the DNA. The parts of the tails most proximal to the histone disc (and therefore not subject to the protease cleavage discussed above) also make some of the many hydrogen bonds between the histones and the DNA as they pass by the DNA.

### Wrapping of the DNA around the Histone Protein Core Stores Negative Superhelicity

Each nucleosome added to a covalently closed circular template changes the linking number of the associated DNA by approximately  $-1.2$ . Because the remainder of the DNA is kept relaxed by topoisomerases, the DNA that is packaged into nucleosomes would become negatively supercoiled if nucleosomes were removed from the DNA. Thus, nucleosomes can be viewed as storing or stabilizing negative superhelicity. Why would the cell want to maintain a stockpile of negative superhelicity? There are many instances when it is useful to drive unwinding of DNA in the cell, including initiation of DNA replication, transcription, and recombination. Importantly,

negatively supercoiled DNA favors DNA unwinding (see Chapter 4, Fig. 4-17). Thus, removal of a nucleosome not only allows increased access to the DNA, but also facilitates DNA unwinding of nearby DNA sequences (Box 8-2, Nucleosomes and Superhelical Density).

If nucleosomes store negative superhelicity in eukaryotic cells, what serves the equivalent function in prokaryotic cells? The answer for many prokaryotic organisms is that the entire genome is maintained in a negatively supercoiled state. This is accomplished by a specialized topoisomerase called **gyrase** that has the ability to introduce negative superhelicity into relaxed DNA by reducing the linking number. For example, in *E. coli* cells, gyrase action results in the genome having an average superhelical density of approximately  $-0.07$ . The addition of negative supercoils into otherwise relaxed DNA is an energy-requiring reaction. Consistent with this, gyrase requires ATP to introduce negative supercoils. In the absence of ATP, gyrase can only relax DNA (e.g., reduce the linking number of positively supercoiled DNA).

Not all bacteria need to maintain their DNA in a negatively supercoiled state. Bacteria that prefer to grow at very high temperatures ( $>80^{\circ}\text{C}$ ) must expend energy to *prevent* their DNA from unwinding due to thermal denaturation. These organisms have a different topoisomerase called **reverse gyrase**. Consistent with its name, reverse gyrase increases the linking number of relaxed DNA in the presence of ATP. By keeping the genome positively supercoiled, reverse gyrase counteracts the effect of thermal denaturation that would ordinarily result in many regions of the genome being unwound.

## HIGHER-ORDER CHROMATIN STRUCTURE

---

### Heterochromatin and Euchromatin

From the earliest observations of chromosomes in the light microscope, it was clear that they were not uniform structures. Early studies of chromosomes divided chromosomal regions into two categories: **euchromatin** and **heterochromatin**. Heterochromatin was characterized by dense staining with a variety of dyes and a more condensed appearance, whereas euchromatin had the opposite characteristics, staining poorly with dyes and having a relatively open structure. As our molecular understanding of genes and their expression advanced, it became clear that heterochromatic regions of chromosomes had very limited gene expression. In contrast, euchromatic regions showed higher levels of gene expression, suggesting that these different structures were connected to global levels of gene expression.

Heterochromatic regions show little gene expression, but this does not mean that these regions are not important. As we shall learn when gene expression is discussed, keeping a gene turned off can be just as important as turning a gene on. In addition, heterochromatin is associated with particular chromosomal regions, including the telomere and the centromere, and is important for the function of both of these key chromosomal elements.

Over the years, researchers have gained a more complete molecular understanding of heterochromatin and euchromatin structure. It is clear that DNA in both types of chromatin is packaged into nucleosomes. The difference between heterochromatin structure and euchromatin structure is how the nucleosomes in these different chromosomal regions are (or are not) assembled into larger assemblies. It has become clear that heterochromatic regions are composed of nucleosomal DNA assembled into

**Box 8-2** Nucleosomes and Superhelical Density

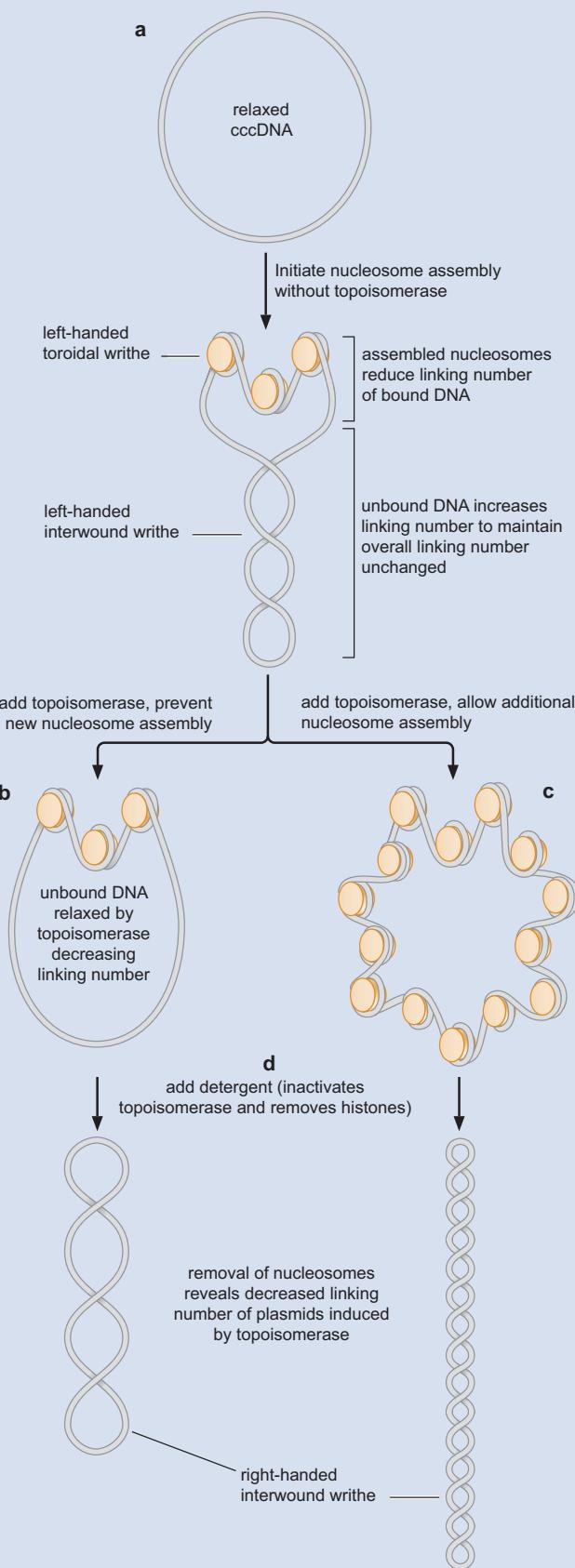
Why do nucleosomes alter the topological state of the DNA they include? As described in Chapter 4, there are two forms of writhe that can contribute to the formation of supercoiled DNA: toroidal and interwound (also referred to as *plectonemic*). The wrapping of DNA around the histone octamer is a form of toroidal writhe. The handedness of the writhe controls whether it introduces positive or negative supercoils (i.e., increases or decreases the linking number of the associated DNA). For toroidal writhe, left-handed wrapping induces negative superhelicity (for interwound writhe, the opposite is true; right-handed pitch is associated with negative superhelicity). Thus, the left-handed toroidal wrapping of DNA around the nucleosome reduces the linking number of the associated DNA. For this reason, nucleosomes preferentially form with DNA that has negative superhelical density. In contrast, assembling nucleosomes on DNA that has positive superhelical density is very difficult.

The assembly of many nucleosomes on covalently closed, circular DNA (cccDNA) requires the presence of a topoisomerase to accommodate changes in the linking number of the DNA bound to histones (see Box 8-2 Fig. 1). Without a topoisomerase present, for every nucleosome formed with the cccDNA, the unbound DNA (not associated with nucleosomes) would have to accommodate an equivalent *increase* in the linking number (remember that the overall linking number of a cccDNA is fixed in the absence of a topoisomerase). Thus, the unbound DNA would accumulate increased linking number and positive superhelical density. The more positively supercoiled the unbound DNA, the more difficult it is for additional nucleosomes to assemble on this DNA.

Addition of a topoisomerase greatly facilitates nucleosome association with cccDNA. When a topoisomerase is present during nucleosome assembly, it cannot act on the DNA bound to the nucleosome. Instead, the topoisomerase relaxes the DNA not included in nucleosomes, reducing the positive superhelical density in these regions by decreasing the linking number. By maintaining the unbound DNA in a relaxed state, topoisomerases facilitate the binding of histones to the DNA and the formation of additional nucleosomes. Importantly, the overall effect on the plasmid is that the linking number is decreased as more nucleosomes are assembled.

The decrease in the linking number caused by topoisomerase during nucleosome assembly can be used as an assay for this

**BOX 8-2 FIGURE 1** Topoisomerase is required for nucleosome assembly using covalently closed, circular DNA (cccDNA). (a) Assembly of nucleosomes using cccDNA in the absence of topoisomerase is limited by the accumulation of positive superhelicity in the DNA not associated with nucleosomes. (b) Addition of topoisomerase without additional nucleosome assembly illustrates how topoisomerase reduces the linking number to relax the DNA not incorporated into nucleosomes. (c) Additional nucleosome assembly in the presence of topoisomerase. (d) Simultaneous removal of histones and inactivation of topoisomerase (e.g., by addition of a strong detergent) reveals the reduced linking number associated with nucleosomal DNA.



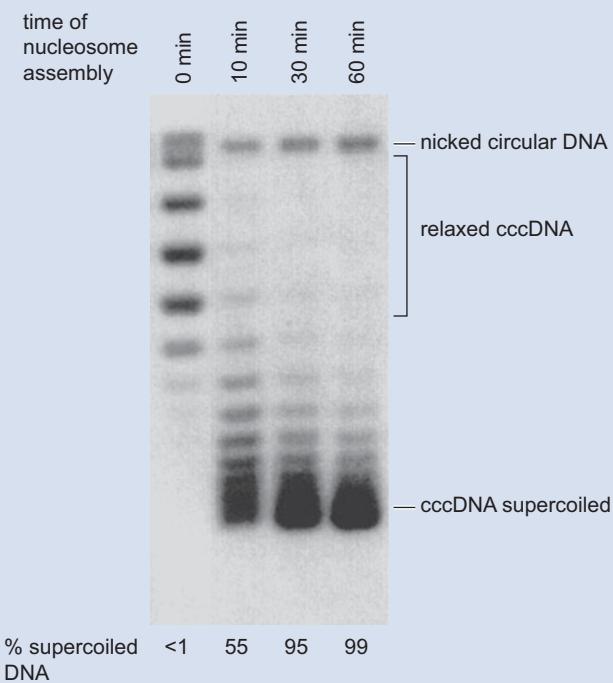
**Box 8-2** (Continued)

event. The assay takes advantage of the ability to distinguish between relaxed and supercoiled cccDNA by gel electrophoresis (see Chapter 4, Fig. 4-27). The first step is to assemble nucleosomes onto a cccDNA in the presence of a topoisomerase. At appropriate times, a strong detergent (e.g., SDS [sodium dodecyl sulfate]) is added to the assembly reaction, rapidly inactivating the topoisomerase and removing histones from the DNA. The resulting DNA is then separated by gel electrophoresis to determine the supercoiled nature of the DNA. Because the detergent inactivates the topoisomerase at the same time as removing the histones from the DNA, the linking number of the DNA assembled into nucleosomes is preserved. On average, the topoisomerase will have decreased the linking number by  $-1.2$  for each nucleosome assembled on the cccDNA. Thus, the more nucleosomes assembled on the cccDNA, the more negatively supercoiled is the cccDNA (Box 8-2 Fig. 1c,d). This can easily be observed by the faster migration of supercoiled DNA during gel electrophoresis (Box 8-2 Fig. 2).

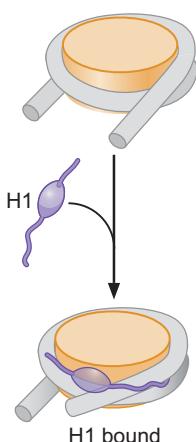
Because nucleosomal DNA wraps around the histone protein 1.65 times, the formation of a single nucleosome using covalently closed, circular plasmid would create a writhe of  $-1.65$  and thus change the linking number by an equivalent amount. As described above, when the change in linking number associated with each nucleosome was measured, the number was lower than this, approximately  $-1.2$  for each nucleosome added. This discrepancy is referred to as the “nucleosome linking number paradox,” and the solution to this conundrum was revealed when the high-resolution crystal structure of the nucleosome was solved. Careful analysis of the DNA associated with the histone protein core showed that the number of bases per turn was reduced relative to naked DNA (from 10.5 to 10.2 bp/turn). A reduction in the number of base pairs per turn results in an increase in the linking number for that DNA. Consider the example of a 10,500-bp cccDNA described in Chapter 4. Normal B-form DNA will have 10.5 bp/turn, resulting in a linking number of +1000 for the plasmid ( $10,500/10.5$ ). In contrast, the same DNA with a pitch of 10.2 bp/turn will have a linking number of approximately +1029 ( $10,500/10.2$ ). Thus, by decreasing the number of base pairs per turn of the helix, binding to the histone octamer causes a slight increase in the linking number over the length of the nucleosome-bound DNA. This change reduces the change in linking number per nucleosome assembled from  $-1.65$  to  $-1.2$ . The difference of approximately +0.4 per nucleosome can be calculated using the difference in the number of base pairs per turn and the length of DNA associated with a nucleosome.

Are these issues relevant to the linear eukaryotic chromosomes? For short linear fragments, superhelicity is not relevant because the ends of the DNA can rotate to accommodate

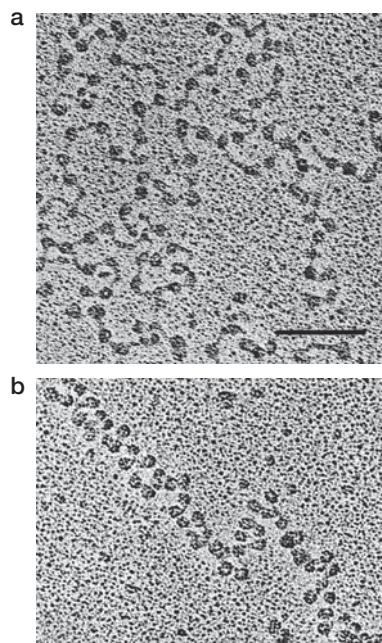
changes in the linking number. But this is not true for the very large linear chromosomes of eukaryotic cells. First, the large size of these chromosomes would not allow rapid enough rotation to dissipate changes in DNA superhelicity easily. More importantly, as we discuss later, the chromosome is not a simple linear strand of DNA. Each chromosomal DNA is folded into a more compact structure composed of large loops that are tethered to a protein structure called the **nuclear scaffold**. These attachments serve to topologically isolate one loop from the next and prevent free rotation of chromosomal DNA.



**BOX 8-2 FIGURE 2** Example of a nucleosome assembly assay that measures the associated decrease in linking number. Nucleosome assembly was performed on a relaxed cccDNA in the presence of a topoisomerase. Before the initiation of assembly (0 min) or at various times during the nucleosome assembly reaction, detergent was added and the DNA was separated on a nondenaturing agarose gel and visualized by staining with ethidium bromide. Although an agarose gel does not distinguish between positively and negatively supercoiled cccDNA, the ability of DNA intercalators that increase the linking number to shift the DNA toward the top of the gel (and a more relaxed state) can be used to show that these are negatively supercoiled cccDNAs (not shown). (Reprinted, with permission, from Ito T. et al. 1997. *Cell* 90: 145–155, Fig. 2c. © Elsevier.)



**FIGURE 8-27** Histone H1 binds two DNA helices. Upon interacting with a nucleosome, histone H1 binds to the linker DNA at one end of the nucleosome and the central DNA helix of the nucleosome bound DNA (the middle of the 147 bp bound by the core histone octamer).



**FIGURE 8-28** Addition of H1 leads to more compact nucleosomal DNA. The two images show an electron micrograph of nucleosomal DNA in the absence (a) and presence (b) of histone H1. Note the more compact and defined structure of the DNA in the presence of histone H1. (Reprinted, with permission, from Thoma F. et al. 1979. *J. Cell Biol.* **83**: 403–427, Figs. 4 and 6. © Rockefeller University Press.)

higher-order structures that result in a barrier to gene expression. In contrast, euchromatic nucleosomes are found in much less organized assemblies. In the following sections, we discuss what is known regarding how nucleosomes are assembled into higher-order structures.

### Histone H1 Binds to the Linker DNA between Nucleosomes

Once nucleosomes are formed, the next step in the packaging of DNA is the binding of histone H1. Like the core histones, H1 is a small, positively charged protein (see Table 8-5). H1 interacts with the linker DNA between nucleosomes, further tightening the association of the DNA with the nucleosome. This can be detected by the increased protection of nucleosomal DNA from micrococcal nuclease digestion. Thus, beyond the 147 bp protected by the core histones, addition of histone H1 to a nucleosome protects an additional 20 bp of DNA from micrococcal nuclease digestion.

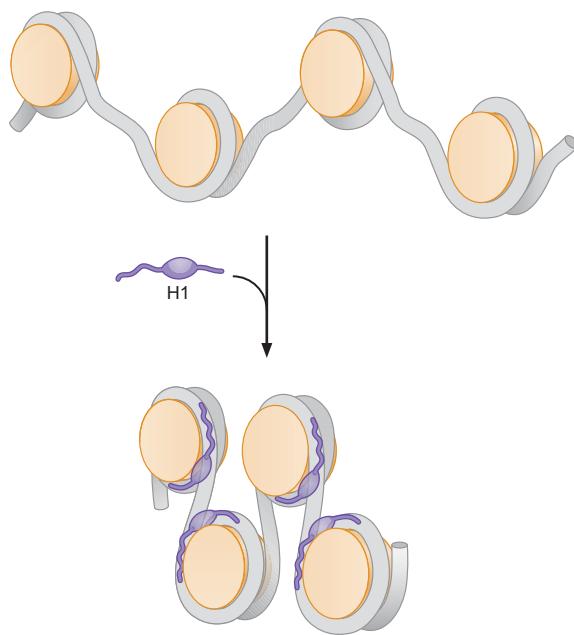
Histone H1 has the unusual property of binding two distinct regions of the DNA duplex. Typically, these two regions are part of a single DNA molecule associated with a nucleosome (Fig. 8-27). The sites of H1 binding are located asymmetrically relative to the nucleosome. One of the two regions bound by H1 is the linker DNA at *one* end of the nucleosome. The second site of DNA binding is in the middle of the associated 147 bp (the only DNA duplex present at the dyad axis). Thus, the additional DNA, protected from the nuclease digestion described above, is restricted to linker DNA on only *one* side of the nucleosome. By bringing these two regions of DNA into close proximity, H1 binding increases the length of the DNA wrapped tightly around the histone octamer.

H1 binding produces a more defined angle of DNA entry and exit from the nucleosome. This effect, which can be visualized in the electron microscope (Fig. 8-28), results in the nucleosomal DNA taking on a distinctly zigzag appearance. The angles of entry and exit observed vary substantially depending on conditions (including salt concentration, pH, and the presence of other proteins). If we assume that these angles are  $\sim 20^\circ$  relative to the dyad axis, this would result in a pattern in which nucleosomes would alternate on either side of a central region of linker DNA bound by histone H1 (Fig. 8-29).

### Nucleosome Arrays Can Form More Complex Structures: The 30-nm Fiber

Binding of H1 stabilizes higher-order chromatin structures. In the test tube, as salt concentrations are increased, the addition of histone H1 results in the nucleosomal DNA forming a **30-nm fiber**. This structure, which can also be observed *in vivo*, represents the next level of DNA compaction. Importantly, the incorporation of DNA into this fiber makes the DNA less accessible to many DNA-dependent enzymes (such as RNA polymerases).

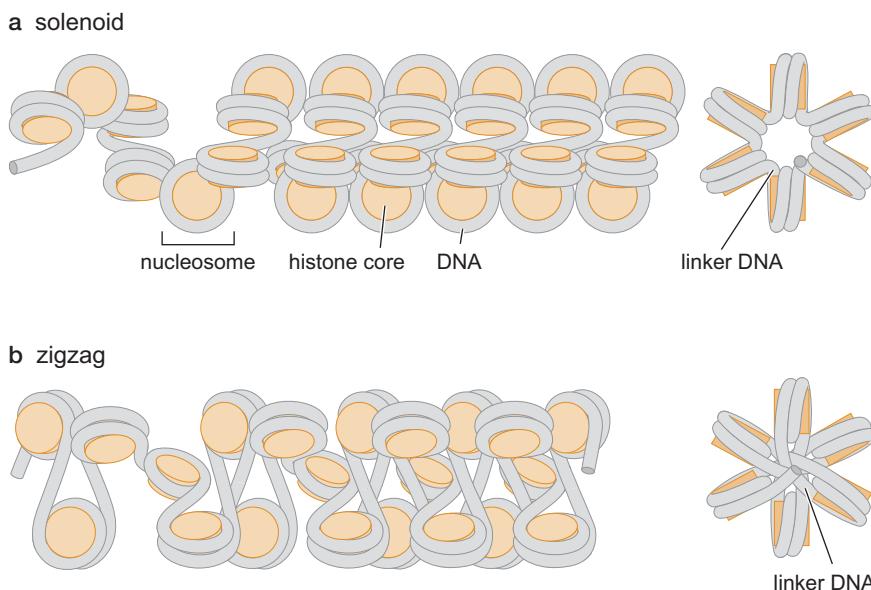
There are two models for the structure of the 30-nm fiber. In the **solenoid model**, the nucleosomal DNA forms a superhelix containing approximately six nucleosomes per turn (Fig. 8-30a). This structure is supported by both electron microscopy and X-ray diffraction studies, which indicate that the 30-nm fiber has a helical pitch of  $\sim 11$  nm. This distance is also the approximate diameter of the nucleosome disc, suggesting that the 30-nm fiber is composed of nucleosome discs stacked on edge in the form of a helix (see Fig. 8-30a). In this model, the flat surfaces on either face of the histone octamer disc are adjacent to each other, and the DNA surface of the nucleosomes forms the accessible surface of the superhelix. The linker DNA is buried in the center of the superhelix, but it never passes through the axis



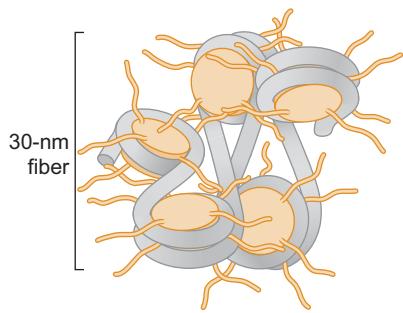
**FIGURE 8-29** Histone H1 induces tighter DNA wrapping around the nucleosome. The two illustrations show a comparison of the wrapping of DNA around the nucleosome in the presence and absence of histone H1. One histone H1 can associate with each nucleosome.

of the fiber. Rather, the linker DNA circles around the central axis as the DNA moves from one nucleosome to the next.

An alternative model for the 30-nm fiber is the “zigzag” model (Fig. 8-30b). This model is based on the zigzag pattern of nucleosomes formed upon H1 addition. In this case, the 30-nm fiber is a compacted form of these zigzag nucleosome arrays. A recent X-ray structure of a single DNA molecule participating in four nucleosomes and biophysical studies of the spring-like nature of isolated 30-nm fibers support the zigzag model. Unlike the solenoid model, the zigzag conformation requires the linker DNA to pass through the central axis of the fiber in a relatively straight form (see Fig. 8-30b). Thus, longer linker DNA favors this conformation. Because the average linker DNA varies between different species (see Table 8-4), the form of the 30-nm fiber may not always be the same, and both forms of the 30-nm fiber may be found in cells depending on the local linker DNA length.



**FIGURE 8-30** Two models for the 30-nm chromatin fiber. In each panel, the left-hand view shows the side of the fiber, and the right-hand view shows a view down the central axis of the fiber. (a) The solenoid model. Note that the linker DNA does not pass through the central axis of the superhelix and that the sides and entry and exit points of the nucleosomes are relatively inaccessible. (b) The “zigzag” model. In this model, the linker DNA frequently passes through the central axis of the fiber, and the sides and even the entry and exit points are more accessible. (Reproduced, with permission, from Pollard T. and Earnshaw W. 2002. *Cell biology*, 1st ed., Fig. 13-6. © Elsevier.)



**FIGURE 8-31** Speculative model for the stabilization of the 30-nm fiber by histone amino-terminal tails. In this model, the 30-nm fiber is illustrated using the zigzag model. Several different tail-histone core interactions are possible. Here, the interactions are shown as between every alternate histone, but they could also be with adjacent or more distant histones.

### The Histone Amino-Terminal Tails Are Required for the Formation of the 30-nm Fiber

Core histones lacking their amino-terminal tails are incapable of forming 30-nm fibers. The most likely role of the tails is to stabilize the 30-nm fiber by interacting with adjacent nucleosomes. This model is supported by the 3D crystal structure of the nucleosome, which shows that each of the amino-terminal tails of H2A, H3, and H4 interacts with adjacent nucleosome cores in the crystal lattice (Fig. 8-31). Recent studies indicate that the interaction between the positively charged amino terminus of histone H4 and a negatively charged region of the histone-fold domain of histone H2A is particularly important for 30-nm fiber formation. Supporting the importance of this interaction, the residues of H2A that interact with the H4 tail are conserved across many eukaryotic organisms but are not involved in DNA binding or formation of the histone octamer. One possibility is that these regions of H2A are conserved to mediate internucleosomal interactions with the H4 tail. As we shall see later, the histone tails are frequent targets for modification in the cell. It is likely that some of these modifications influence the ability to form the 30-nm fiber and other higher-order nucleosome structures.

### Further Compaction of DNA Involves Large Loops of Nucleosomal DNA

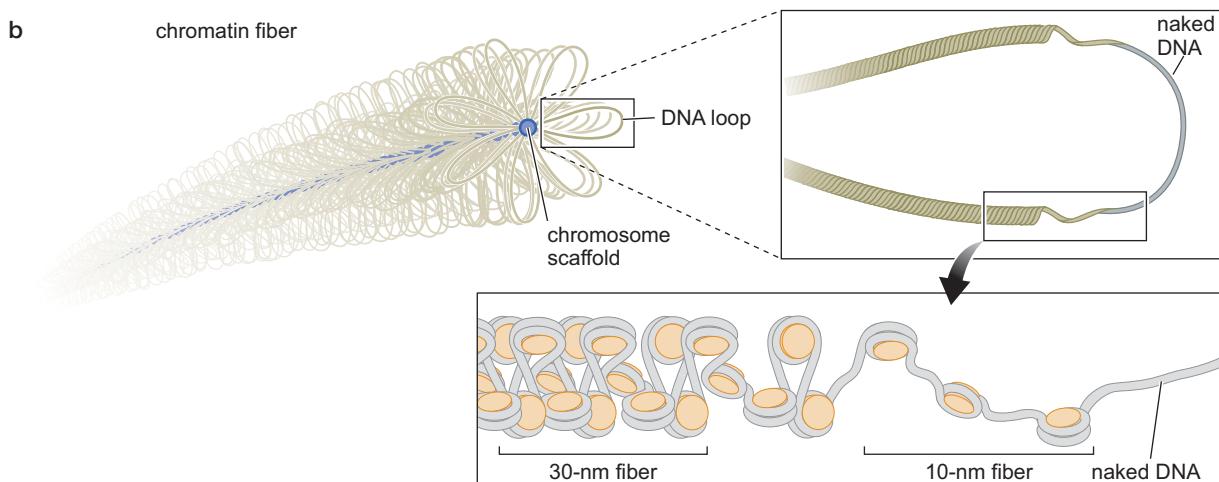
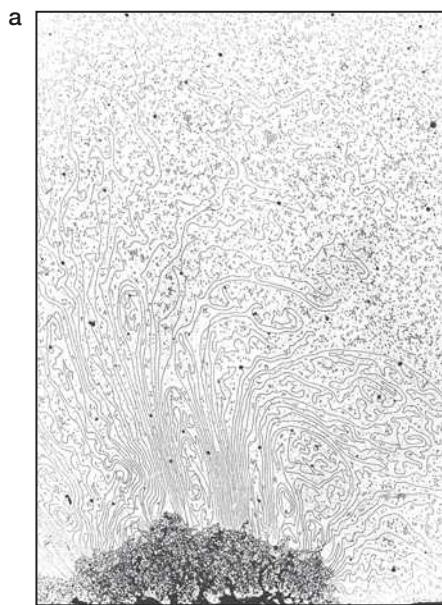
Together, the packaging of DNA into nucleosomes and the 30-nm fiber results in the compaction of the linear length of DNA by ~40-fold. This is still insufficient to fit 1–2 m of DNA into a nucleus  $\sim 10^{-5}$  m across. Additional folding of the 30-nm fiber is required to compact the DNA further. Although the exact nature of this folded structure remains unclear, one popular model proposes that the 30-nm fiber forms loops of 40–90 kb that are held together at their bases by a proteinaceous structure referred to as the **nuclear scaffold** (Fig. 8-32). A variety of methods have been developed to identify proteins that are part of this structure, although the true nature of the nuclear scaffold remains mysterious.

Two classes of proteins that contribute to the nuclear scaffold have been identified. One of these is topoisomerase II (Topo II), which is abundant in both scaffold preparations and purified mitotic chromosomes. Treating cells with drugs that result in DNA breaks at the sites of Topo II DNA binding generates DNA fragments that are ~50 kb in size. This is similar to the size range observed for limited nuclease digestion of chromosomes and suggests that Topo II may be part of the mechanism that holds the DNA at the base of these loops. In addition, the presence of Topo II at the bottom of each loop would ensure that the loops are topologically isolated from one another.

The SMC proteins are also abundant components of the nuclear scaffold. As we discussed above (see above section on Chromosome Duplication and Segregation), these proteins are key components of the machinery that condenses and holds sister chromatids together after chromosome duplication. The associations of these proteins with the nuclear scaffold may serve to enhance their functions by providing an underlying foundation for their interactions with chromosomal DNA.

### Histone Variants Alter Nucleosome Function

The core histones are among the most conserved eukaryotic proteins; therefore, the nucleosomes formed by these proteins are very similar in all

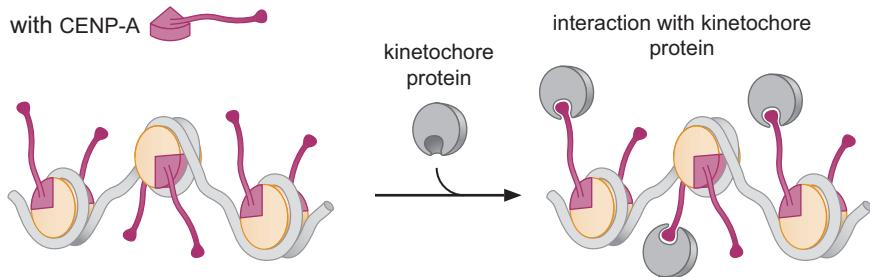


**FIGURE 8-32 Higher-order structure of chromatin.** (a) A transmission electron micrograph shows chromatin emerging from a central structure of a chromosome. The electron-dense regions are the nuclear scaffold that acts to organize the large amounts of DNA found in eukaryotic chromosomes. (b) A model for the structure of a eukaryotic chromosome shows that the majority of the DNA is packaged into large loops of 30-nm fiber that are tethered to the nuclear scaffold at their base. Sites of active DNA manipulation (e.g., sites of transcription or DNA replication) are in the form of 10-nm fiber or even naked DNA. (a, Courtesy of J.R. Paulson and U.K. Laemmli.)

eukaryotes. But there are numerous histone variants found in eukaryotic cells. Such unorthodox histones can replace one of the four standard histones to form alternate nucleosomes. Such nucleosomes may serve to demarcate particular regions of chromosomes or confer specialized functions to the nucleosomes into which they are incorporated. For example, H2A.X is a variant of H2A that is widely distributed in eukaryotic nucleosomes. When chromosomal DNA is broken (referred to as a double-strand break), H2A.X located adjacent to the break is phosphorylated at a serine residue that is not present in H2A. Phosphorylated H2A.X is specifically recognized by DNA repair enzymes leading to their localization at the site of DNA damage.

A second histone H3 variant, CENP-A, is associated with nucleosomes that include centromeric DNA. In this chromosomal region, CENP-A replaces the histone H3 subunits in nucleosomes. These nucleosomes are incorporated into the kinetochore that mediates attachment of the chromosome to the mitotic spindle (see Fig. 8-12). Compared with H3, CENP-A includes an extended amino-terminal tail region but has an otherwise similar histone-fold region. Thus, it is unlikely that incorporation of CENP-A changes the

**FIGURE 8-33** Alteration of chromatin by incorporation of histone variants. Incorporation of CENP-A in place of histone H3 is proposed to act as a binding site for one or more protein components of the kinetochore.



core structure of the nucleosome. Instead, the extended tail of CENP-A is a binding sites for another protein component of the kinetochore called CENP-C (Fig. 8-33). Consistent with this interaction being critical for kinetochore formation, loss of CENP-A interferes with the association of kinetochore components with centromeric DNA.

## REGULATION OF CHROMATIN STRUCTURE

### The Interaction of DNA with the Histone Octamer Is Dynamic

As discussed in detail in Chapter 19, the incorporation of DNA into nucleosomes can have a profound impact on the expression of the genome. In many instances, it is critical that nucleosomes can be moved or that their grip on the DNA can be loosened to allow other proteins access to the DNA. Consistent with this requirement, the association of the histone octamer with the DNA is inherently dynamic. In addition, there are factors that act on the nucleosome to increase or decrease the dynamic nature of this association. Together, these properties allow changes in nucleosome position and DNA association in response to the frequently changing needs for DNA accessibility.

Like all interactions mediated by noncovalent bonds, the association of any particular region of DNA with the histone octamer is not permanent: any individual region of the DNA will transiently be released from tight interaction with the octamer now and then. This release is analogous to the occasional opening of the DNA double helix (as we discussed in Chapter 4). The dynamic nature of DNA binding to the histone core structure is important, because many DNA-binding proteins strongly prefer to bind to histone-free DNA. Such proteins can recognize their binding site only when it is released from the histone octamer or is contained in linker or nucleosome-free DNA.

As a result of intermittent, spontaneous unwrapping of DNA from the nucleosome, the DNA-binding site for a given protein will be released from the histone octamer with a probability of 1 in 50 to 1 in 100,000, depending on where the binding site is within the nucleosome. The more central the binding site, the less frequently it is accessible. Thus, a binding site near position 73 of the 147 bp tightly associated with a nucleosome is rarely accessible, whereas binding sites near the ends (position 1 or 147) of the nucleosomal DNA are most frequently accessible. These findings indicate that the mechanism of exposure is due to unwrapping of the DNA from the nucleosome, rather than to the DNA briefly coming off the surface of the histone octamer (Fig. 8-34). It is important to note that these studies were performed on a population of individual nucleosomes in a test tube: the ability of DNA to unwrap from the nucleosome may be different for the large

---

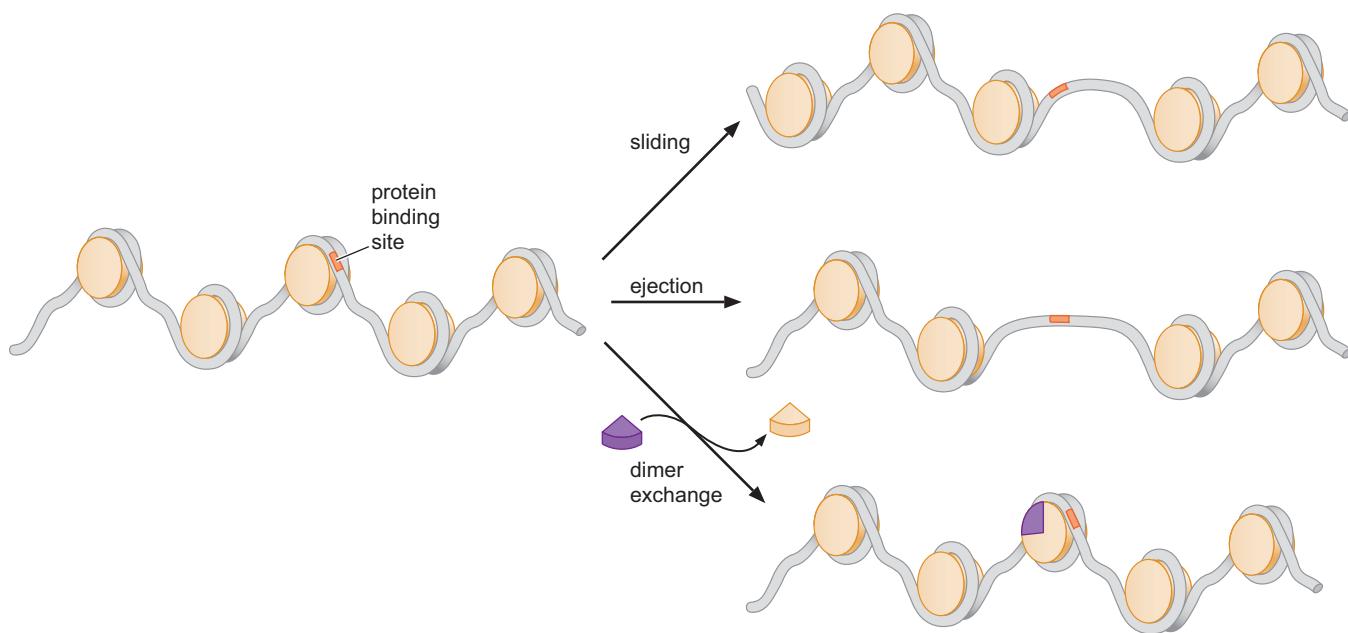
frequent event

stretches of DNA participating in many adjacent nucleosomes (called **nucleosome arrays**) present in cells. Association of H1 and incorporation of nucleosomes into the 30-nm fiber will also alter these probabilities. Nevertheless, the dynamic nature of nucleosome structure indicates that nucleosomes only look like the structure revealed in the X-ray crystallography studies for short periods of time and instead spend much of their time in other conformations.

### Nucleosome-Remodeling Complexes Facilitate Nucleosome Movement

In addition to the intrinsic dynamics shown by the nucleosome, the stability of the histone octamer–DNA interaction is influenced by large protein complexes called **nucleosome-remodeling complexes**. These multiprotein complexes facilitate changes in nucleosome location or interaction with the DNA using the energy of ATP hydrolysis. There are three basic types of nucleosome changes mediated by these enzymes (Fig. 8-35). All nucleosome-remodeling complexes can catalyze the “**sliding**” of DNA along the surface of the histone octamer. A subset of nucleosome-remodeling complexes can catalyze a second, more extreme change in which a histone octamer is ejected into solution or “**transferred**” from one DNA helix to another. Finally, some of these enzymes can facilitate the exchange of the H2A/H2B dimer within a nucleosome with variants of the dimer (e.g., H2A.X/H2B exchanged for H2A/H2B at double-strand breaks).

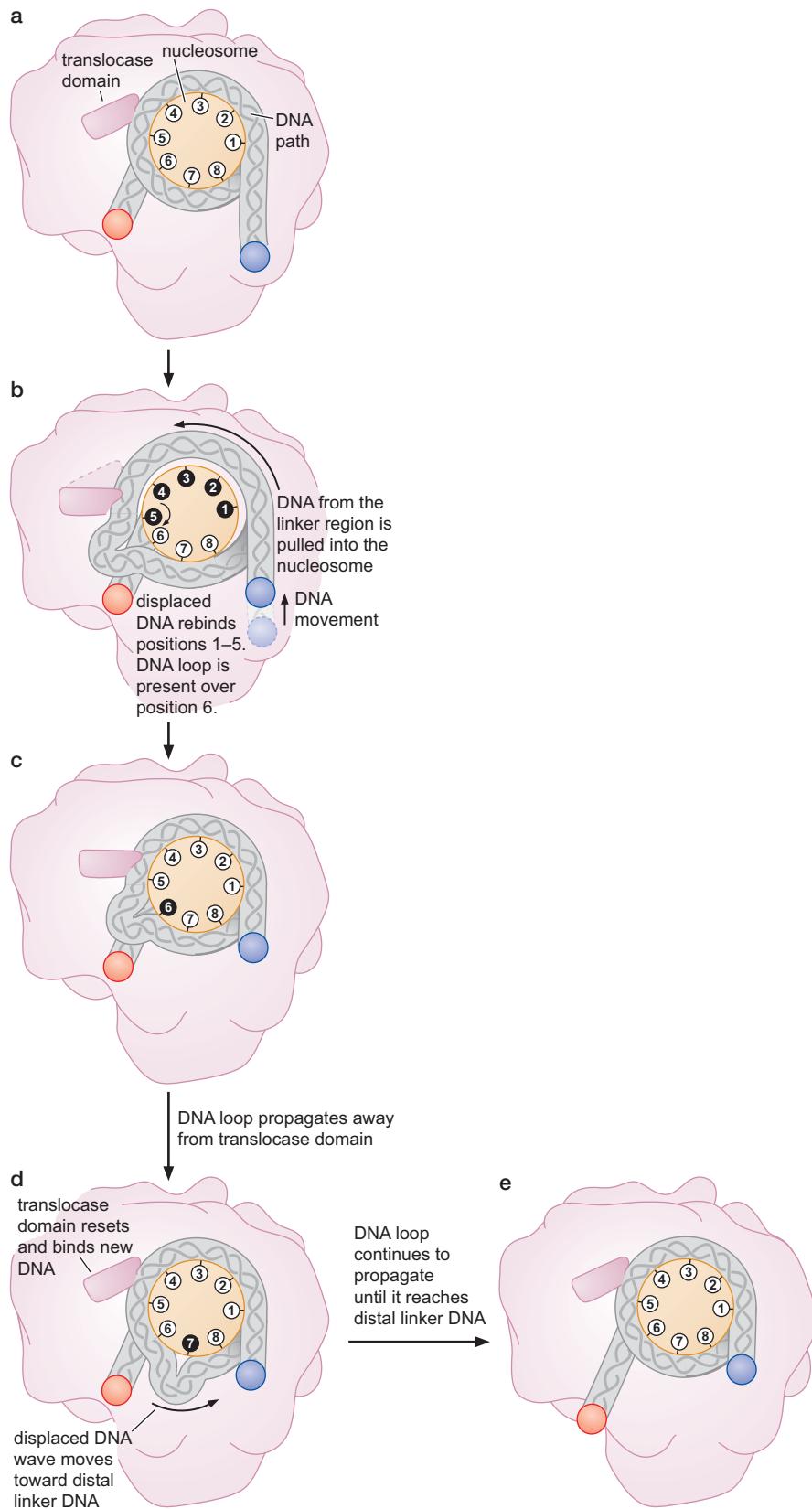
Recent studies have begun to reveal how nucleosome-remodeling complexes move DNA on the surface of the histone octamer (Fig. 8-36). Each of these multi-subunit enzymes contains an ATP-hydrolyzing DNA translocase subunit that is capable of moving in a directional manner (also called *translocating*) on double-stranded DNA when separated from the rest of the nucleosome-remodeling complex. Current models suggest that nucleosome-remodeling complexes bind the histone octamer tightly and position the DNA translocase subunit adjacent to the nucleosomal DNA. By holding the translocase in place relative to the histone octamer, the net result of ATP hydrolysis by the nucleosome-remodeling complex is to move the DNA relative to the surface of the histone octamer. DNA translocation generates a loop of DNA that is released from the surface of the nucleosome near the



**FIGURE 8-35** Nucleosome movement catalyzed by nucleosome-remodeling activities. (Top) Nucleosome movement by sliding along a DNA molecule exposes sites for DNA-binding proteins. (Middle) Nucleosome-remodeling complexes can also eject a nucleosome from the DNA creating larger nucleosome-free regions of DNA. (Bottom) A subset of nucleosome-remodeling complexes catalyzes the exchange of H2A/H2B dimers with either unmodified or variant H2A/H2B dimers (e.g., H2A-X).

site of the translocation. This loop is proposed to propagate on the surface of the histone octamer until it reaches the other end of the nucleosomal DNA. Although this loop movement could potentially proceed in either direction, it is thought that other interactions between the nucleosome-remodeling complex and the nucleosomal DNA prevent propagation toward the proximal DNA linker (which would result in no change in nucleosome positioning). Importantly, this approach does not demand that all interactions between the histone octamer and nucleosomal DNA be broken simultaneously. Instead, the “inchworm-like” movement of the DNA on the surface of the histone octamer allows the majority of the histone-DNA interactions to be maintained throughout the remodeling process. It is important to keep in mind that different DNA sequences interact with the histone octamer with roughly equal affinities. Thus, a DNA molecule that is sliding across a histone octamer can be viewed as binding to the octamer in many different energetic equivalent states and the nucleosome-remodeling complex is allowing DNA to access these different states more easily.

There are multiple types of nucleosome-remodeling complexes in any given cell (Table 8-6). They can have as few as two subunits or more than 10 subunits. Each of these complexes contains a related ATP-hydrolyzing subunit that catalyzes the DNA movement described above and in Figure 8-36. Although the ATP-hydrolyzing subunit is similar among the different nucleosome-remodeling complexes, the other subunits associated with each complex modulate their function. For example, these complexes can include subunits that target them to particular chromosomal locations. In some instances, this targeting is mediated by interactions between subunits of the remodeling complex and DNA-bound transcription factors. In other instances, nucleosome-remodeling complexes are localized by subunits that bind to specific histone-tail modifications (via chromodomains or bromodomains, as we discuss later).



**FIGURE 8-36** A model for nucleosomal DNA sliding catalyzed by nucleosome-remodeling complexes. (a) The model proposes that a DNA translocating domain of the ATP-hydrolyzing subunit of the nucleosome-remodeling complex binds the nucleosomal DNA two helical turns from the central dyad (e.g., at position 52 out of the total of 147 bp associated with the nucleosome). Other subunits of the nucleosome-remodeling complex bind tightly to the histones. The illustration shows each of the contacts between the DNA and the histones from the dyad to the closest unbound DNA (one contact per helical turn, seven of the 14 total). (b) Using the ATP-dependent DNA translocating activity, the nucleosome-remodeling complex first pulls the DNA from the nearest linker domain into the nucleosome. This breaks the five histone–DNA contacts between the ATP-hydrolyzing subunit and the linker DNA (broken contacts are shown in black, intact contacts in white) and creates a loop of DNA on the opposite side of the translocase domain. (c) The broken contacts re-form with the translocated DNA (positions 1–5), leaving the loop of DNA next to the ATP-hydrolyzing subunit (disrupting the contacts at position 6). (d) To remove the loop of DNA, the model proposes that the loop moves like a “wave” across the surface of the histones, breaking one or two contacts at a time (first contact 6 and then 7, etc.) until all of the contacts have re-formed with the appropriate amount of DNA between them, at which point the excess DNA is no longer present within the histone-associated DNA and the nucleosome has shifted its position on the DNA. (e) After the loop of DNA has propagated to the distal linker, and the nucleosome has shifted its position on the DNA. (Adapted, with permission, from Saha A. et al. 2006. *Nat. Rev. Mol. Cell Biol.* 7: 437–447, Fig. 4a. © Macmillan.)

**TABLE 8-6 ATP-Dependent Nucleosome-Remodeling Complexes**

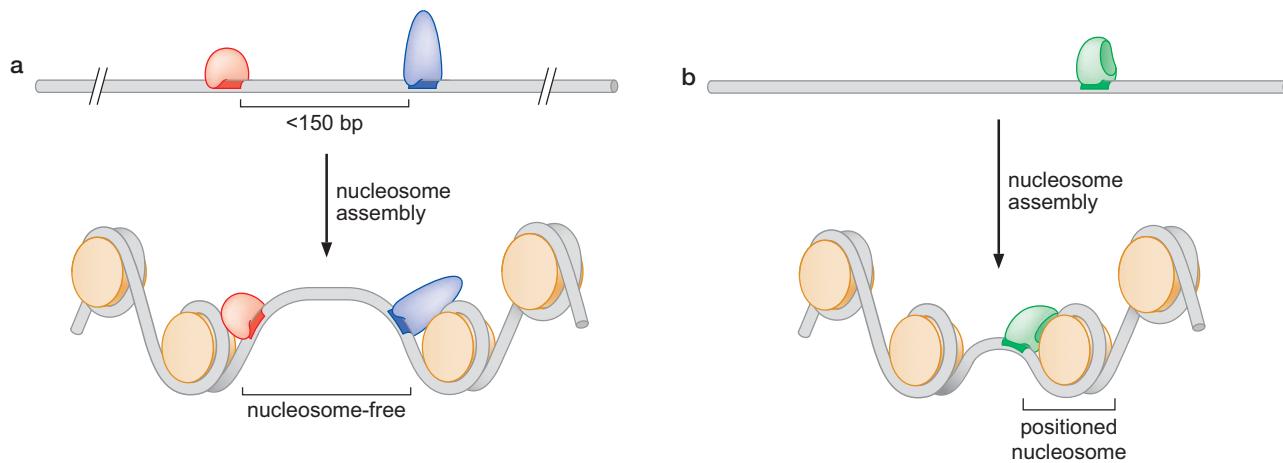
Type	Number of Subunits	Histone-Binding Domains	Slide	Transfer
SWI/SNF	8–14	Bromodomain	Yes	Yes
ISWI	2–4	Bromodomain, SANT domain, PHD finger	Yes	No
CHD	1–10	Chromodomain, PHD finger, SANT domain	Yes	Yes
INO80	10–16	Bromodomain	Yes	nd

nd, Not determined.

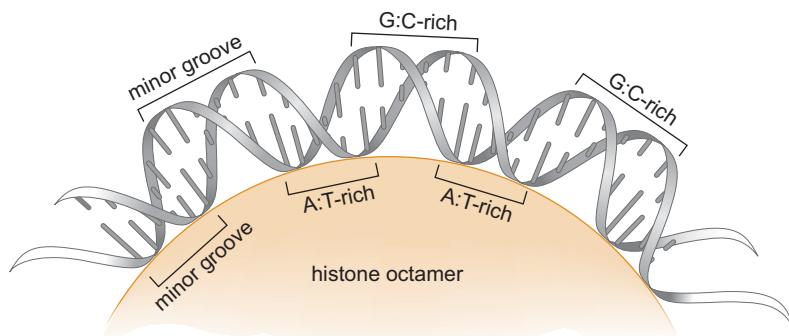
### Some Nucleosomes Are Found in Specific Positions: Nucleosome Positioning

Because of their sequence-nonspecific and dynamic interactions with DNA, most nucleosomes are not fixed in their locations. But there are occasions when restricting nucleosome location, or **positioning** nucleosomes as it is called, is beneficial. Typically, positioning a nucleosome allows the DNA-binding site for a regulatory protein to remain in the accessible linker DNA region. In many instances, such nucleosome-free regions are larger to allow the binding sites for multiple regulatory proteins to remain accessible. For example, the regions upstream of active transcription start sites are frequently associated with large nucleosome-free regions.

Nucleosome positioning can be directed by DNA-binding proteins or particular DNA sequences. In the cell, one frequent method involves competition between nucleosomes and DNA-binding proteins. Just as many proteins cannot bind to DNA within a nucleosome, binding of a protein to the DNA can prevent the subsequent association of the core histones with that stretch of DNA. If two such DNA-binding proteins are bound to sites closer than the minimal region of DNA required to assemble a nucleosome ( $\sim 150$  bp), the DNA between the proteins will remain nucleosome-free (Fig. 8-37a). Binding of additional proteins to adjacent DNA can further increase the size of a nucleosome-free region. In addition to this inhibitory mechanism of protein-



**FIGURE 8-37** Two modes of DNA-binding protein-dependent nucleosome positioning. (a) Association of many DNA-binding proteins with DNA is incompatible with the association of the same DNA with the histone octamer. Because a nucleosome requires more than 147 bp of DNA to form, if two such factors bind to the DNA less than this distance apart, the intervening DNA cannot assemble into a nucleosome. (b) A subset of DNA-binding proteins has the ability to bind to nucleosomes. Once bound to DNA, such proteins will facilitate the assembly of nucleosomes immediately adjacent to the protein's DNA-binding site.



**FIGURE 8-38** Nucleosomes prefer to bind bent DNA. Specific DNA sequences can position nucleosomes. Because the DNA is bent severely during association with the nucleosome, DNA sequences that position nucleosomes are intrinsically bent. A:T base pairs have an intrinsic tendency to bend toward the minor groove and G:C base pairs have the opposite tendency. Sequences that alternate between A:T- and G:C-rich sequences with a periodicity of ~5 bp will act as preferred nucleosome-binding sites. (Adapted, with permission, from Alberts B. et al. 2002. *Molecular biology of the cell*, 4th ed., Fig. 4-28. © Garland Science/Taylor & Francis LLC.)

dependent nucleosome positioning, some DNA-binding proteins interact tightly with adjacent nucleosomes, leading to nucleosomes *preferentially* assembling immediately adjacent to these proteins (Fig. 8-37b).

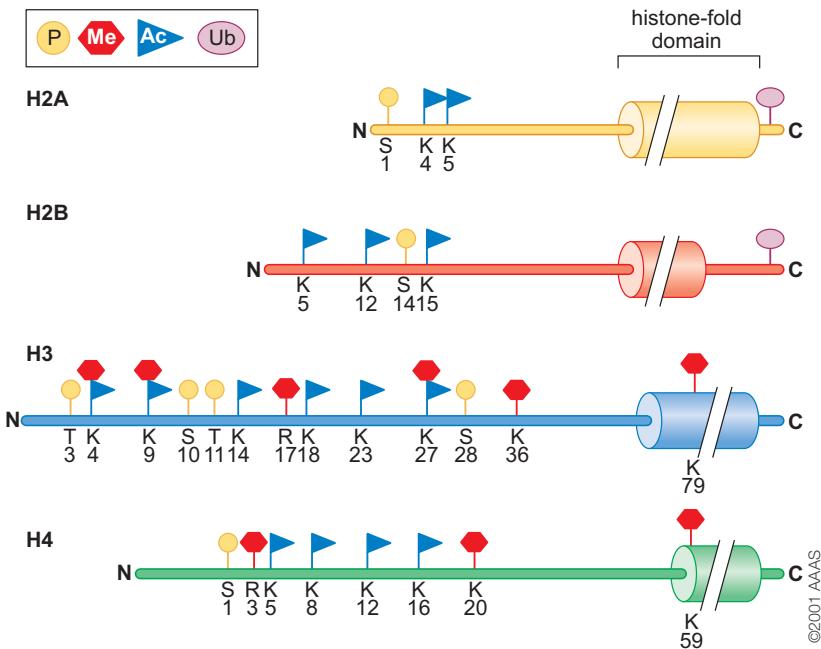
A second method of nucleosome positioning involves particular DNA sequences that have a high affinity for the nucleosome. Because DNA bound in a nucleosome is bent, nucleosomes preferentially form on DNA that bends easily. A:T-rich DNA has an intrinsic tendency to bend toward the minor groove. Thus, A:T-rich DNA is favored in positions in which the minor groove faces the histone octamer. G:C-rich DNA has the opposite tendency and is therefore favored when the minor groove is facing away from the histone octamer (Fig. 8-38). Each nucleosome will try to maximize this arrangement of A:T-rich and G:C-rich sequences. Recent studies of nucleosome positioning in the yeast *S. cerevisiae* suggest that as many as 50% of tightly positioned nucleosomes can be attributed to preferential binding of the histone core to the sequences they include. It is important to note that, despite being favored, such sequences are not required for nucleosome assembly, and the action of other proteins including chromatin-remodeling and transcription factors can move nucleosomes from such preferred positions.

These mechanisms of nucleosome positioning influence the organization of nucleosomes in the genome. Despite this, many nucleosomes are not tightly positioned. As discussed in the chapters on eukaryotic transcription (Chapters 13 and 19), tightly positioned nucleosomes are most often found at sites directing the initiation of transcription. Although we have discussed positioning primarily as a method to ensure that a regulatory DNA sequence is accessible, a positioned nucleosome can just as easily prevent access to specific DNA sites by being positioned in a manner that overlaps the same sequence. Thus, positioned nucleosomes can have either a positive or negative effect on the accessibility of nearby DNA sequences. An approach to mapping nucleosome locations is described in Box 8-3, Determining Nucleosome Position in the Cell.

### The Amino-Terminal Tails of the Histones Are Frequently Modified

When histones are isolated from cells, a subset of their amino-terminal tails is typically modified with a variety of small molecules (Fig. 8-39). Lysines in

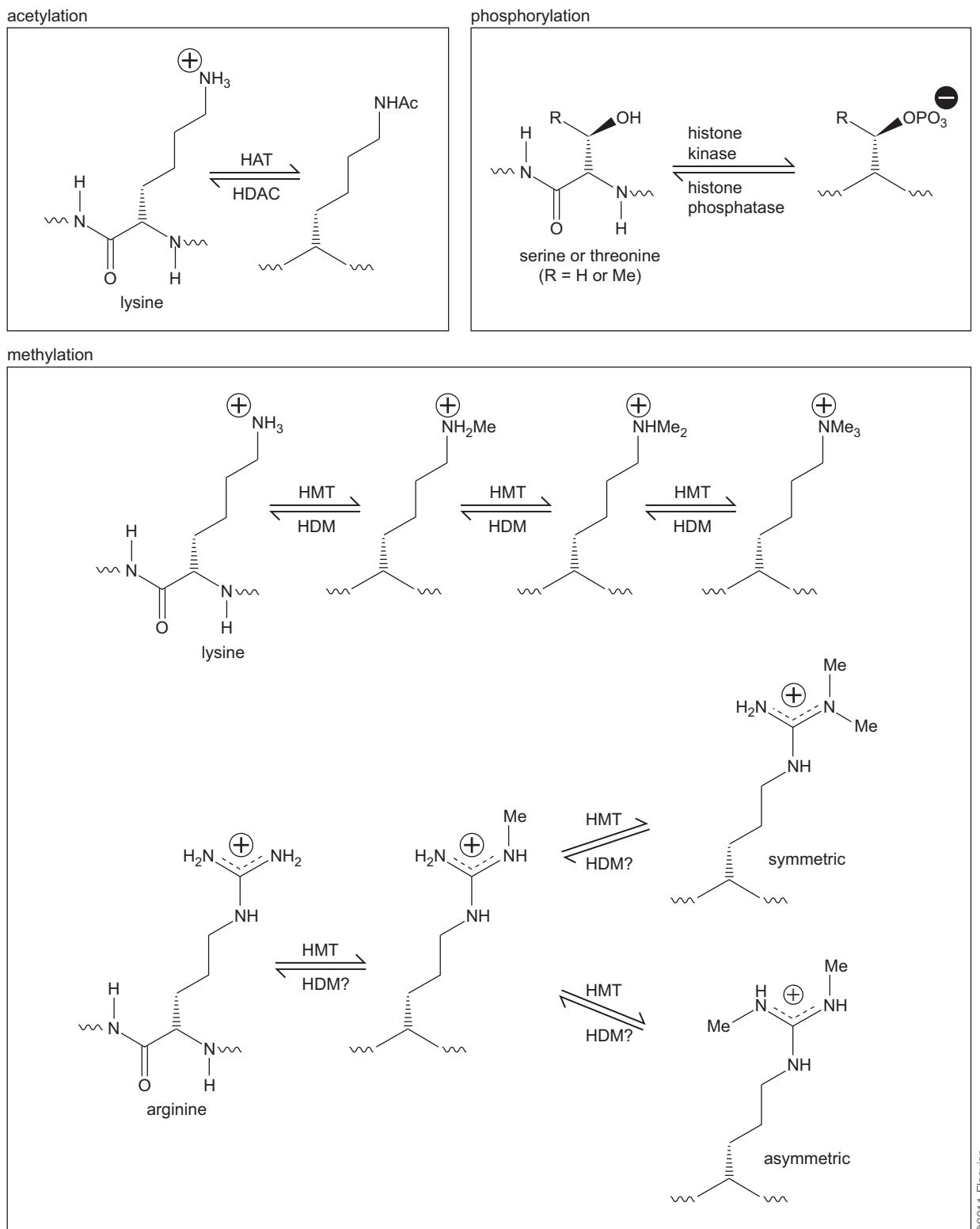
**FIGURE 8-39** Modifications of the histone amino-terminal tails alters the function of chromatin. The sites of known histone modifications are illustrated on each histone. Although the types of histone modifications continue to grow, for simplicity, only sites of acetylation, methylation, phosphorylation, and ubiquitylation are shown. The majority of these modifications occur on the tail regions, but there are occasional modifications within the histone fold (e.g., methylation of lysine 79 of histone H3). (Adapted, with permission, from Alberts B. et al. 2002. *Molecular biology of the cell*, 4th ed., Fig. 4-35. © Garland Science/Taylor & Francis LLC; and, with permission, from Jenuwein T. and Allis C.D. 2001. *Science* 293: 1074–1080, Figs. 2 and 3. © AAAS.)



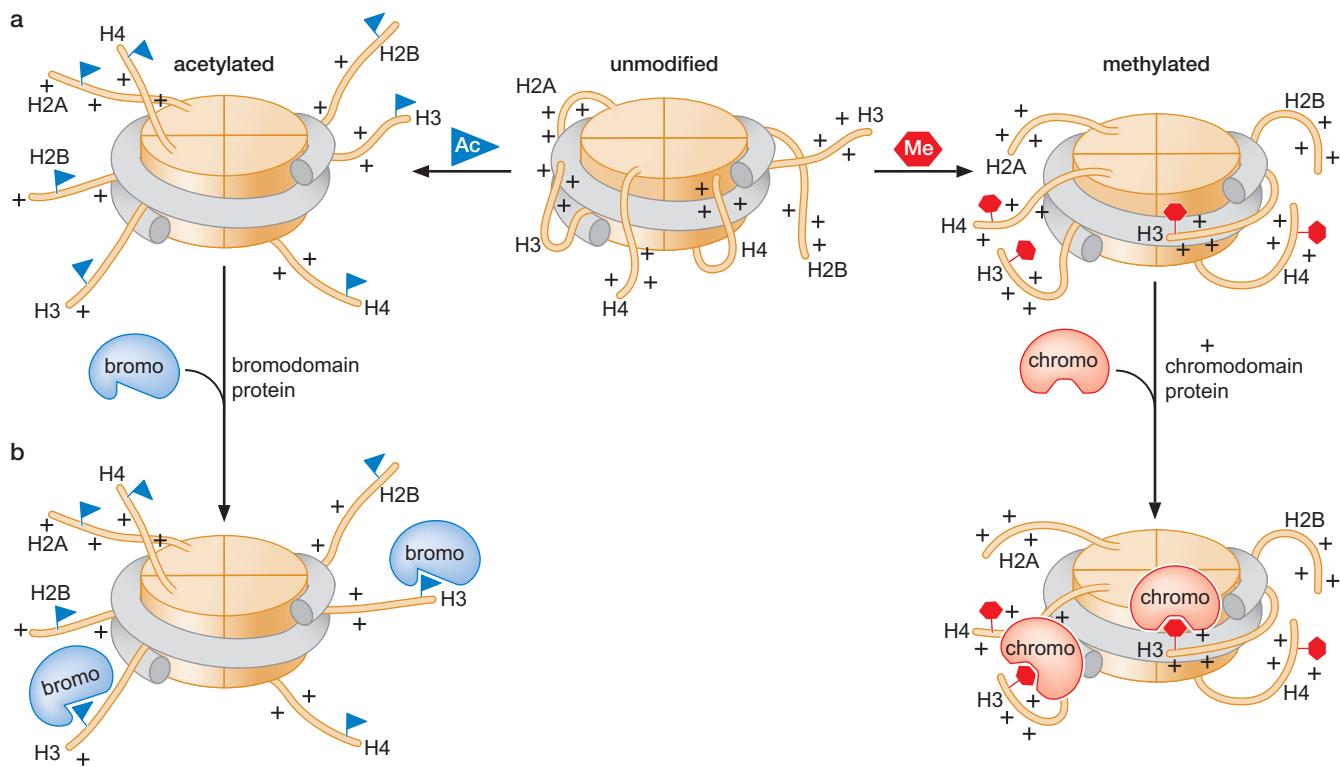
the tails are frequently modified with a single acetyl or methyl group, and arginines are found to be modified with one, two, or three methyl groups (Fig. 8-40). Similarly, serines and threonines (and one tyrosine) are subject to modification with phosphate. Although less common, other modifications with larger moieties including ADP-ribose and the small proteins ubiquitin and sumo are also found attached to histones.

Importantly, specific modifications are associated with histones involved in different cellular events. For example, acetylation of lysines at positions 8 and 16 of the histone H4 amino-terminal tail is associated with the start sites of expressed genes, but acetylation at lysines 5 and 12 is not. Instead, acetylation of these other lysines (5 and 12) marks newly synthesized H4 molecules that are ready to be deposited onto DNA as part of a new nucleosome. Similarly, methylation of lysines 4, 36, or 79 of histone H3 typically is associated with expressed genes, whereas methylation of lysines 9 or 27 of the same histone frequently is associated with transcriptional repression. The observation that particular histone modifications have a high probability of occurring at specific functional regions of chromatin (e.g., transcription start sites) has led to the hypothesis that histone tail modifications constitute a biological code that can be written, read, and erased by specific proteins in the cell. For a full discussion of this hypothesis, see Box 19-5.

How does histone modification alter nucleosome function? One obvious change is that acetylation and phosphorylation each acts to reduce the overall positive charge of the histone tails; acetylation of lysine neutralizes its positive charge (Fig. 8-41). This loss of positive charge reduces the affinity of the tails for the negatively charged backbone of the DNA. More importantly, modification of the histone tails affects the ability of nucleosome arrays to form more repressive higher-order chromatin structure. As we described above, histone amino-terminal tails are required to form the 30-nm fiber, and modification of the tails modulates this function. For example, consistent with the association of some types of acetylated histones with expressed regions of the genome, acetylation of the H4 amino-terminal tail interferes with the ability of nucleosomes to be incorporated into the



**FIGURE 8-40** Structure of histone tail modifications. The molecular structure of the small molecule histone modifications and the class of enzyme responsible are illustrated (histone acetyl transferase [HAT]; histone deacetylase [HDAC]; histone methyltransferase [HMT]; histone demethylase [HDM]). Only the affected amino acid is shown. Currently there is no known histone demethylase for arginine methylation, suggesting that these marks are only lost when the histone is removed from the DNA. (Adapted, with permission, from Lohse B. et al. 2011. *Bioorg. Med. Chem.* **19:** 3625–3636, Fig. 1. © Elsevier.)



**FIGURE 8-41** Effects of histone tail modifications. (a) The effect on the association with nucleosome-bound DNA. Unmodified and methylated histone tails are thought to associate more tightly with nucleosomal DNA than acetylated histone tails. (b) Modification of histone tails creates binding sites for chromatin-modifying enzymes.

repressive 30-nm fiber. As we described above, formation of the 30-nm fiber is facilitated by an interaction between the positively charged H4 amino-terminal tail and the negatively charged surface of the H2A histone-fold domain. Acetylation interferes with this association by altering the charge of the H4 tail.

### Protein Domains in Nucleosome-Remodeling and -Modifying Complexes Recognize Modified Histones

Modified histone tails can also act to recruit specific proteins to the chromatin (Fig. 8-41b). Protein domains called **bromodomains**, **chromodomains**, **TUDOR domains**, and **PHD** (for plant homeodomain) **fingers** specifically recognize modified forms of histone tails. Bromodomain-containing proteins interact with acetylated histone tails, and chromodomain-TUDOR domains and PHD-finger-containing proteins interact with methylated histone tails. Yet another protein domain, called a **SANT domain**, has the opposite property. SANT-domain-containing proteins interact preferentially with unmodified histone tails. Consistent with these protein domains being important for interpreting histone modifications, in many instances proteins containing these domains specifically recognize the modified form of only one of the many possible sites of histone modification. For example, the protein HP1 contains a chromodomain that will bind to methylated lysine 9 of histone H3 but not to any other site of histone methylation. Intriguingly, there are proteins that include more than one of these domains,

suggesting that they are specialized for recognizing histone tails that are multiply modified. For example, there are proteins that contain a PHD finger specific for methylated lysine 4 of histone H3 immediately next to a bromo-domain capable of recognizing an acetylated lysine.

How do the domains that recognize modified histones alter the function of the associated nucleosomes? One important way is that modified histones recruit enzymes that will further modify adjacent nucleosomes. For example, many of the enzymes that acetylate histone tails (called histone acetyltransferases or HATs) include bromodomains that recognize the same histone modifications that they create (Table 8-7). In this case, the bromodomain facilitates the maintenance and propagation of acetylated histones by modifying nucleosomes that are adjacent to the already acetylated histones (as we shall discuss later).

Modified histones can also recruit other proteins that act on chromatin. Many nucleosome-remodeling complexes include one or more subunits with domains that recognize modified histones (see Table 8-6) allowing modified histones to recruit these enzymes. Several proteins involved in

## ► KEY EXPERIMENTS

### Box 8-3 Determining Nucleosome Position in the Cell

The significance of the location of nucleosomes adjacent to important regulatory sequences has led to the development of methods to monitor the location of nucleosomes in cells. Many of these methods exploit the ability of nucleosomes to protect DNA from digestion by micrococcal nuclease. As described in Box 8-1, micrococcal nuclease has a strong preference to cleave DNA between nucleosomes, rather than DNA tightly associated with nucleosomes. This property can be used to map nucleosomes that are associated with the same position throughout a cell population (Box 8-3 Fig. 1).

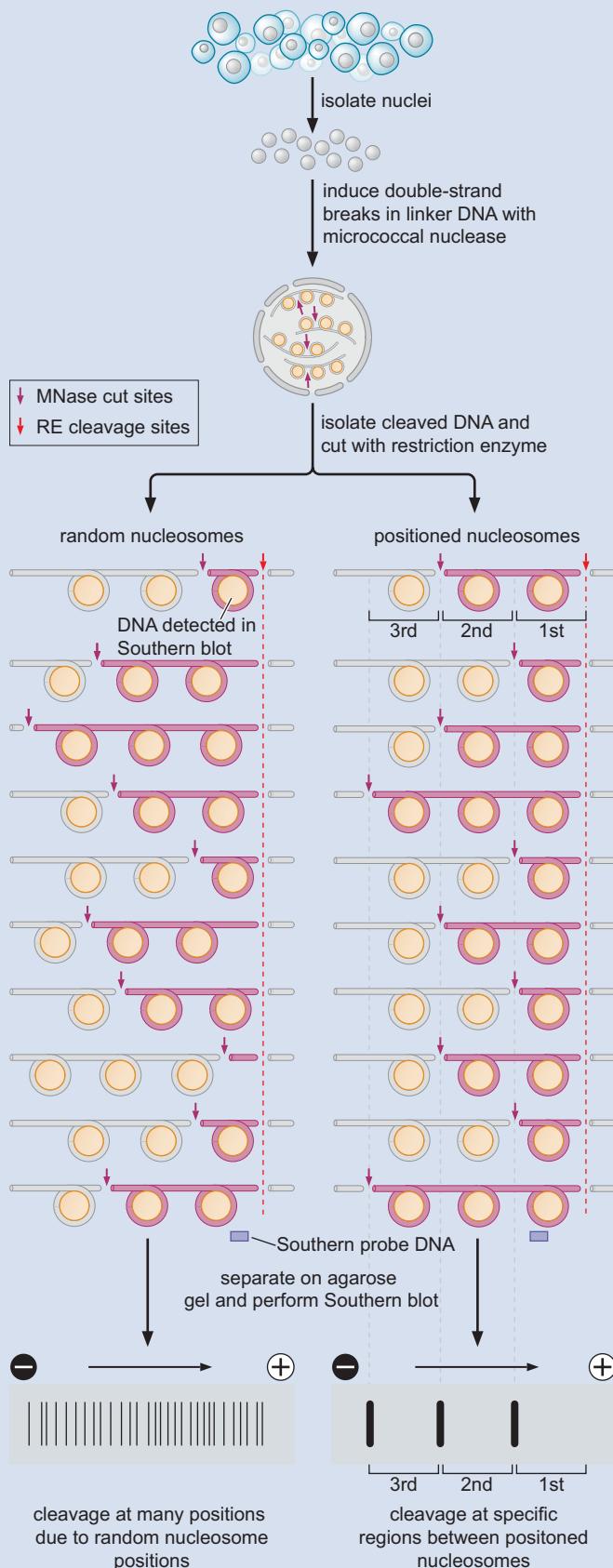
To map nucleosome location accurately, it is important to isolate the cellular chromatin and treat it with the appropriate amount of micrococcal nuclease with minimal disruption of the overall chromatin structure. This is typically achieved by gently lysing cells while leaving the nuclei intact. The nuclei are then briefly treated (typically for 1 min) with several different concentrations of micrococcal nuclease, a protein small enough to diffuse rapidly into the nucleus. The goal of the titration is for micrococcal nuclease to cleave the region of interest only once in each cell. Once the DNA has been digested, the nuclei can be lysed, and all of the protein can be removed from the DNA. The sites of cleavage (and, more importantly, the sites not cleaved) leave a record of the protein bound to DNA.

To identify the sites of cleavage in a particular region, it is necessary to create a defined end point for all of the cleaved fragments and exploit the specificity of DNA hybridization. To create a defined end point, the purified DNA from each sample is cut with a restriction enzyme known to cleave adjacent to the site of interest. After separation by size using agarose gel electrophoresis, the DNA is denatured and transferred to a nitrocellulose membrane such that its position in the gel is retained. A labeled DNA probe of specific sequence is then hybridized to the nitrocellulose-bound DNA (this is called a Southern blot

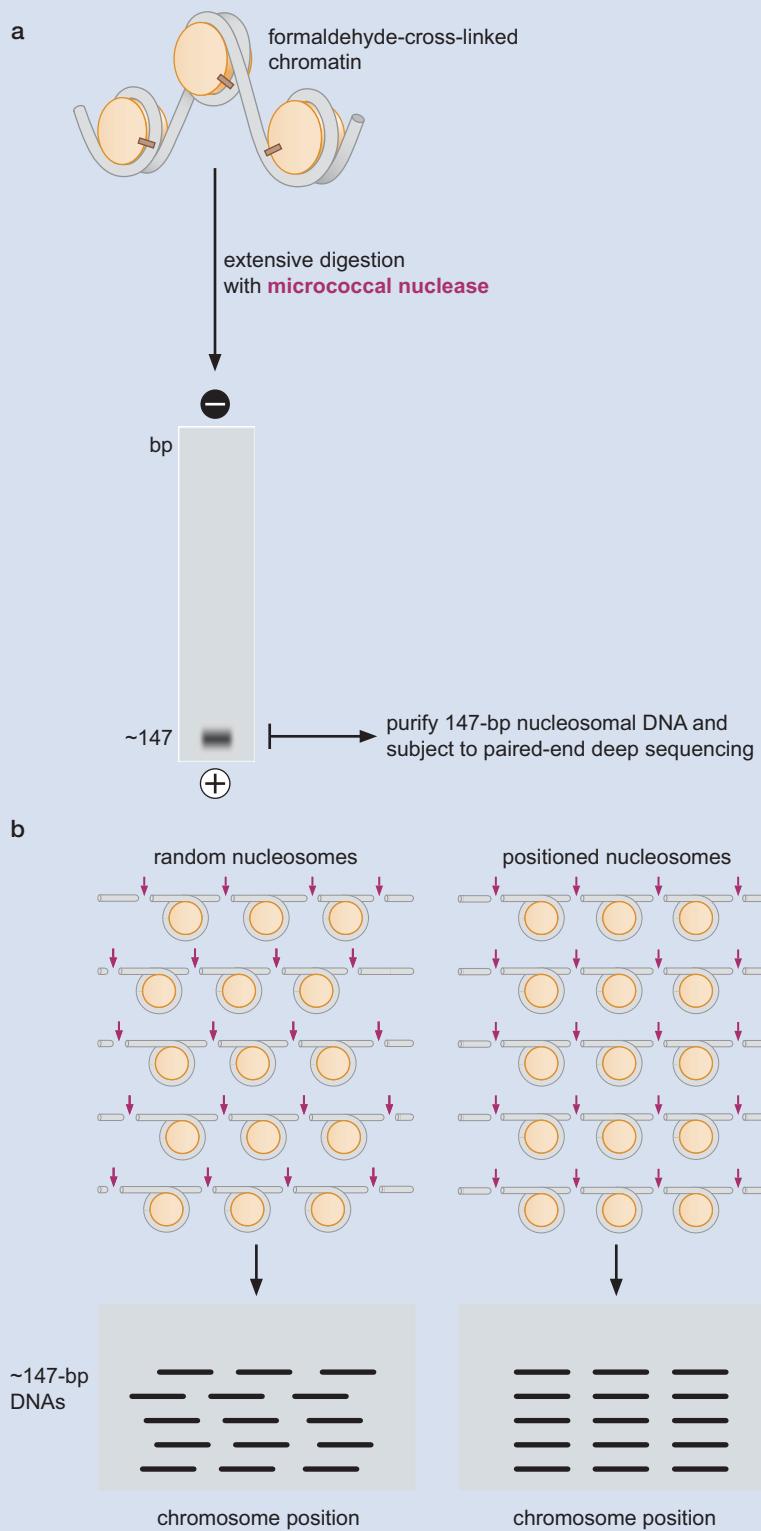
and is described in more detail in Chapter 7). In this case, the DNA probe is chosen to hybridize immediately adjacent to the restriction enzyme cleavage site at the site of interest. After hybridization and washing, the DNA probe will show the size of the fragments generated by micrococcal nuclease in the region of interest.

How do the fragment sizes reveal the location of positioned nucleosomes? DNA associated with positioned nucleosomes will be resistant to micrococcal nuclease digestion, leaving an ~160–200-bp region of DNA that is not cleaved. This will appear as a gap in the ladder of DNA bands detected on the Southern blot. The location of these gaps reveals the position of the nucleosomes adjacent to the restriction site/labeled DNA probe.

More recently, a related approach has been developed to identify positioned nucleosomes across entire genomes. This method starts by cross-linking histones to the DNA by treating the cells of interest with formaldehyde (Box 8-3 Fig. 2a). Next, the cells are lysed, and chromatin is isolated and treated with micrococcal nuclease until the majority of the DNA is the size of a mononucleosome (~147 bp). After reversing the cross-linking, the DNA is separated using gel electrophoresis, and the resulting 147-bp DNA fragments are purified and subjected to paired-end deep sequencing. This method of deep sequencing not only sequences both ends of each DNA fragment but also keeps track of which ends are from the same DNA fragment. Thus, paired-end sequencing reveals both the genomic location and the length of the sequenced DNA fragment. The nucleosome-sized DNA fragments sequenced reveal the location of a nucleosome. These locations can then be plotted along the length of each chromosome. The location of positioned nucleosomes is revealed by sites with many DNA fragments that are derived from the same 147-bp region (Box 8-3 Fig. 2b). Using this approach, all of the positioned nucleosomes across an entire genome can be mapped.

**Box 8-3** (Continued)

**BOX 8-3 FIGURE 1** Analysis of nucleosome positioning at a defined chromosomal position. The experimental steps in determining nucleosome positioning in the cell are illustrated. See box text for details.

**Box 8-3** (Continued)

**BOX 8-3 FIGURE 2** Genome-wide analysis of nucleosome positioning. (a) After cells are cross-linked with formaldehyde and chromatin is isolated, extensive treatment of the cross-linked chromatin with micrococcal nuclease results in the generation of predominantly nucleosome-core particles. Following the reversal of the cross-links, the predominant band of 147-bp DNA is isolated using gel electrophoresis and subjected to paired-end deep sequencing. (b) Illustration of the chromosomal mapping of nucleosome-associated DNAs at a site with random and positioned nucleosomes.

**TABLE 8-7** Histone-Modifying Enzymes

Histone Acetyltransferase Complexes				
Type	Number of Subunits	Catalytic Subunit	Histone-Binding Domains	Target Histones
SAGA	15	Gcn5	Bromodomain, chromodomain	H3 and H2B
PCAF	11	PCAF	Bromodomain	H3 and H4
NuA3	5	Sas3	PHD finger	H3
NuA4	6	Esa1	Chromodomain, SANT domain, PHD finger	H4 and H2A
P300/CBP	1	P300/CBP	Bromodomain, PHD finger	H2A, H2B, H3, and H4
Histone Deacetylase Complexes				
Type	Number of Subunits	Catalytic Subunit(s)	Histone-Binding Domains	
NuRD	9	HDAC1/HDAC2	Chromodomain, PHD finger	
SIR2 complex	3	Sir2	Neither	
Rpd3 large	12	Rpd3	PHD finger	
Rpd3 small	5	Rpd3	Chromodomain, PHD finger	
Histone Methyltransferases				
Name	Histone-Binding Domains			Target Histone
SET1	None			H3 (lysine 4)
SUV39/CLR4	Chromodomain			H3 (lysine 9)
SET2	None			H3 (lysine 36)
DOT1	None			H3 (lysine 79)
PRMT	None			H3 (arginine 3)
SET9/SUV4-20	None			H4 (lysine 20)
Histone Demethylases				
Name	Histone-Binding Domains			Target Methylated Histone
LSD1	PHD finger, SANT domain			H3 (lysine 4)
JHDM1	PHD finger			H3 (lysine 36)
JHDM3	PHD finger, TUDOR domain			H3 (lysines 9 and 36)

regulating transcription also include these domains. For example, a key component of the eukaryotic transcription machinery called TFIID includes a bromodomain. This domain directs the transcription machinery to sites of histone acetylation, which is an additional way that histone acetylation contributes to the increased transcriptional activity of the associated DNA. Chromodomains that recognize sites of histone methylation associated with transcriptionally repressed genes are found in several proteins that are important for the establishment of heterochromatin, including the HP1 protein and Polycomb proteins (see Chapters 19 and 21, respectively).

### Specific Enzymes Are Responsible for Histone Modification

The histone modifications we have just described are dynamic and are catalyzed by specific enzymes (Fig. 8-40). Histone acetyltransferases (HATs) catalyze the addition of acetyl groups to histones, whereas histone deacetylases (HDacs) remove these modifications. Similarly, histone methyltransferases (HMTs) add methyl groups to histones, and histone demethylases (HDMs) remove these modifications. A number of different histone acetyltransferases and deacetylases have been identified and are distinguished by their abilities

to target a different subset of histones or in some cases specific lysines in one histone. Histone methyltransferases and demethylases appear to be much more specific, always targeting only one of the many lysines or arginines on a specific histone (Table 8-7). Because these different modifications have different effects on nucleosome function, the modification of a nucleosome with different histone acetyltransferases or methyltransferases (or the removal of modifications by histone deacetylases or demethylases) can modulate chromatin structure and influence a wide array of DNA transactions.

Like their nucleosome-remodeling complex counterparts, these modifying enzymes are part of large multiprotein complexes. Additional subunits play important roles in recruiting these enzymes to specific regions of the DNA. Similar to the nucleosome-remodeling complexes, these interactions can be with transcription factors bound to DNA or directly with specifically modified nucleosomes. The recruitment of these enzymes to particular DNA regions is responsible for the distinct patterns of histone modification observed along the chromatin and is a major mechanism for modulating the levels of gene expression along the eukaryotic chromosome (see Chapter 19).

### Nucleosome Modification and Remodeling Work Together to Increase DNA Accessibility

The combination of amino-terminal tail modifications and nucleosome remodeling can dramatically change the accessibility of the DNA. As discussed in Chapters 13 and 19, the protein complexes involved in these modifications are frequently recruited to sites of active transcription. Although the order of their function is not always the same, the combined action can result in a profound, but localized, changes in DNA accessibility. Modification of amino-terminal tails can reduce the ability of nucleosome arrays to form repressive structures. This change creates sites that can recruit other proteins, including nucleosome remodelers. Remodeling of the nucleosomes can then further increase the accessibility of the nucleosomal DNA to allow DNA-binding proteins access to their binding sites. In combination with the appropriate DNA-binding proteins or DNA sequences, these changes can result in the positioning or release of nucleosomes at specific sites on the DNA (Fig. 8-42).

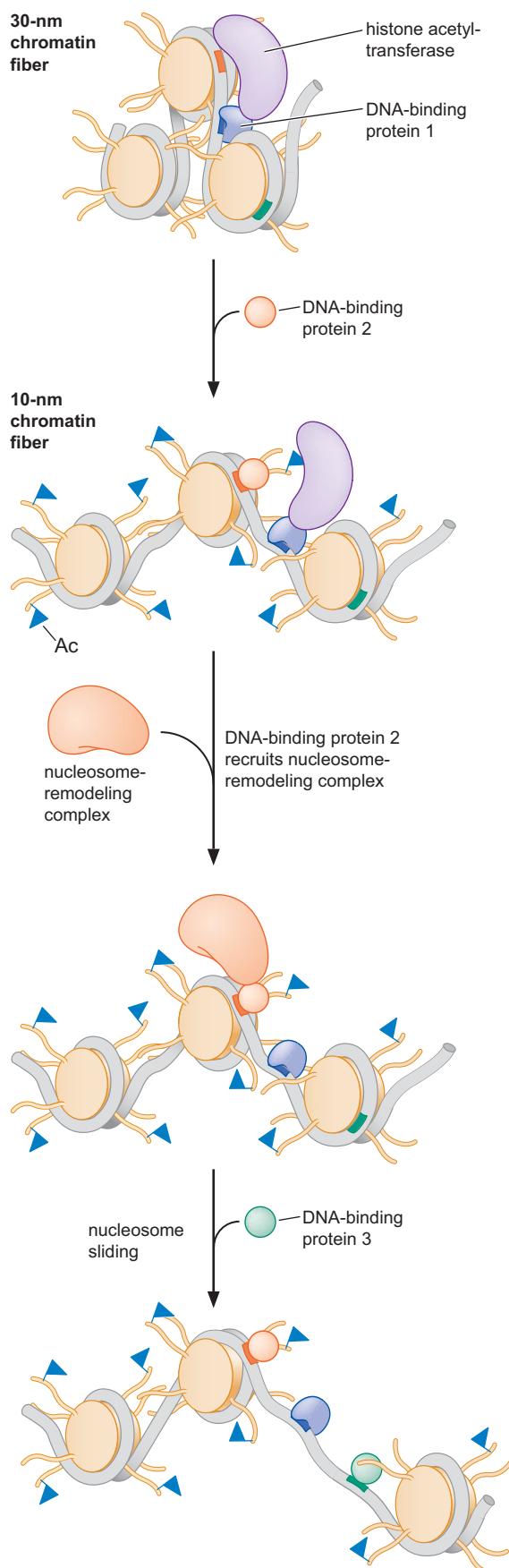
## NUCLEOSOME ASSEMBLY

### Nucleosomes Are Assembled Immediately after DNA Replication

The duplication of a chromosome requires replication of the DNA *and* the reassembly of the associated proteins on each daughter DNA molecule. The latter process is tightly linked to DNA replication to ensure that the newly replicated DNA is rapidly packaged into nucleosomes. In Chapter 9, we discuss the mechanisms of DNA replication in detail. Here, we discuss the mechanisms that direct the assembly of nucleosomes after the DNA is replicated (see Interactive Animation 8-2).

Although the replication of DNA requires the nucleosome disassembly, the DNA is rapidly repackaged into nucleosomes in an ordered series of events. As discussed above, the first step in the assembly of a nucleosome is the binding of an H3·H4 tetramer to the DNA. Once the tetramer is bound, two H2A·H2B dimers associate to form the final nucleosome. H1 joins this





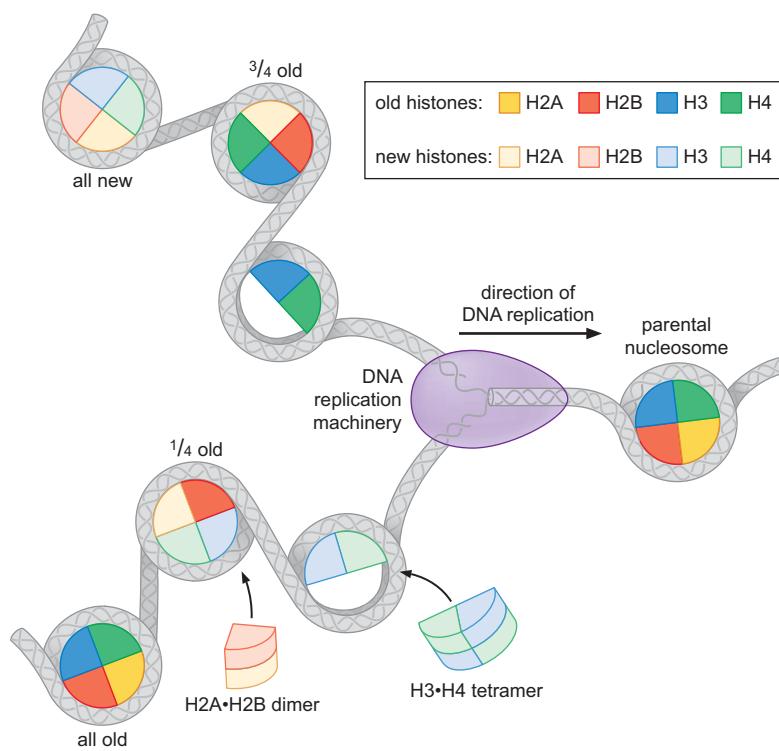
**FIGURE 8-42** Chromatin-remodeling and histone-modifying complexes work together to alter chromatin structure. Sequence-specific DNA-binding proteins typically recruit these enzymes to specific regions of a chromosome. In the illustration, the blue DNA-binding protein first recruits a histone acetyltransferase that modifies the adjacent nucleosomes, increasing the accessibility of the associated DNA by locally converting the chromatin fiber from the 30-nm fiber to the more accessible 10-nm form. This increased accessibility allows the binding of a second DNA-binding protein (orange) that recruits a nucleosome-remodeling complex. Localization of the nucleosome-remodeling complex facilitates the sliding of the adjacent nucleosomes, which allows the binding site for a third DNA-binding protein (green) to be exposed. For example, this could be the binding site for the TATA-binding protein at a start site of transcription. Although we show the order of association as histone acetylation complex and then nucleosome-remodeling complex, both orders are observed and can be equally effective. It is also true that recruitment of a different histone-modifying complex could result in the formation of more compact and inaccessible chromatin.

complex last, presumably during the formation of higher-order chromatin assemblies.

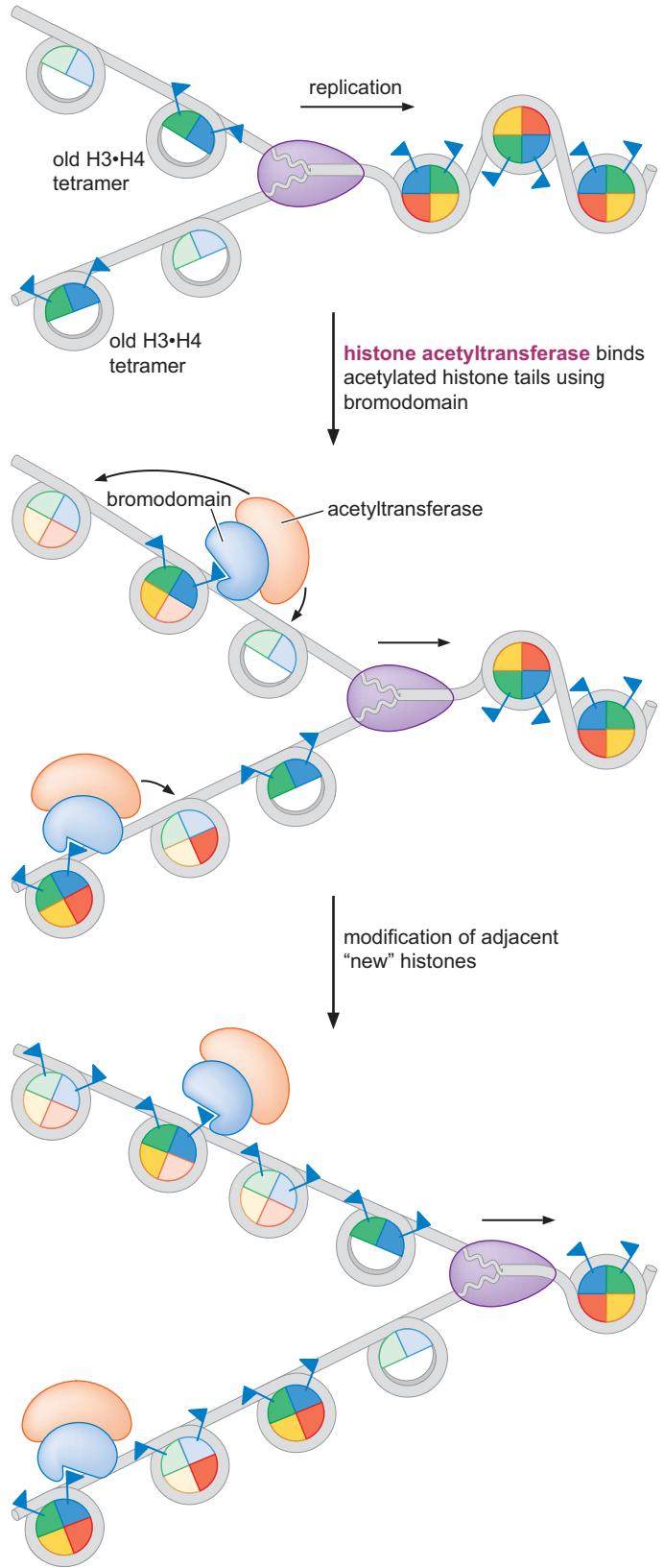
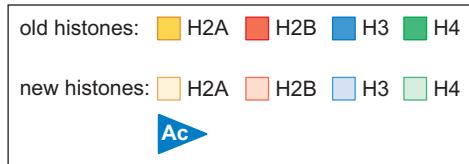
To duplicate a chromosome, at least half of the nucleosomes on the daughter chromosomes must be newly synthesized. Are all of the old histones lost and only new histones assembled into nucleosomes? If not, how are the old histones distributed between the two daughter chromosomes? The fate of the old histones is a particularly important issue given the effects that histone modification can have on the accessibility of the resulting chromatin. If the old histones were lost completely, then chromosome duplication would erase any “memory” of the previously modified nucleosomes. In contrast, if the old histones were retained on a single chromosome, that chromosome would have a distinct set of modifications relative to the other copy of the chromosome.

In experiments that differentially labeled old and new histones, it was found that the old histones are present on both of the daughter chromosomes (Fig. 8-43). Mixing is not entirely random, however. H3-H4 tetramers and H2A·H2B dimers are composed of either all new or all old histones. Thus, as the replication fork passes, nucleosomes are broken down into their component subassemblies. H3-H4 tetramers appear to remain bound to one of the two daughter duplexes at random and are never released from DNA into the free pool of histones. In contrast, the H2A·H2B dimers are released and enter the local pool, available for new nucleosome assembly.

The distributive inheritance of old histones during chromosome duplication provides a mechanism for the propagation of the parental pattern of histone modification. By this mechanism, old modified histones will tend to rebind one of the daughter chromosomes at a position near their previous position on the parental chromosome (Fig. 8-44). The old histones have an equal probability of binding either daughter chromosome. This localized inheritance of modified histones ensures that a subset of the modified histones is located in similar positions on each daughter chromosome.



**FIGURE 8-43** Inheritance of histones after DNA replication. As the chromosome is replicated, histones that were associated with the parental chromosome are differently distributed. The histone H3-H4 tetramers are randomly transferred to one of the two daughter strands but do not enter into the soluble pool of H3-H4 tetramers. Newly synthesized H3-H4 tetramers form the basis of the nucleosomes on the strand that does not inherit the parental tetramer. In contrast, H2A and H2B dimers are released into the soluble pool and compete for H3-H4 association with newly synthesized H2A and H2B. As a consequence of this type of distribution, on average, every second H3-H4 tetramer on newly synthesized DNA will be derived from the parental chromosome. These tetramers will include all of the modifications added to the parental nucleosomes. The H2A-H2B dimers are more likely to be derived from newly synthesized protein.



**FIGURE 8-44** Inheritance of parental H3-H4 tetramers facilitates the inheritance of chromatin states. As a chromosome is replicated, the distribution of the parental H3-H4 tetramers results in the daughter chromosomes receiving the same modifications as the parent. The ability of these modifications to recruit enzymes that perform the same modifications facilitates the propagation of the modification to the two daughter chromosomes. For simplicity, acetylation is shown on the core regions of the histones. In reality, this modification is generally on the amino-terminal tails.

The ability of these modified histones to recruit enzymes that add similar modifications to adjacent nucleosomes (see the discussion of modified histone-binding domains above) provides a simple mechanism to maintain states of modification after DNA replication has occurred (Fig. 8-44). Such mechanisms are likely to play a critical role in the inheritance of chromatin states from one generation to another. Given the importance of histone modification in controlling gene expression (see Chapter 19) as well as other DNA transactions, the maintenance of such modification states is critical to maintaining cell identity as cells replicate their DNA and divide.

### Assembly of Nucleosomes Requires Histone “Chaperones”

The assembly of nucleosomes is not a spontaneous process. Early studies found that the simple addition of purified histones to DNA resulted in little or no nucleosome formation. Instead, the majority of the histones aggregate in a nonproductive form. For correct nucleosome assembly, it was necessary to raise salt concentrations to very high levels ( $>1\text{ M NaCl}$ ) and then slowly reduce the concentration over many hours. Although useful for assembling nucleosomes for *in vitro* studies (such as for the structural studies of the nucleosome described above), elevated salt concentrations are not involved in nucleosome assembly *in vivo*.

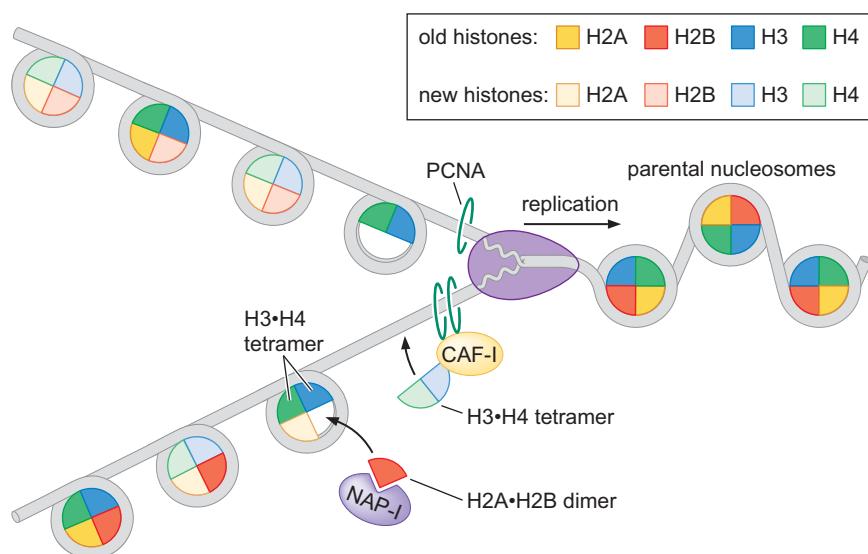
Studies of nucleosome assembly under physiological salt concentrations identified factors required to direct the assembly of histones onto the DNA. These factors are negatively charged proteins that form complexes with either H3·H4 or H2A·H2B dimers (see Table 8-8) and escort them to sites of nucleosome assembly. Because they act to keep histones from interacting with the DNA nonproductively, these factors have been referred to as **histone chaperones** (see Fig. 8-45).

How do the histone chaperones direct nucleosome assembly to sites of new DNA synthesis? Studies of the histone H3·H4 chaperone CAF-I reveal a likely answer. Nucleosome assembly directed by CAF-I requires that the target DNA be replicating. Thus, replicating DNA is marked in some way for nucleosome assembly. Interestingly, this mark is gradually lost after replication is completed. Studies of CAF-I-dependent assembly have determined that the mark is a ring-shaped sliding clamp protein called PCNA. As we discuss in detail in Chapter 9, this factor forms a ring around the DNA duplex and is responsible for holding DNA polymerase on the DNA during DNA synthesis. After the polymerase is finished, PCNA is released from the DNA polymerase but still encircles the DNA. In this condition, PCNA is available to interact with other proteins. CAF-I associates with the released PCNA and assembles H3·H4 preferentially on the PCNA-bound DNA. Thus, by associating with a component of the DNA replication machinery, CAF-I is directed to assemble nucleosomes at sites of recent DNA replication.

**TABLE 8-8** Properties of Histone Chaperones

Name	Number of Subunits	Histones Bound	Interaction with Sliding Clamp
CAF-1	4	H3·H4	Yes
HIRA	4	H3·H4	No
RCAF	1	H3·H4	No
NAP-1	1	H2A·H2B	No

**FIGURE 8-45 Chromatin assembly factors facilitate the assembly of nucleosomes.** After the replication fork has passed, chromatin assembly factors chaperone free H3-H4 tetramers (e.g., CAF-I) and H2A-H2B dimers (NAP-I) to the site of newly replicated DNA. Once at the newly replicated DNA, these factors transfer their associated histone to the DNA. CAF-I is recruited to the newly replicated DNA by interactions with DNA sliding clamps. These ring-shaped, auxiliary replication factors encircle the DNA and are released from the replication machinery as the replication fork moves. For a more detailed description of DNA sliding clamps and their function in DNA replication, see Chapter 9.



## SUMMARY

Within the cell, DNA is organized into large structures called chromosomes. Although the DNA forms the foundations for each chromosome, approximately half of each chromosome is composed of protein. Chromosomes can be either circular or linear; however, each cell has a characteristic number and composition of chromosomes. We now know the sequence of the entire genome of thousands of organisms. These sequences have revealed that the underlying DNA of each organism's chromosomes is used more or less efficiently to encode proteins. Simple organisms tend to use the majority of DNA to encode protein; however, more complex organisms use only a small portion of their DNA to encode proteins. The increased complexity of regulatory sequences, the appearance of introns, and the presence of additional regulatory RNAs (e.g., miRNAs) all contribute to the expansion of the non-coding regions of the genomes of more complex organisms.

Cells must carefully maintain their complement of chromosomes as they divide. Each chromosome must have DNA elements that direct chromosome maintenance during cell division. All chromosomes must have one or more origins of replication. In eukaryotic cells, centromeres play a critical role in the segregation of chromosomes, and telomeres help to protect and replicate the ends of linear chromosomes. Eukaryotic cells carefully separate the events that duplicate and segregate chromosomes as cell division proceeds. Chromosome segregation can occur in one of two ways. During mitosis, a highly specialized apparatus ensures that one copy of each duplicated chromosome is delivered to each daughter cell. During meiosis, an additional round of chromosome segregation (without DNA replication) reduces the number of chromosomes in the resulting daughter cells by half to generate haploid gametes.

The combination of eukaryotic DNA and its associated proteins is referred to as chromatin. The fundamental unit of chromatin is the nucleosome, which is made up of two copies each

of the core histones (H2A, H2B, H3, and H4) and ~147 bp of DNA. This protein-DNA complex serves two important functions in the cell: it compacts the DNA to allow it to fit into the nucleus, and it restricts the accessibility of the DNA. This latter function is extensively exploited by cells to regulate many different DNA transactions including gene expression.

The atomic structure of the nucleosome shows that the DNA is wrapped about 1.7 times around the outside of a disc-shaped, histone protein core. The interactions between the DNA and the histones are extensive but uniformly base-nonspecific. The nature of these interactions explains both the bending of the DNA around the histone octamer and the ability of virtually all DNA sequences to be incorporated into a nucleosome. This structure also reveals the location of the amino-terminal tails of the histones and their role in directing the path of the DNA around the histones.

Once DNA is packaged into nucleosomes, it has the ability to form more complex structures that further compact the DNA. This process is facilitated by a fifth histone called H1. By binding the DNA both within and adjacent to the nucleosome, H1 causes the DNA to wrap more tightly around the octamer. A more compact form of chromatin, the 30-nm fiber, is readily formed by arrays of H1-bound nucleosomes. This structure is more repressive than DNA packaged into nucleosomes alone. The incorporation of DNA into this structure results in a dramatic reduction in its accessibility to the enzymes and proteins involved in transcription of the DNA.

The interaction of the DNA with the histones in the nucleosome is dynamic, allowing DNA-binding proteins intermittent access to the DNA. Nucleosome-remodeling complexes increase the accessibility of DNA incorporated into nucleosomes by increasing the mobility of nucleosomes. Two forms of mobility can be observed: sliding of the histone octamer along the DNA or complete release of the histone octamer from the DNA. In addition, these complexes facilitate the

exchange of H2A/H2B dimers. Nucleosome-remodeling complexes are recruited to particular regions of the genome to facilitate alterations in chromatin accessibility. A subset of nucleosomes is restricted to fixed sites in the genome and is said to be “positioned.” Nucleosome positioning can be directed by DNA-binding proteins or particular DNA sequences.

Modification of the histone amino-terminal tails also alters the accessibility of chromatin. The types of modifications include acetylation and methylation of lysines, methylation of arginines, and phosphorylation of serines, threonines, and tyrosines. Acetylation of amino-terminal tails is frequently associated with regions of active gene expression and inhibits formation of the 30-nm fiber. Histone modifications alter the properties of the nucleosome itself, as well as acting as binding sites for proteins that influence the accessibility of the chromatin. In addition, these modifications

recruit enzymes that perform the same modification, leading to similar modification of adjacent nucleosomes and facilitating the stable propagation of regions of modified nucleosomes/chromatin as the chromosomes are duplicated.

Nucleosomes are assembled immediately after the DNA is replicated, leaving little time during which the DNA is unpackaged. Assembly involves the function of specialized histone chaperones that escort the H3·H4 tetramers and H2A·H2B dimers to the replication fork. During the replication of the DNA, nucleosomes are transiently disassembled. Histone H3·H4 tetramers and H2A·H2B dimers are randomly distributed to one or the other daughter molecule. On average, each new DNA molecule receives half old and half new histones. Thus, both chromosomes inherit modified histones that can then act as “seeds” for the similar modification of adjacent histones.

## BIBLIOGRAPHY

### Books

- Allis C.D., Jenuwein T., Reinberg D., and Caparros M.-L., eds. 2007. *Epigenetics*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
- Brown T.A. 2007. *Genomes 3*, 2nd ed. Garland Science, New York.
- Morgan D.O. 2007. *The cell cycle: Principles of control*. New Science Press Ltd., London.

### Chromosomes

- Bendich A.J. and Drlica K. 2000. Prokaryotic and eukaryotic chromosomes: What's the difference? *Bioessays* **22**: 481–486.
- Thanbichler M., Wang S.C., and Shapiro L. 2005. The bacterial nucleoid: A highly organized and dynamic structure. *J. Cell Biochem.* **96**: 506–521.

### Nucleosomes

- Clapier C.R. and Cairns B.R. 2009. The biology of chromatin remodeling complexes. *Annu. Rev. Biochem.* **78**: 273–304.

Gardner K.E., Allis C.D., and Strahl B.D. 2011. Operating on chromatin, a colorful language where context matters. *J. Mol. Biol.* **409**: 36–46.

Li G. and Reinberg D. 2011. Chromatin higher-order structures and gene regulation. *Curr. Opin. Genet. Dev.* **21**: 175–186.

Luger K., Madev A.W., and Richmond R.K. 1997. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**: 251–260.

Narlikar G.J., Fan H.-Y., and Kingston R.E. 2002. Cooperation between complexes that regulate chromatin structure and transcription. *Cell* **108**: 475–487.

Rando O. 2012. Combinatorial complexity in chromatin structure and function: Revisiting the histone code. *Curr. Opin. Genet. Dev.* **22**: 148–155.

Shahbazian M.D. and Grunstein M. 2007. Functions of site-specific histone acetylation and deacetylation. *Annu. Rev. Biochem.* **76**: 75–100.

Thiriet C. and Hayes J.J. 2005. Chromatin in need of a fix: Phosphorylation of H2AX connects chromatin to DNA repair. *Mol. Cell* **18**: 617–622.

## QUESTIONS

### MasteringBiology®

For instructor-assigned tutorials and problems, go to *MasteringBiology*.

For answers to even-numbered questions, see Appendix 2: Answers.

**Question 1.** List at least three properties that differ between the chromosome makeup in *E. coli* compared to human cells.

**Question 2.** Explain where the chromosomal DNA is located in prokaryotic versus eukaryotic cells.

**Question 3.** Does genome size correlate directly with organism complexity? Explain your reasoning.

**Question 4.** Intergenic sequences make up >60% of the human genome. Where do these intergenic sequences come from and what are some of their functions?

**Question 5.** Explain why each chromosome in a eukaryotic cell contains multiple origins of replication but includes one and only one centromere.

**Question 6.** How does the sister chromatid cohesion ensure that each daughter cell receives one copy of each chromosome?

**Question 7.** For a diploid human cell, state how many copies of each chromosome are present in each cell (or soon to be daughter cell).

Start of mitosis

End of mitosis

Start of meiosis

End of meiosis I

End of meiosis II

**Question 8.** For humans, what cells undergo mitosis? What cells undergo meiosis?

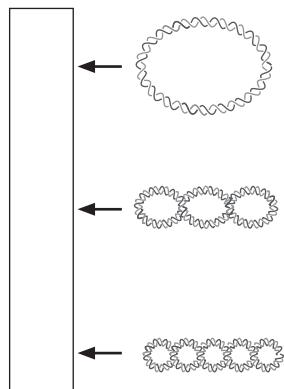
**Question 9.** Describe the components of a nucleosome.

**Question 10.** Name the types of bonds that occur between histone proteins and DNA and the region of DNA where these bonds form. Are these interactions sequence-specific? Explain why or why not.

**Question 11.** Explain why stored negative superhelicity from packaging DNA into nucleosomes is advantageous for cellular functions.

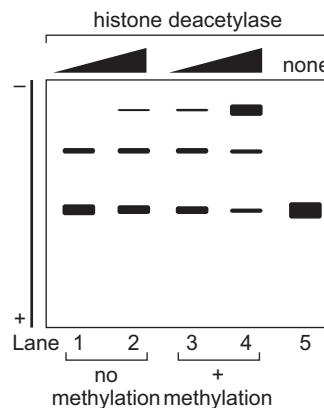
**Question 12.** Which protein domain(s) recognize the acetylation of histone amino-terminal tails? Which protein domain(s) recognize methylated histone amino-terminal tails?

**Question 13.** Review Box 8-2, Figure 1. For each of the DNAs described below, predict where the DNA would migrate on an agarose gel. Use the image of a gel below as a guide.



- A. Starting relaxed cccDNA as shown in Box 8-2, Figure 1a.
- B. Initiate nucleosome assembly without topoisomerase (as shown in Box 8-2, Fig. 1a), treat with detergent before running the products on an agarose gel.
- C. Add topoisomerase to the previous reaction but prevent additional nucleosome assembly as shown in Box 8-2, Figure 1b. Add detergent before running the products on an agarose gel.
- D. Add topoisomerase to the reaction described in B and allow additional nucleosome assembly as shown in Box 8-2, Figure 1c. Add detergent before running the products on an agarose gel.

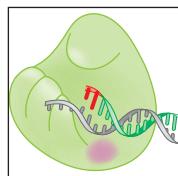
**Question 14.** You want to study the potential interaction between nucleosome-bound DNA and a specific histone deacetylase. You decide to perform an electrophoretic mobility shift assay (EMSA). For a review of this technique, see Chapter 7. You use a  $^{32}\text{P}$  end-labeled, linear template DNA that contains two nucleosome positioning sites. You assemble two nucleosomes on the DNA template before incubation with and without the histone deacetylase. For some reactions, you use unmodified nucleosomes. For other reactions, you use nucleosomes that are methylated at lysine 36 of the histone protein H3.



- A. Based on the data, propose a model for the interaction between the histone deacetylase and nucleosome-bound DNA.
- B. What type of protein domain do you predict allows the histone deacetylase to interact with the nucleosomes?

Adapted from Huh et al. (2012. *EMBO J.* **31**: 3564–3574).

CHAPTER 9



# The Replication of DNA

WHEN THE DNA DOUBLE HELIX WAS DISCOVERED, the feature that most excited biologists was the complementary relationship between the bases on its intertwined polynucleotide chains. It seemed unimaginable that such a complementary structure would not be used as the basis for DNA replication. In fact, it was the self-complementary nature revealed by the DNA structure that finally led most biologists to accept Oswald T. Avery's conclusion that DNA, not some form of protein, was the carrier of genetic information (Chapter 2).

In our discussion of how templates act (Chapter 4), we emphasized that two identical surfaces will not attract each other. Instead, it is much easier to visualize the attraction of groups with opposite shape or charge. Thus, without any detailed structural knowledge, we might guess that a molecule as complicated as the gene could not be copied directly. Instead, replication would involve the formation of a molecule complementary in shape, and this, in turn, would serve as a template to make a replica of the original molecule. Therefore, in the days before detailed knowledge of protein or nucleic acid structure, some geneticists wondered whether DNA served as a template for a specific protein that, in turn, served as a template for a corresponding DNA molecule.

But as soon as the self-complementary nature of DNA became known, the idea that protein templates might play a role in DNA replication was discarded. It was immensely simpler to postulate that each of the two strands of every parental DNA molecule served as a template for the formation of a complementary daughter strand. Although from the start this hypothesis seemed too good not to be true, experimental support nevertheless had to be generated. Happily, within 5 years of the discovery of the double helix, decisive evidence emerged for the separation of the complementary strands during DNA replication (see discussion of the Meselson and Stahl experiment in Chapter 2), and enzymological studies showed that DNA alone is the template for the synthesis of new DNA strands.

With these results, the problem of how genes replicate was in one sense solved. But in another sense, the study of DNA replication had only begun. How does DNA replication begin? How are the intertwined DNA strands separated so that they can act as template? What regulates the extent of replication so that daughter cells neither accumulate nor lose chromosomes? Study of these and other questions has revealed that the replication of even the simplest DNA molecule is a complex, multistep process, involving many more enzymes than was initially anticipated following the discovery of the first DNA polymerizing enzyme. The replication of the large, linear chromosomes of eukaryotes is still more challenging. These chromosomes require

## O U T L I N E

- The Chemistry of DNA Synthesis, 258
  - The Mechanism of DNA Polymerase, 260
    - The Replication Fork, 269
    - The Specialization of DNA Polymerases, 277
  - DNA Synthesis at the Replication Fork, 283
    - Initiation of DNA Replication, 288
    - Binding and Unwinding: Origin Selection and Activation by the Initiator Protein, 293
    - Finishing Replication, 302
  - Visit Web Content for Structural Tutorials and Interactive Animations

many start sites of replication to ensure that the entire chromosome is duplicated in a timely fashion, and the initiation of replication from these sites must be carefully coordinated to ensure that all sequences are replicated exactly once. Moreover, because conventional DNA replication cannot completely replicate the chromosome ends (called telomeres), cells have developed a novel method to maintain the integrity of this part of the chromosome.

In this chapter, we first describe the basic chemistry of DNA synthesis and the function of the enzymes that catalyze this reaction. We then discuss how the synthesis of DNA occurs in the context of an intact chromosome at structures called “replication forks.” We then focus on the initiation of DNA replication. DNA replication is tightly controlled in all cells and the initiation step is most highly regulated. We describe how replication initiation proteins unwind the DNA duplex at specific sites in the genome called “origins of replication” and how the replication fork proteins are recruited to these sites and assembled into the replisome. Finally, we describe how DNA replication is terminated and the special problems of replicating the ends of linear chromosomes. Together, the study of DNA replication reveals how multiple proteins come together to form a complex, multi-enzyme machine that performs this critical process with astounding speed, accuracy, and completeness.

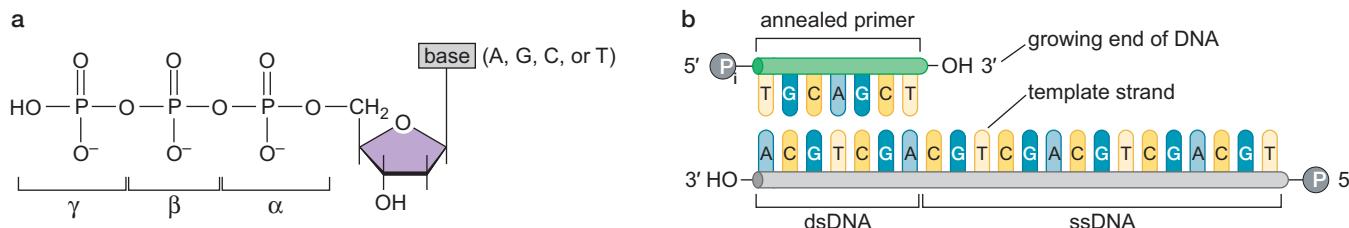
## THE CHEMISTRY OF DNA SYNTHESIS

### DNA Synthesis Requires Deoxynucleoside Triphosphates and a Primer:Template Junction



For the synthesis of DNA to proceed, two key substrates must be present (see Interactive Animation 9-1). First, new synthesis requires the four deoxynucleoside triphosphates—dGTP, dCTP, dATP, and dTTP (Fig. 9-1a). Nucleoside triphosphates have three phosphoryl groups that are attached via the 5'-hydroxyl of the 2'-deoxyribose. The phosphoryl group proximal to the deoxyribose is called the  $\alpha$ -phosphate, whereas the middle and distal groups are called the  $\beta$ -phosphate and the  $\gamma$ -phosphate, respectively.

The second essential substrate for DNA synthesis is a particular arrangement of single-stranded DNA (ssDNA) and double-stranded DNA (dsDNA) called a **primer:template junction** (Fig. 9-1b). As suggested by its name, the primer: template junction has two key components. The **template** provides the ssDNA that directs the addition of each complementary deoxynucleotide. The **primer** is complementary to, but shorter than, the template. The primer



**FIGURE 9-1** Substrates required for DNA synthesis. (a) The general structure of the 2'-deoxynucleoside triphosphates. The positions of the  $\alpha$ -phosphate,  $\beta$ -phosphate, and  $\gamma$ -phosphate are labeled. (b) The structure of a generalized primer:template junction. The shorter primer strand is completely annealed to the longer DNA strand and must have a free 3'-OH adjacent to an ssDNA region of the template. The longer DNA strand includes a region annealed to the primer and an adjacent ssDNA region that acts as the template for new DNA synthesis. New DNA synthesis extends the 3' end of the primer.

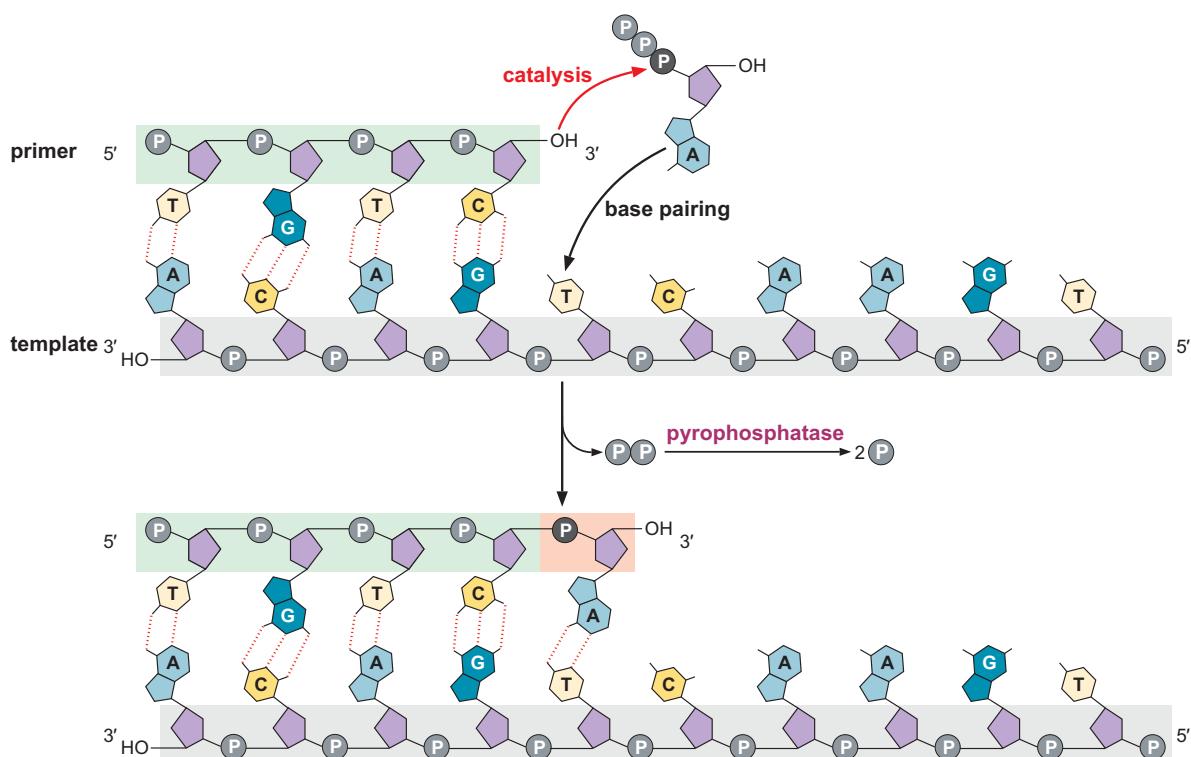
must have an exposed 3'-OH adjacent to the single-strand region of the template. It is this 3'-OH that will be extended by nucleotide addition.

Formally, only the primer portion of the primer:template junction is a substrate for DNA synthesis because only the primer is chemically modified during DNA synthesis. The template provides only the information necessary to select which nucleotides are added. Nevertheless, both a primer and a template are essential for all DNA synthesis.

### DNA Is Synthesized by Extending the 3' End of the Primer

The chemistry of DNA synthesis requires that the new chain grows by extending the 3' end of the primer (Fig. 9-2). Indeed, this is a feature of the synthesis of both RNA and DNA. The phosphodiester bond is formed in an  $S_N2$  reaction in which the hydroxyl group at the 3' end of the primer strand attacks the  $\alpha$ -phosphoryl group of the incoming nucleoside triphosphate. The leaving group for the reaction is pyrophosphate, which is composed of the  $\beta$ -phosphate and  $\gamma$ -phosphate of the nucleotide substrate.

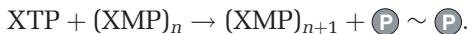
The template strand directs which of the four nucleoside triphosphates is added. The nucleoside triphosphate that base-pairs with the template strand is highly favored for addition to the primer strand. Recall that the two strands of the double helix have an antiparallel orientation. This arrangement means that the template strand for DNA synthesis has the opposite orientation of the growing DNA strand.



**FIGURE 9-2** Diagram of the mechanism of DNA synthesis. DNA synthesis is initiated when the 3'-OH of the primer mediates the nucleophilic attack of the  $\alpha$ -phosphate of the incoming dNTP. This results in the extension of the 3' end of the primer by one nucleotide and releases one molecule of pyrophosphate. Pyrophosphatase rapidly hydrolyzes released pyrophosphate into two phosphate molecules.

## Hydrolysis of Pyrophosphate Is the Driving Force for DNA Synthesis

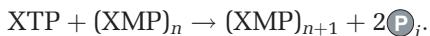
The addition of a nucleotide to a growing polynucleotide chain of length  $n$  is indicated by the following reaction:



But the free energy for this reaction is rather small ( $\Delta G = -3.5 \text{ kcal/mol}$ ). What, then, is the driving force for the polymerization of nucleotides into DNA? Additional free energy is provided by the rapid hydrolysis of the pyrophosphate into two phosphate groups by an enzyme known as pyrophosphatase:



The net result of nucleotide addition *and* pyrophosphate hydrolysis is the breaking of two high-energy phosphate bonds. Therefore, DNA synthesis is a coupled process, with an overall reaction of



This is a highly favorable reaction with a  $\Delta G$  of  $-7 \text{ kcal/mol}$ , which corresponds to an equilibrium constant ( $K_{\text{eq}}$ ) of  $\sim 10^5$ . Such a high  $K_{\text{eq}}$  means that the DNA synthesis reaction is effectively irreversible.

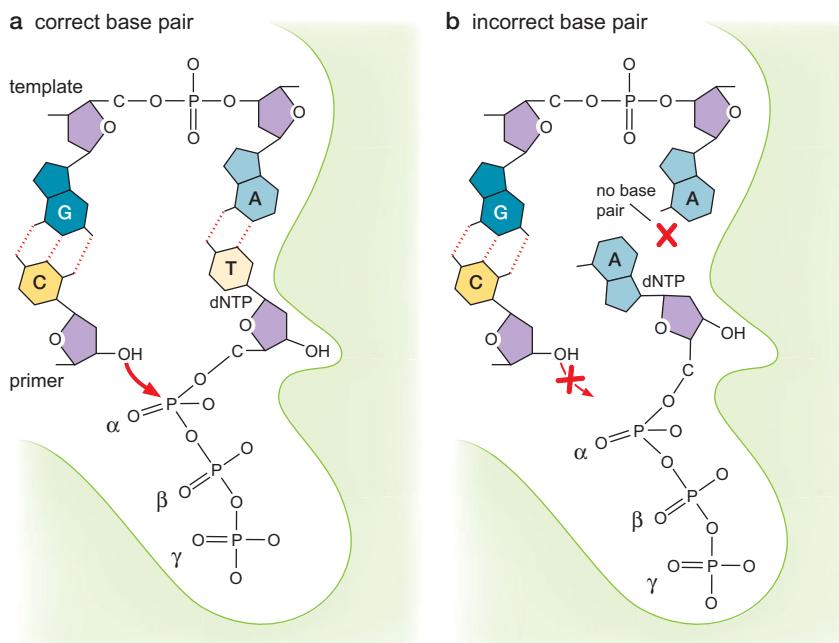
## THE MECHANISM OF DNA POLYMERASE

### DNA Polymerases Use a Single Active Site to Catalyze DNA Synthesis

The synthesis of DNA is catalyzed by a class of enzymes called **DNA polymerases**. Unlike most enzymes, which have one active site that catalyzes one reaction, DNA polymerase uses a single active site to catalyze the addition of any of the four deoxynucleoside triphosphates. DNA polymerase accomplishes this catalytic flexibility by exploiting the nearly identical geometry of the A:T and G:C base pairs (remember that the dimensions of the DNA helix are largely independent of the DNA sequence).

The DNA polymerase monitors the ability of the incoming nucleotide to form an A:T or G:C base pair, rather than detecting the exact nucleotide that enters the active site (Fig. 9-3). *Only* when a correct base pair is formed are the 3'-OH of the primer and the  $\alpha$ -phosphate of the incoming nucleoside triphosphate in the optimum position for catalysis to occur. Incorrect base pairing leads to dramatically lower rates of nucleotide addition as a result of a catalytically unfavorable alignment of these substrates (see Fig. 9-3b). This is an example of kinetic proofreading, in which an enzyme favors catalysis using one of several possible substrates by dramatically increasing the rate of bond formation only when the correct substrate is present. Indeed, the rate of incorporation of an incorrect nucleotide is as much as 10,000-fold slower than when base pairing is correct. A common method to monitor synthesis of new DNA is described in Box 9-1, Incorporation Assays Can Be Used to Measure Nucleic Acid and Protein Synthesis.

DNA polymerases show an impressive ability to distinguish between ribonucleoside and deoxyribonucleoside triphosphates (rNTPs and dNTPs). Although rNTPs are present at approximately 10-fold higher concentration in the cell, they are incorporated at a rate that is more than 1000-fold lower than dNTPs. This discrimination is mediated by the steric exclusion of



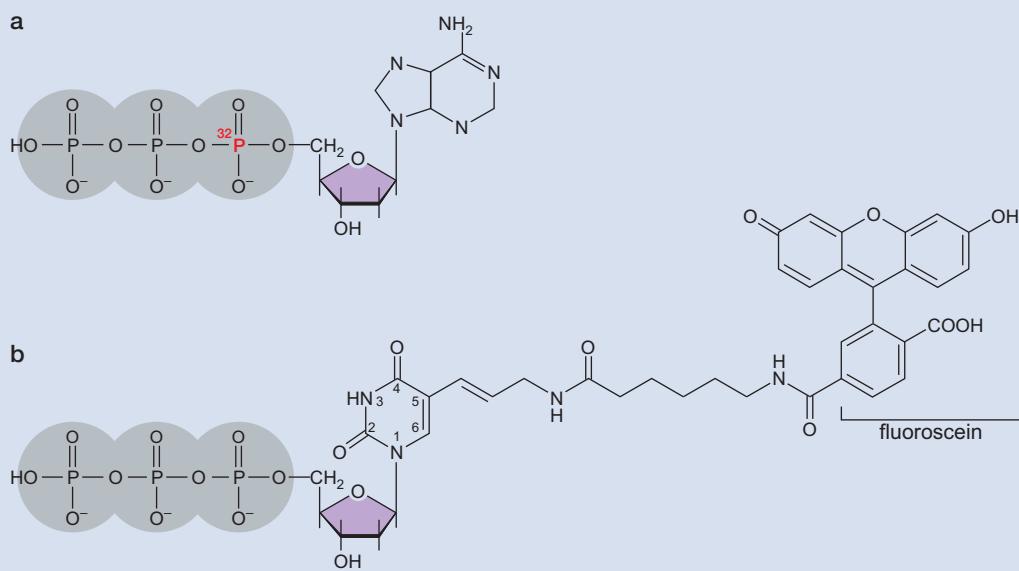
**FIGURE 9-3** Correctly paired bases are required for DNA-polymerase-catalyzed nucleotide addition. (a) Schematic diagram of the attack of a primer 3'-OH end on a correctly base-paired dNTP. (b) Schematic diagram of the consequence of incorrect base pairing on catalysis by DNA polymerase. In the example shown, the incorrect A:A base pair displaces the  $\alpha$ -phosphate of the incoming nucleotide. This incorrect alignment reduces the rate of catalysis dramatically, resulting in the DNA polymerase preferentially adding correctly base-paired dNTPs. (Adapted, with permission, from Brautigam C.A. and Steitz T.A. 1998. *Curr. Opin. Struct. Biol.* **8**: 54–63, Fig. 4d. © Elsevier.)

## ► TECHNIQUES

### Box 9-1 Incorporation Assays Can Be Used to Measure Nucleic Acid and Protein Synthesis

How can the activity of a DNA polymerase be measured? The simplest assay used to measure the synthesis of a polymer is an incorporation assay. In the case of DNA polymerase, this type of assay measures the incorporation of labeled dNTP precursors into DNA molecules. Typically, dNTPs are labeled by including

radioactive atoms in a part of the nucleotide that will be retained in the final DNA product (e.g., by replacing the phosphorous atom in the  $\alpha$ -phosphate with the radioactive isotope  $^{32}\text{P}$ ) (Box 9-1 Fig. 1a). Alternatively, nucleotides can be synthesized with fluorescent molecules in the place of the methyl group on



**BOX 9-1 FIGURE 1** Two forms of labeled deoxynucleoside triphosphates. (a)  $[\alpha\text{-}^{32}\text{P}]$ dATP. In this nucleotide, the  $\alpha$ -phosphorous is replaced with the radioactive isotope  $^{32}\text{P}$ . Note that only this phosphorous atom will become part of the DNA after nucleotide incorporation. (b) Fluorescently labeled thymidine triphosphate analog. In this labeled precursor, the fluorescent compound fluorescein has been attached via a linker to the 5 position of the thymine ring that is normally attached to a methyl group.

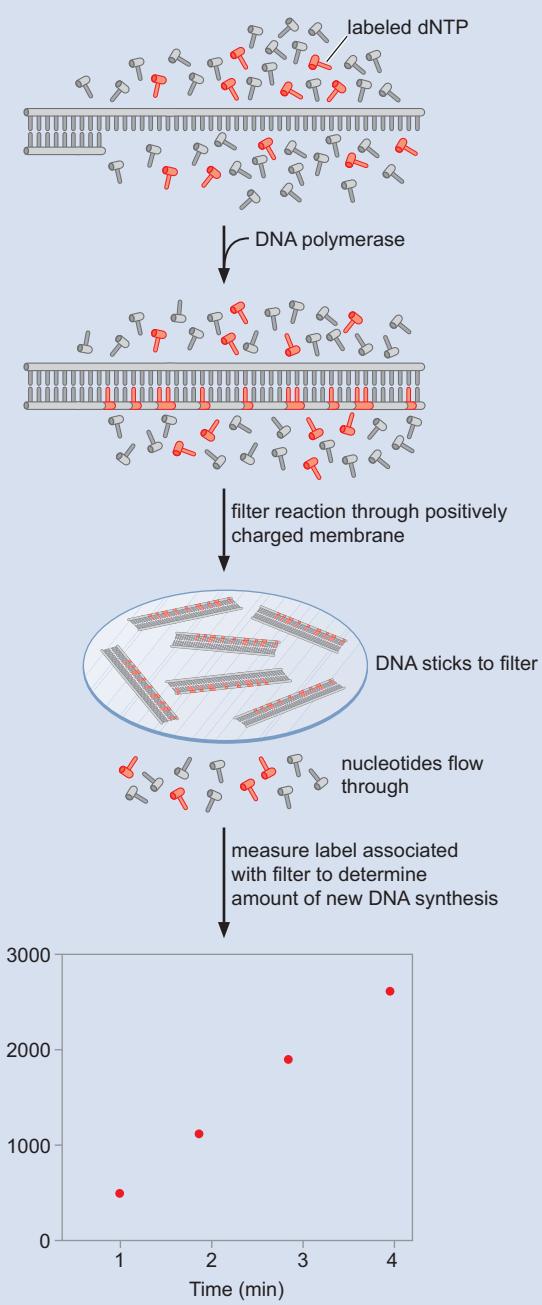
**Box 9-1 (Continued)**

dTTP (Box 9-1 Fig. 1b). This methyl group is not involved in base pairing, and DNA polymerases can readily accommodate much larger moieties in this location. In either case, these modifications allow easy monitoring of the labeled nucleotide using film or sensitive photomultipliers to detect emitted electrons or photons.

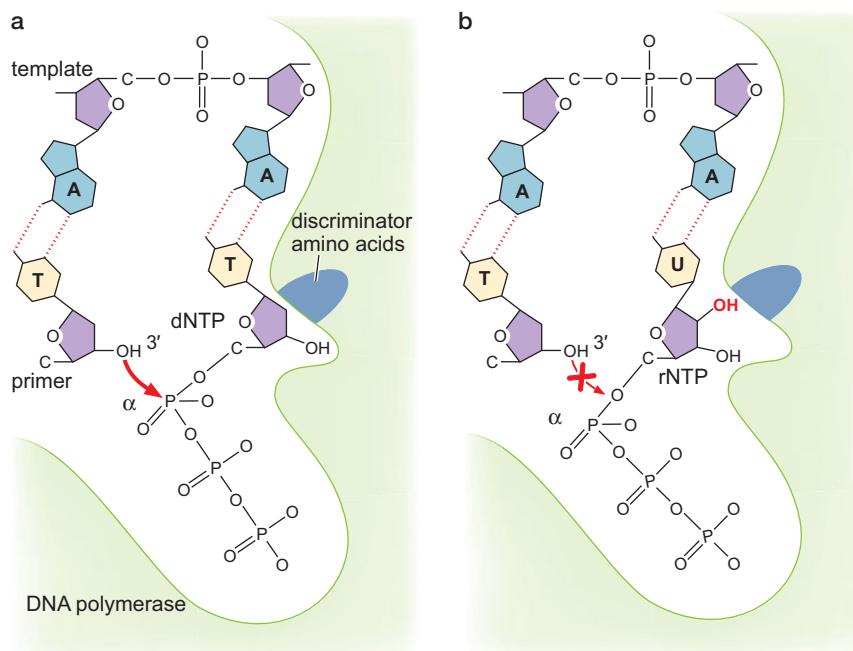
An incorporation assay requires two steps (Box 9-1 Fig. 2). First, the precursor is incorporated into polymers. In the case of DNA polymerase, this is accomplished by incubating the polymerase with a primer/template junction and the labeled dNTP precursor(s) for an appropriate period of time. In most instances, only one of the four dNTPs is labeled because this generally provides easily detectable levels of incorporated nucleotides. Second, the resulting polymers must be separated from unincorporated precursors. In the case of DNA, this can be accomplished in one of two ways. The DNA polymerase reaction can be passed through a positively charged filter in the presence of salt concentrations that allow binding of the highly negatively charged DNA backbone, but not of the less-charged single nucleotides. Alternatively, gel electrophoresis can be used to separate the DNA products by size, because the unincorporated nucleotides will migrate much faster than the DNA product. In either case, the amount of DNA product synthesized can be measured by determining the amount of labeled nucleotide incorporated into the DNA polymer.

We have described an incorporation assay in the context of a DNA polymerase reaction; however, comparable approaches are used to measure the activities of enzymes that direct the synthesis of RNA or proteins. For example, labeled amino acids can be similarly used to analyze their incorporation into proteins.

**BOX 9-1 FIGURE 2** Incorporation assay to measure DNA synthesis. In the example shown, filter binding is used to separate unincorporated from DNA-incorporated labeled nucleotides.



rNTPs from the DNA polymerase active site (Fig. 9-4). In DNA polymerase, the nucleotide-binding pocket cannot accommodate a 2'-OH on the incoming nucleotide. This space is occupied by two amino acids that make van der Waals contacts with the sugar ring. Changing these amino acids to other amino acids with smaller side chains (e.g., by changing a glutamate to an alanine) results in a DNA polymerase with significantly reduced discrimination between dNTPs and rNTPs. Nucleotides that meet some but not all of the requirements for use by DNA polymerase can inhibit DNA synthesis by terminating elongation. Such nucleotides represent an important class of drugs used to treat cancer and viral infections (see Box 9-2, Anticancer and Antiviral Agents Target DNA Replication).



**FIGURE 9-4** Schematic illustration of the steric constraints preventing DNA polymerase from using rNTP precursors. (a) Binding of a correctly base-paired dNTP to the DNA polymerase. Under these conditions, the 3'-OH of the primer and the  $\alpha$ -phosphate of the dNTP are in close proximity. (b) Addition of a 2'-OH results in a steric clash with amino acids (the discriminator amino acids) in the nucleotide-binding pocket. This results in the  $\alpha$ -phosphate of the dNTP being displaced. In this state, the  $\alpha$ -phosphate is incorrectly aligned with the 3'-OH of the primer, dramatically reducing the rate of catalysis.

### DNA Polymerases Resemble a Hand That Grips the Primer:Template Junction

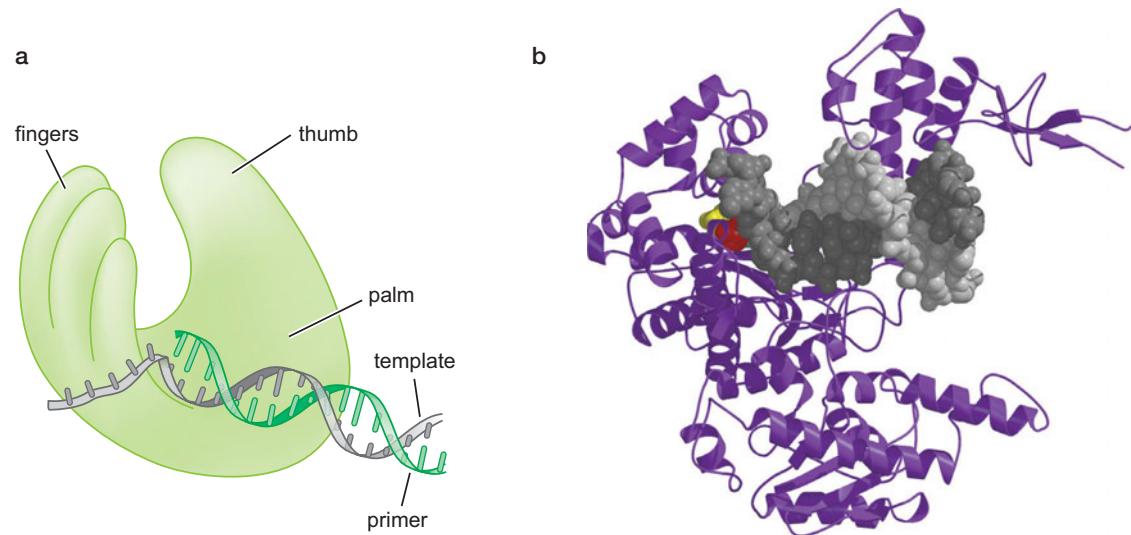
A molecular understanding of how the DNA polymerase catalyzes DNA synthesis has emerged from studies of the atomic structure of various DNA polymerases bound to primer:template junctions (see Structural Tutorial 9-1). These structures reveal that the DNA substrate sits in a large cleft that resembles a partially closed right hand (Fig. 9-5). Based on the hand analogy, the three domains of the polymerase are called the thumb, fingers, and palm.

The palm domain is composed of a  $\beta$  sheet and contains the primary elements of the catalytic site. In particular, this region of DNA polymerase binds two divalent metal ions (typically  $Mg^{2+}$  or  $Zn^{2+}$ ) that alter the chemical environment around the correctly base-paired dNTP and the 3'-OH of the primer (Fig. 9-6). One metal ion reduces the affinity of the 3'-OH for its hydrogen. This generates a  $3' O^-$  that is primed for the nucleophilic attack of the  $\alpha$ -phosphate of the incoming dNTP. The second metal ion coordinates the negative charges of the  $\beta$ -phosphate and  $\gamma$ -phosphate of the dNTP and stabilizes the pyrophosphate produced by joining the primer and the incoming nucleotide.

In addition to its role in catalysis, the palm domain also monitors the base pairing of the most recently added nucleotides. This region of the polymerase makes extensive hydrogen-bond contacts with base pairs in the minor groove of the newly synthesized DNA. These contacts are not base-specific (see Fig. 4-10) but only form if the recently added nucleotides (whichever they may be) are correctly base-paired. Mismatched DNA in this region interferes with these minor-groove contacts and dramatically slows catalysis. The combination of the slowed catalysis and reduced affinity for newly synthesized mismatched DNA allows the release of the primer strand from the polymerase active site, and, in many cases, this strand binds and is acted on by a proofreading nuclease that removes the mismatched DNA (as discussed later).

What are the roles of the fingers and the thumb? The fingers are also important for catalysis. Several residues located within the fingers bind to





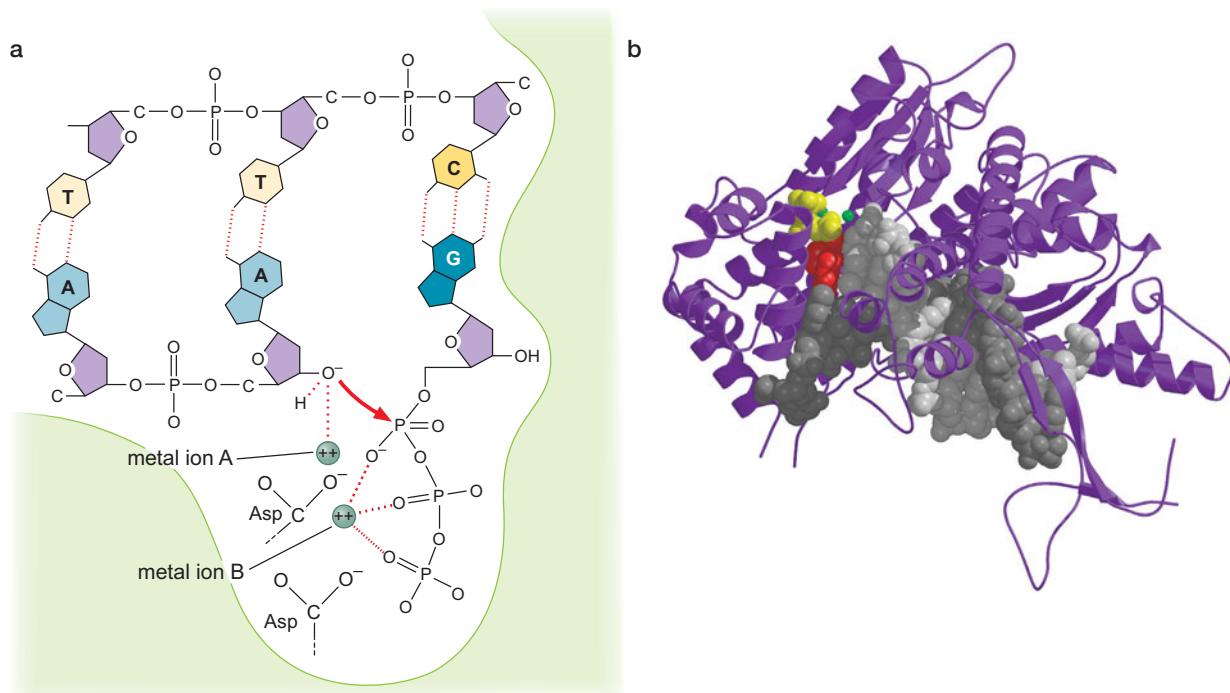
**FIGURE 9-5** 3D structure of DNA polymerase resembles a right hand. (a) Schematic of DNA polymerase bound to a primer:template junction. The fingers, thumb, and palm are noted. The recently synthesized DNA is associated with the palm, and the site of DNA catalysis is located in the crevice between the fingers and the thumb. The single-stranded region of the template strand is bent sharply and does not pass between the thumb and the fingers. (b) A similar view of the T7 DNA polymerase bound to DNA. The DNA is shown in a space-filling manner, and the protein is shown as a ribbon diagram. The fingers and the thumb are composed of  $\alpha$  helices. The palm domain is obscured by the DNA. The incoming dNTP is shown in red (for the base and the deoxyribose) and yellow (for the triphosphate moiety). The template strand of the DNA is shown in dark gray, and the primer strand is shown in light gray. (Adapted from Doublie S. et al. 1998. *Nature* **391**: 251–258. Image prepared with MolScript, BobScript, and Raster3D.)

the incoming dNTP. More importantly, once a correct base pair is formed between the incoming dNTP and the template, the finger domain moves to enclose the dNTP (Fig. 9-7). This closed form of the polymerase “hand” stimulates catalysis by moving the incoming nucleotide into close contact with the catalytic metal ions.

The finger domain also associates with the template region, leading to a nearly  $90^\circ$  turn of the phosphodiester backbone between the first and second bases of the template. This bend serves to expose only the first template base after the primer at the catalytic site and avoids any confusion concerning which template base should pair with the next nucleotide to be added (Fig. 9-8).

In contrast to the fingers and the palm, the thumb domain is not intimately involved in catalysis. Instead, the thumb interacts with the DNA that has been most recently synthesized (see Fig. 9-5). This serves two purposes. First, it maintains the correct position of the primer and the active site. Second, the thumb helps to maintain a strong association between the DNA polymerase and its substrate. This association contributes to the ability of the DNA polymerase to add many dNTPs each time it binds a primer:template junction (as discussed later).

To summarize, an ordered series of events occurs each time the DNA polymerase adds a nucleotide to the growing DNA chain. The incoming nucleotide base-pairs with the next available template base. This interaction causes the fingers of the polymerase to close around the base-paired dNTP. This conformation of the enzyme places the critical catalytic metal ions in a position to catalyze formation of the next phosphodiester bond. Attachment of the base-paired nucleotide to the primer leads to the reopening of the fingers and the movement of the primer:template junction by one base pair. The



**FIGURE 9-6** Two metal ions bound to DNA polymerase catalyze nucleotide addition. (a) Illustration of the active site of a DNA polymerase. The two metal ions (shown in green) are held in place by interactions with two highly conserved aspartate residues. Metal ion A primarily interacts with the 3'-OH, resulting in reduced association between the O and the H. This leaves a nucleophilic  $3'\text{O}^-$ . Metal ion B interacts with the triphosphates of the incoming dNTP to neutralize their negative charge. After catalysis, the pyrophosphate product is stabilized through similar interactions with metal ion B (not shown). (b) 3D structure of the active-site metal ions associated with the T7 DNA polymerase, the 3'-OH end of the primer, and the incoming nucleotide. The metal ions are shown in green and the remaining elements are shown in the same colors as those in Figure 9-5b. The view of the polymerase shown here is roughly equivalent to rotating the image shown in Figure 9-5b  $\sim 180^\circ$  around the axis of the DNA helix. (Adapted from Doublie S. et al. 1998. *Nature* 391: 251–258. Image prepared with MolScript, BobScript, and Raster3D.)

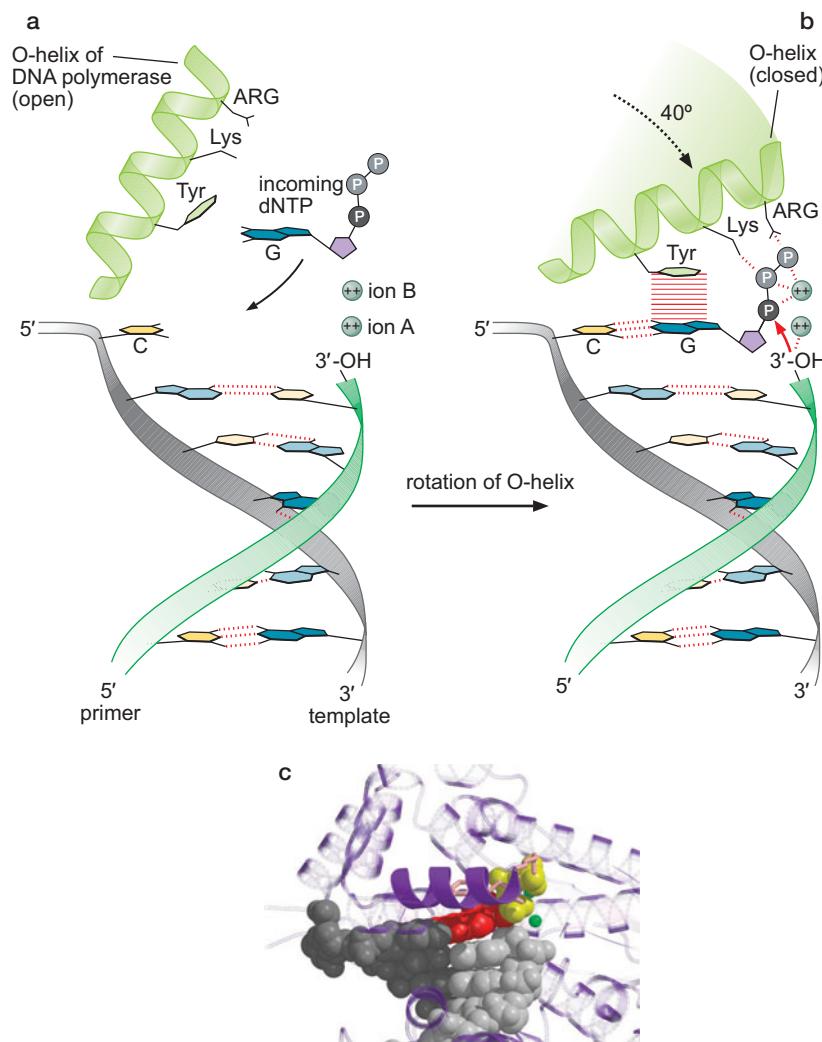
polymerase is then ready for the next cycle of addition. Importantly, each of these events is strongly stimulated by correct base pairing between the incoming dNTP and the template.

### DNA Polymerases Are Processive Enzymes

Catalysis by DNA polymerase is rapid. DNA polymerases are capable of adding as many as 1000 nucleotides/sec to a primer strand. The speed of DNA synthesis is largely due to the processive nature of DNA polymerase. **Processivity** is a characteristic of enzymes that operate on polymeric substrates. In the case of DNA polymerases, the **degree of processivity** is defined as the *average number of nucleotides added each time the enzyme binds a primer:template junction*. Each DNA polymerase has a characteristic processivity that can range from only a few nucleotides to more than 50,000 bases added per binding event (Fig. 9-9).

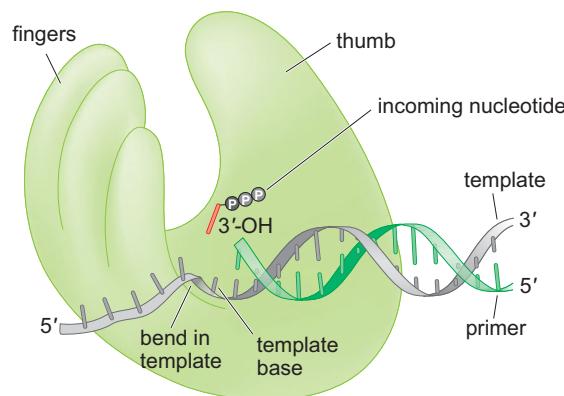
The rate of DNA synthesis is dramatically increased by adding multiple nucleotides per binding event. It is the initial binding of polymerase to the primer:template junction that is rate-limiting for DNA synthesis. In a typical DNA polymerase reaction, it takes  $\sim 1$  sec for the DNA polymerase to locate and bind a primer:template junction. Once bound, addition of a nucleotide

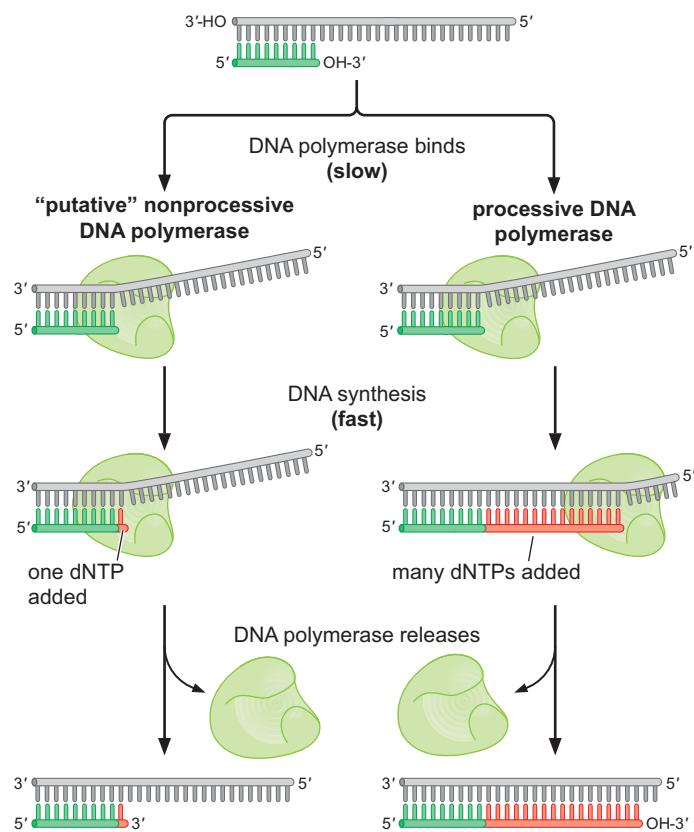
**FIGURE 9-7** DNA polymerase “grips” the template and the incoming nucleotide when a correct base pair is made. (a) An illustration of the changes in DNA polymerase structure after the incoming nucleotide base-pairs correctly to the template DNA. The primary change is a 40° rotation of one of the helices in the finger domain called the O-helix. In the open conformation, this helix is distant from the incoming nucleotide. When the polymerase is in the closed conformation, this helix moves and makes several important interactions with the incoming dNTP. A tyrosine makes stacking interactions with the base of the dNTP, and two charged residues associate with the triphosphate. The combination of these interactions positions the dNTP for catalysis mediated by the two metal ions bound to the DNA polymerase. (Based on Doublie S. et al. 1998. *Nature* 391: 251–258, Fig. 5. © 1998.) (b) The structure of T7 DNA polymerase bound to its substrates in the closed conformation. (Purple) The O-helix; the rest of the protein structure is shown as transparent for clarity. (Pink) The critical tyrosine, lysine, and arginine can be seen behind the O-helix. (Red) The base and the deoxyribose of the incoming dNTP; (light gray) the primer; (dark gray) template strand; (green) the two catalytic metal ions; (yellow) phosphates. (Adapted from Doublie S. et al. 1998. *Nature* 391: 251–258. Image prepared with MolScript, BobScript, and Raster3D.)



is very fast (in the millisecond range). Thus, a completely nonprocessive DNA polymerase would add ~1 bp/sec. In contrast, the fastest DNA polymerases add as many as 1000 nucleotides/sec by remaining associated with the template for thousands of rounds of dNTP addition. Consequently, a highly processive polymerase increases the overall rate of DNA synthesis by as much as 1000-fold compared with a nonprocessive enzyme.

**FIGURE 9-8** Illustration of the path of the template DNA through the DNA polymerase. The recently replicated DNA is associated with the palm region of the DNA polymerase. At the active site, the first base of the single-stranded region of the template is in a position expected for dsDNA. As one follows the template strand toward its 5' end, the phosphodiester backbone abruptly bends 90°. This results in the second and all subsequent single-stranded bases being placed in a position that prevents any possibility of base pairing with a dNTP bound at the active site.



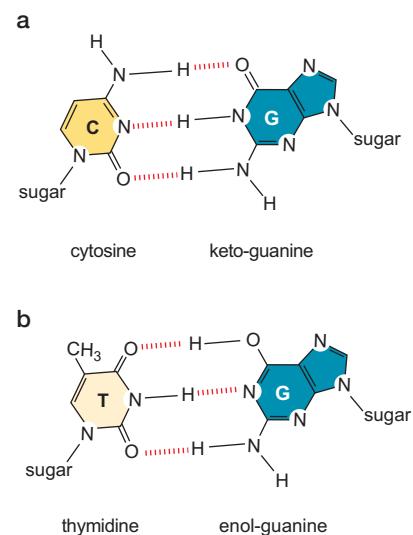


**FIGURE 9-9** DNA polymerases synthesize DNA in a processive manner. This illustration shows the difference between a processive and a nonprocessive DNA polymerase. Both DNA polymerases bind the primer:template junction. Upon binding, the nonprocessive enzyme adds a single dNTP to the 3' end of the primer and then is released from the new primer:template junction. In contrast, a processive DNA polymerase adds many dNTPs each time it binds to the template.

Processivity is facilitated by sliding of DNA polymerases along the DNA template. Once bound to a primer:template junction, DNA polymerase interacts tightly with much of the double-stranded portion of the DNA in a sequence-nonspecific manner. These interactions include electrostatic interactions between the phosphate backbone and the thumb domain and interactions between the minor groove of the DNA and the palm domain (described above). The sequence-independent nature of these interactions permits the easy movement of the DNA even after it binds to polymerase. Each time a nucleotide is added to the primer strand, the DNA partially releases from the polymerase. (The hydrogen bonds with the minor groove are broken, but the electrostatic interactions with the thumb are maintained.) The DNA then rapidly rebinds to the polymerase in a position that is shifted by 1 bp using the same sequence-nonspecific mechanism. Further increases in processivity are achieved through interactions between the DNA polymerase and accessory proteins, which we discuss later.

### Exonucleases Proofread Newly Synthesized DNA

A system based only on base-pair geometry and the complementarity between the bases cannot reach the extraordinarily high levels of accuracy that are observed for DNA synthesis in the cell (approximately one mistake in every  $10^{10}$  bp added). A major limit to DNA polymerase accuracy is the occasional (about one in  $10^5$  times) flickering of the bases into the “wrong” tautomeric form (imino or enol) (see Chapter 4, Fig. 4-5). These alternate forms of the bases permit incorrect base pairs to be correctly positioned for catalysis (Fig. 9-10). When the nucleotide returns to its “correct” state, the incorporated nucleotide is mismatched with the template and must be eliminated.



**FIGURE 9-10** The tautomeric shift of guanine results in mispairing with thymidine. (a) Base pairing between the normal keto form of guanine with cytosine. (b) In the rare instance that guanine assumes the enol tautomer, it now base-pairs with thymidine instead of cytosine. Although we have illustrated the mispairing of the alternate tautomer of guanine, each of the bases can form alternate tautomers that change its base-pairing specificity.

## ► MEDICAL CONNECTIONS

**Box 9-2** Anticancer and Antiviral Agents Target DNA Replication

The central role of DNA replication during cell division makes it a common target for chemotherapeutic drugs that aim to prevent the growth of tumors. These drugs target DNA replication at various stages.

Several common chemotherapeutic reagents target the biosynthesis of the nucleotide precursors for DNA, thus starving DNA polymerase for new building blocks. For example, the drugs 5-fluorouracil (5-FU) and 6-mercaptopurine (6-MP) are analogs of nucleotide precursors that inhibit the synthesis of pyrimidine and purine nucleotides, respectively (Box 9-2 Fig. 1a,b). 5-FU is the major agent used in the treatment of colorectal cancer and is also used in the treatment of stomach, pancreatic, and advanced breast cancer. 6-MP is primarily used to treat patients with acute leukemia (blood cell cancers).

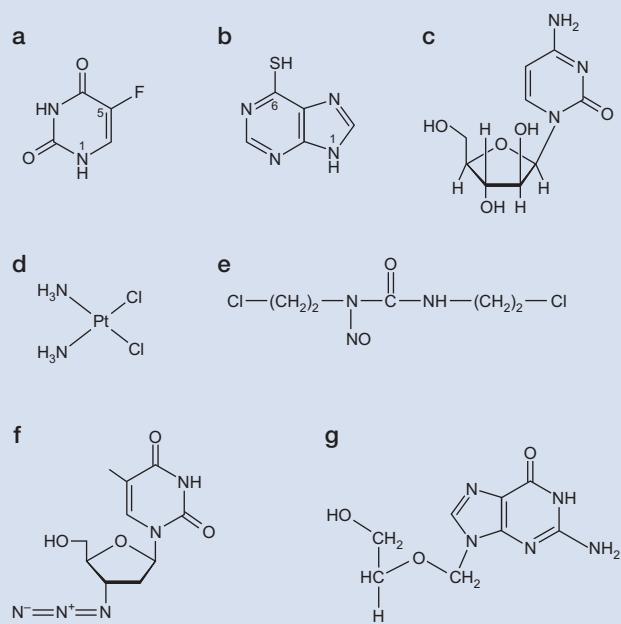
Other anticancer drugs target DNA synthesis more directly. Cytosine arabinoside (AraC) is a deoxycytidine analog that after conversion to a nucleoside triphosphate is incorporated into DNA in place of dCTP (Box 9-2 Fig. 1c). Once incorporated into the DNA, the difference between the deoxyribose sugar of

dCTP and the arabinose sugar of AraCTP leads to incorrect positioning of the 3'-hydroxyl of the primer DNA and termination of elongation. Like 6-MP, AraC is primarily used in the treatment of acute leukemia.

A third class of chemotherapies damages DNA to block DNA replication. Cisplatin and bischloroethyl nitrosourea (BCNU) cause intrastrand and interstrand DNA cross-links when G residues are adjacent to one another (Box 9-2 Fig. 1d,e). These cross-links (particularly the interstrand variety) interfere with DNA elongation. Cisplatin is a major drug used to treat metastatic testicular cancer, and BCNU is used to treat brain tumors and leukemias. Similarly, camptothecin and etoposide are inhibitors of topoisomerases that block the ability of these proteins to re-form a phosphodiester bond after cleaving the DNA backbone (see Chapter 4, Fig. 4-24). Treatment with either of these inhibitors leaves a break in the DNA that terminates DNA replication when DNA polymerase attempts to use it as a template.

As a class, these drugs target cells that are replicating their DNA and therefore are frequently dividing. Although the rapidly dividing nature of cancer cells makes them particularly susceptible to such drugs, other cells in the body are affected also. Not surprisingly, these DNA replication inhibitors are also toxic toward rapidly growing host cells such as red and white blood cells, hair cells, and gastrointestinal mucosal cells. Inhibiting the growth of these cells leads to the now familiar side effects of many chemotherapies, including immunosuppression (due to loss of white blood cells), anemia (due to loss of red blood cells), diarrhea (due to gastrointestinal defects), and hair loss.

Replication inhibitors have also been used as antiviral agents. The first drug found to be effective against HIV infection was azidothymidine (AZT), a thymidine analog that inhibits the specialized DNA polymerase (called a reverse transcriptase) (see Chapter 12) that copies the RNA genome of HIV into DNA after infection. More recently, a guanine nucleoside analog called acyclovir has replaced AZT as the preferred HIV DNA polymerase inhibitor. In this analog, the ribose of a normal nucleoside is replaced with an open-chain structure that resembles the part of ribose closest to the base (Box 9-2 Fig. 1f,g). Nevertheless, this analog can be modified to a triphosphate form that can be incorporated by the viral DNA polymerase into DNA. Once incorporated, these analogs act as chain terminators because of their lack of a ribose group and therefore the 3'-OH required for further nucleotide addition. Importantly, these drugs are poorly recognized by cellular DNA polymerases and, thus, have fewer side effects than chemotherapeutic nucleotide analogs.



**BOX 9-2 FIGURE 1** Structures of common chemotherapeutic reagents that target DNA replication. (a) 5-Fluorouracil, (b) 6-mercaptopurine, (c) cytosine arabinoside, (d) cisplatin, (e) bischloroethylnitrosourea, (f) azidothymidine, and (g) acyclovir.

Removal of these incorrectly base-paired nucleotides is mediated by a type of nuclease that was originally identified in the same polypeptide as the DNA polymerase. Referred to as **proofreading exonuclease**, these enzymes degrade DNA starting from a 3' DNA end (i.e., from the growing end of the new DNA strand). (Nucleases that can only degrade from a DNA

end are called **exonucleases**; nucleases that can cut within a DNA strand are called **endonucleases**.)

Initially, the presence of a 3' exonuclease as part of the same polypeptide as a DNA polymerase made little sense. Why would the DNA polymerase need to degrade the DNA it had just synthesized? The role for these exonucleases became clear when it was determined that they have a strong preference to degrade DNA containing mismatched base pairs. Thus, in the rare event that an incorrect nucleotide is added to the primer strand, the exonuclease removes this nucleotide from the 3' end of the primer strand. This "proofreading" of the newly added DNA gives the DNA polymerase a second chance to add the correct nucleotide.

The removal of mismatched nucleotides is facilitated by the reduced ability of DNA polymerase to add a nucleotide adjacent to an incorrectly base-paired primer. Mispaired DNA alters the geometry between the 3'-OH and the incoming nucleotide because of poor interactions with the palm region. This altered geometry reduces the rate of nucleotide addition in much the same way that addition of an incorrectly paired dNTP reduces catalysis. Thus, when a mismatched nucleotide is added, it both decreases the rate of new nucleotide addition and increases the rate of proofreading exonuclease activity.

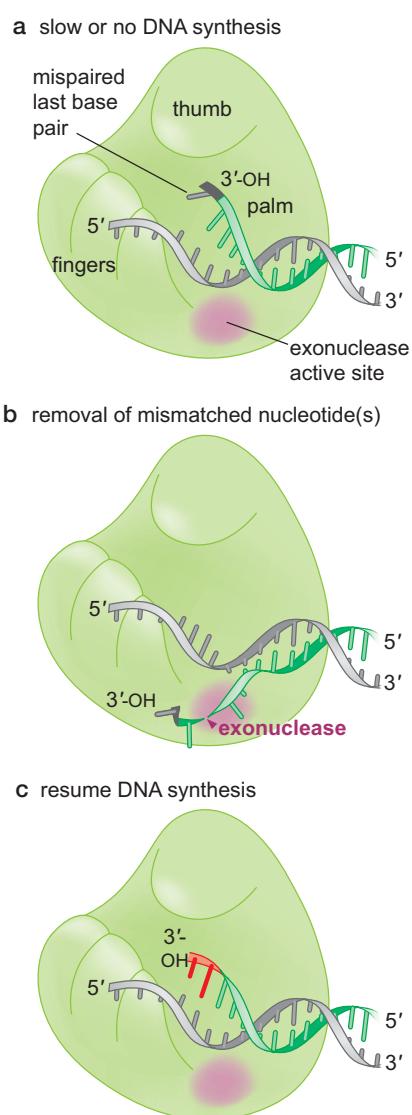
As for processive DNA synthesis, proofreading occurs without releasing the DNA from the polymerase (Fig. 9-11). When a mismatched base pair is present in the polymerase active site, the primer:template junction is destabilized, creating several base pairs of unpaired DNA. The DNA polymerase active site binds such a mismatched template poorly, but the exonuclease active site has a 10-fold higher affinity for single-stranded 3' ends. Thus, the newly unpaired 3' end moves from the polymerase active site to the exonuclease active site. The incorrect nucleotide is removed by the exonuclease (an additional nucleotide may also be removed). The removal of the mismatched base allows the primer:template junction to re-form and rebind the polymerase active site, enabling DNA synthesis to continue.

In essence, proofreading exonucleases work like a "delete key," removing only the most recent errors. The addition of a proofreading exonuclease greatly increases the accuracy of DNA synthesis. On average, DNA polymerase inserts one incorrect nucleotide for every  $10^5$  nucleotides added. Proofreading exonucleases decrease the appearance of incorrect base pairs to 1 in every  $10^7$  nucleotides added. This error rate is still significantly short of the actual rate of mutation observed in a typical cell (approximately one mistake in every  $10^{10}$  nucleotides added). This additional level of accuracy is provided by the postreplication mismatch repair process described in Chapter 10.

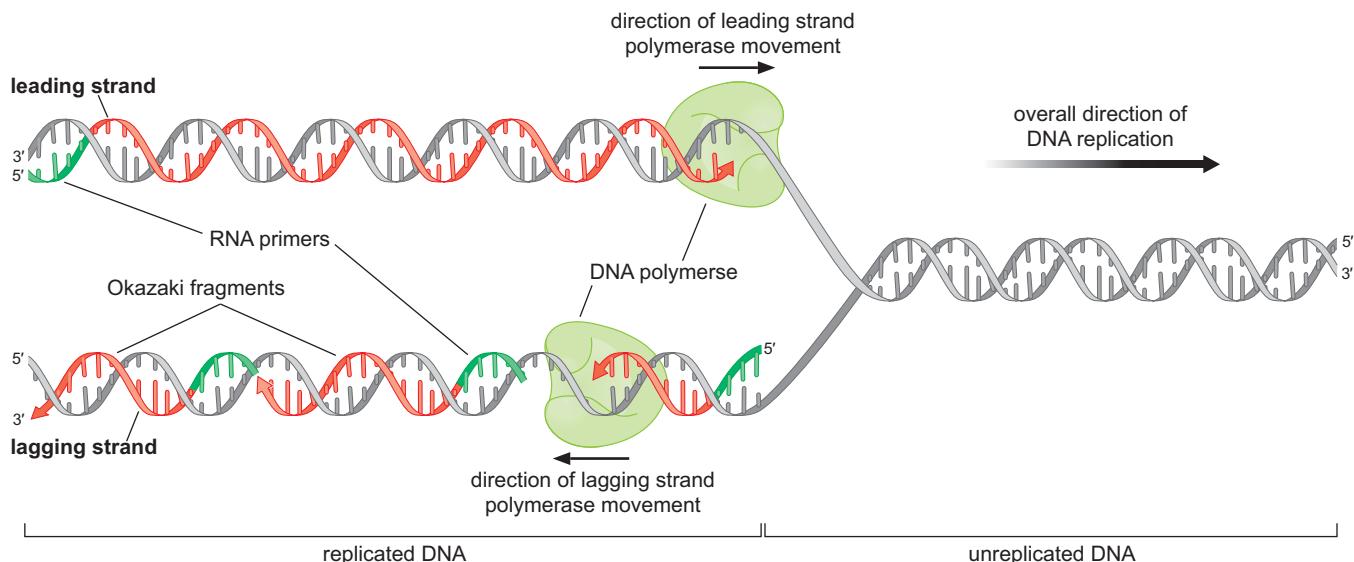
## THE REPLICATION FORK

### Both Strands of DNA Are Synthesized Together at the Replication Fork

Thus far, we have discussed DNA synthesis in a relatively artificial context, that is, at a primer:template junction that is producing only one new strand of DNA. In the cell, both strands of the DNA duplex are replicated at the same time (see Interactive Animation 9-2). This requires separation of the two strands of the double helix to create two template DNAs. The junction between the newly separated template strands and the unreplicated duplex DNA is known as the **replication fork** (Fig. 9-12). The replication fork moves



**FIGURE 9-11** Proofreading exonucleases removes bases from the 3' end of mismatched DNA. (a) When an incorrect nucleotide is incorporated into the DNA, the rate of DNA synthesis is reduced because of the incorrect positioning of the 3'-OH. (b) In the presence of a mismatched 3' end, the last 3–4 nucleotides of the primer become single-stranded, resulting in an increased affinity for the exonuclease active site. Once bound at this active site, the mismatched nucleotide (and frequently an additional nucleotide) is removed from the primer. (c) Once the mismatched nucleotide is removed, a properly base-paired primer:template junction is re-formed, and polymerization resumes (newly synthesized DNA is shown in red). (Adapted, with permission, from Baker T.A. and Bell S.P. 1998. *Cell* 92: 295–305, Fig. 1b. © Elsevier.)



**FIGURE 9-12** Replication fork. (Red) Newly synthesized DNA; (green) RNA primers. The Okazaki fragments shown are artificially short for illustrative purposes. In the cell, Okazaki fragments can vary between 100 and 2000 bases depending on the organism.

continuously toward the duplex region of unreplicated DNA, leaving in its wake two ssDNA templates that each direct the synthesis of a complementary DNA strand.

The antiparallel nature of DNA creates a complication for the simultaneous replication of the two exposed templates at the replication fork. Because DNA is synthesized only by elongating a 3' end, only one of the two exposed templates can be replicated continuously as the replication fork moves. On this template strand, the polymerase simply “chases” the moving replication fork. The newly synthesized DNA strand directed by this template is known as the **leading strand**.

Synthesis of the new DNA strand directed by the other ssDNA template is more complicated. This template directs the DNA polymerase to move in the opposite direction of the replication fork. The new DNA strand directed by this template is known as the **lagging strand**. As shown in Figure 9-12, this strand of DNA must be synthesized in a discontinuous fashion.

Although the leading-strand DNA polymerase can replicate its template as soon as it is exposed, synthesis of the lagging strand must wait for movement of the replication fork to expose a substantial length of template before it can be replicated. Each time a substantial length of new lagging-strand template is exposed, DNA synthesis is initiated and continues until it reaches the 5' end of the previous newly synthesized stretch of lagging-strand DNA.

The resulting short fragments of new DNA formed on the lagging strand are called **Okazaki fragments** and vary in length from 1000 to 2000 nucleotides in bacteria and from 100 to 400 nucleotides in eukaryotes. Shortly after being synthesized, Okazaki fragments are covalently joined together to generate a continuous, intact strand of new DNA (see later discussion). Okazaki fragments are therefore transient intermediates in DNA replication.

### The Initiation of a New Strand of DNA Requires an RNA Primer

As described above, all DNA polymerases require a primer with a free 3'-OH. They cannot initiate a new DNA strand de novo. How, then, are new strands

of DNA synthesis started? To accomplish this, the cell takes advantage of the ability of RNA polymerases to do what DNA polymerases cannot: start new RNA chains de novo. **Primase** is a specialized RNA polymerase dedicated to making short RNA primers (5–10 nucleotides long) on an ssDNA template. These primers are subsequently extended by DNA polymerase. Although DNA polymerases incorporate only deoxyribonucleotides into DNA, they can initiate synthesis using either an RNA primer or a DNA primer annealed to the DNA template.

Although both the leading and lagging strands require primase to initiate DNA synthesis, the frequency of primase function on the two strands is dramatically different (see Fig. 9-12). Each leading strand requires only a single RNA primer. In contrast, the discontinuous synthesis of the lagging strand means that new primers are needed for each Okazaki fragment. Because a single replication fork can add hundreds of thousands of nucleotides to a primer, synthesis of the lagging strand can require hundreds of Okazaki fragments and their associated RNA primers.

Unlike the RNA polymerases involved in messenger RNA (mRNA), ribosomal RNA (rRNA), and transfer RNA (tRNA) synthesis (see Chapter 15), primase does not require an extended DNA sequence to initiate RNA synthesis. Instead, primases prefer to initiate RNA synthesis using an ssDNA template containing a particular trimer (GTA in the case of *Escherichia coli* primase). Consistent with this preference, analysis of the *E. coli* genome sequence shows that the GTA target sequence for *E. coli* primase is overrepresented in the portions of the genome that will be the template for lagging-strand DNA synthesis.

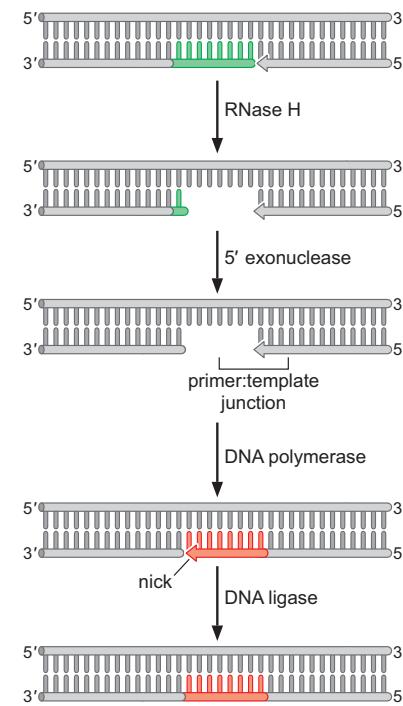
Primase activity is dramatically increased when it associates with another protein that acts at the replication fork called **DNA helicase**. This protein unwinds the DNA at the replication fork, creating an ssDNA template that can be acted on by primase. DNA helicase function is considered in more detail later. The requirement for an ssDNA template and DNA helicase association ensures that primase is only active at the replication fork.

### RNA Primers Must Be Removed to Complete DNA Replication

To complete DNA replication, the RNA primers used for initiation must be removed and replaced with DNA (Fig. 9-13). Removal of the RNA primers can be thought of as a DNA-repair event, and this process shares many of the properties of excision DNA repair, a process covered in detail in Chapter 10.

To replace the RNA primers with DNA, an enzyme called **RNase H** recognizes and removes most of each RNA primer. This enzyme specifically degrades RNA that is base-paired with DNA (the *H* in its name stands for “hybrid” in RNA:DNA hybrid). RNase H removes all of the RNA primer except the ribonucleotide directly linked to the DNA end. This is because RNase H can only cleave bonds between two ribonucleotides. The final ribonucleotide is removed by a 5' exonuclease that degrades RNA or DNA from their 5' ends.

Removal of the RNA primer leaves a gap in the dsDNA that is an ideal substrate for DNA polymerase—a primer:template junction (see Fig. 9-13). DNA polymerase fills this gap until every nucleotide is base-paired, leaving a DNA molecule that is complete except for a break in the phosphodiester backbone between the 3'-OH and 5'-phosphate of the repaired strand. This “nick” in the DNA can be repaired by an enzyme called **DNA ligase**. DNA ligases use high-energy co-factors (such as ATP) to create a phosphodiester bond between an adjacent 5'-phosphate and 3'-OH. Only after all RNA



**FIGURE 9-13** Removal of RNA primers from newly synthesized DNA. The sequential function of RNase H, 5' exonuclease, DNA polymerase, and DNA ligase during the removal of RNA primers is illustrated. (Gray) DNA present before RNA primer removal; (green) RNA primer; (red) the newly synthesized DNA that replaces the RNA primer.

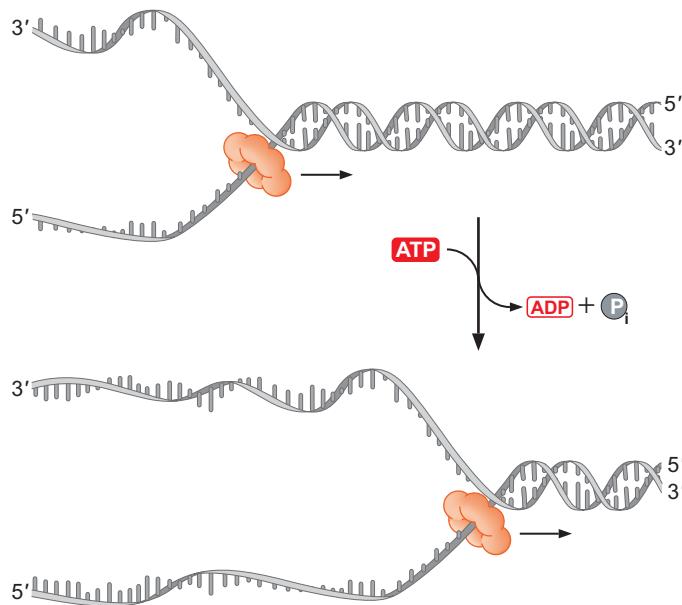
primers are replaced by DNA and the associated nicks are sealed is DNA synthesis complete.

### DNA Helicases Unwind the Double Helix in Advance of the Replication Fork

DNA polymerases are generally poor at separating the two base-paired strands of duplex DNA. Therefore, at the replication fork, a third class of enzymes, called **DNA helicases**, catalyze the separation of the two strands of duplex DNA. These enzymes bind to and move directionally along ssDNA using the energy of nucleoside triphosphate (usually ATP) binding and hydrolysis to displace any DNA strand that is annealed to the bound ssDNA. Typically, DNA helicases that act at replication forks are hexameric proteins that assume the shape of a ring (Fig. 9-14). These ring-shaped protein complexes encircle one of the two single strands at the replication fork adjacent to the single-stranded:double-stranded junction.

Like DNA polymerases, DNA helicases act processively. Each time they associate with substrate, they unwind multiple base pairs of DNA. The ring-shaped hexameric DNA helicases found at replication forks exhibit high processivity because they encircle the DNA. Release of the helicase from its DNA substrate therefore requires the opening of the hexameric protein ring, which is a rare event. Alternatively, the helicase can dissociate when it reaches the end of the DNA strand that it has encircled.

Of course, this arrangement of enzyme and DNA poses problems for the binding of the DNA helicase to the DNA substrate in the first place. This problem is most obvious for circular chromosomes, where there is no DNA end for the DNA helicase to thread onto. However, because helicases are almost always loaded onto the DNA at internal sites of linear chromosomes, the same problem exists during the replication of these DNAs. Thus, there are specialized mechanisms that open the DNA helicase ring and place it around the DNA before re-forming the ring (see section on Initiation of DNA Replication). This topological linkage between proteins involved in DNA replication and their DNA substrates is a common mechanism to increase processivity.



**FIGURE 9-14** DNA helicases separate the two strands of the double helix. When ATP is added to a DNA helicase bound to ssDNA, the helicase moves with a defined polarity on the ssDNA. In the instance illustrated, the DNA helicase has a  $5' \rightarrow 3'$  polarity. This polarity means that the DNA helicase would be bound to the lagging-strand template at the replication fork.

Each DNA helicase moves along ssDNA in a defined direction. This property is referred to as the **polarity** of the DNA helicase. DNA helicases can have a polarity of either  $5' \rightarrow 3'$  or  $3' \rightarrow 5'$ . This direction is always defined according to the strand of DNA bound (or encircled for a ring-shaped helicase), rather than the strand that is displaced. In the case of a DNA helicase that functions on the lagging-strand template of the replication fork, the polarity is  $5' \rightarrow 3'$  to allow the DNA helicase to proceed toward the duplex region of the replication fork (see Fig. 9-14). As is true for all enzymes that move along DNA in a directional manner, movement of the helicase along ssDNA requires the input of chemical energy. For helicases, this energy is provided by ATP hydrolysis.

### DNA Helicase Pulls Single-Stranded DNA through a Central Protein Pore

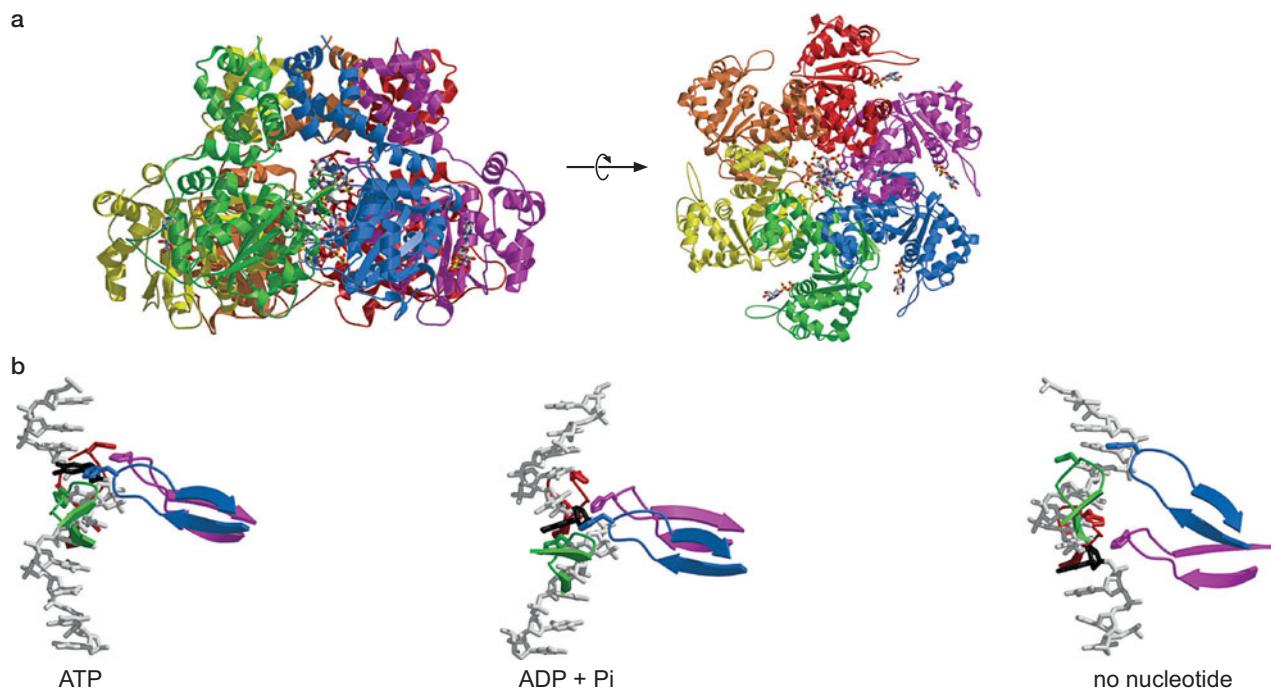
How does a hexameric DNA helicase use the energy of ATP hydrolysis to move along the DNA? The determination of the atomic structure of a viral hexameric helicase bound to a single-stranded DNA substrate provides insights into this question. In this structure, the ssDNA is encircled by the six subunits of the helicase (Fig. 9-15). Each subunit has a “hairpin” protein loop that binds a phosphate of the DNA backbone and its two adjacent ribose components. Interestingly, these DNA-binding loops are found in a right-handed spiral staircase, each binding the next phosphate along the ssDNA. As shown in Figure 9-15, the top of the staircase is associated with the 5' end and the bottom with the 3' end of the ssDNA.

The atomic structure is only a single snapshot; however, each of the six different subunits is at a different stage in the DNA translocation process. Together the interactions of the different subunits with DNA and ATP/ADP reveal how the coordinated movement of these protein hairpins can pull the ssDNA through the central pore of the helicase. A subunit first binds the ssDNA at the top of the structure (Fig. 9-15b), and the DNA-binding loop moves through successive conformations toward the bottom, bringing the bound DNA along with it. Importantly, each of these conformations is associated with a different nucleotide-bound state; the top conformation is in an ATP-bound state, the middle is in an ADP-bound state, and the bottom lacks a bound nucleotide. Thus, as a single subunit binds, hydrolyzes, and releases ATP, it will cycle through the top, middle, and bottom conformations. Overall, one can think of the helicase as having six hands pulling on a rope in a hand-over-hand manner.

In addition to translocating along ssDNA, a helicase must also displace the complementary strand to cause DNA unwinding. In the case of this hexameric helicase, the structure of the central channel shows that strand displacement must occur for DNA to pass through the channel. At its narrowest point, the central channel has a diameter of 13 Å, large enough to fit ssDNA, but much too small to fit the 20 Å diameter of dsDNA.

### Single-Stranded DNA-Binding Proteins Stabilize ssDNA before Replication

After the DNA helicase has passed, the newly generated ssDNA must remain free of base pairing until it can be used as a template for DNA synthesis. To stabilize the separated strands, **ssDNA-binding proteins** (SSBs) rapidly bind to the separated strands. Binding of one SSB promotes the binding of another SSB to the immediately adjacent ssDNA (Fig. 9-16).

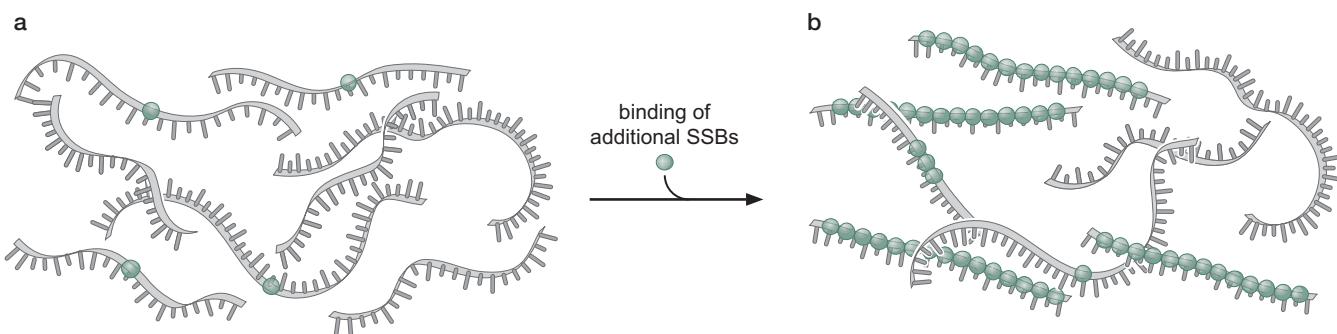


**FIGURE 9-15 Structure and proposed mechanisms of a DNA helicase.** (a) Overall structure of the bovine papillomavirus E1 hexameric helicase. Each subunit is shown in a different color, and the complex is shown from the side (left) and looking down the central channel of the hexamer (right). The protein subunits are shown in ribbon form and the ssDNA as a stick diagram. (b) Illustration of the proposed movement of DNA-binding hairpins. In these views, we are only showing the ssDNA in the central channel of the helicase and two of the six critical protein hairpins that bind this DNA during translocation. The three views show the purple hairpin interacting with the phosphate associated with the black nucleotide moving from the top (the subunit associated with the hairpin is ATP-bound) state to the middle (ADP-bound) state to the bottom (nucleotide-unbound) state. Note how the black base in the ssDNA moves with the blue hairpins in the ATP and ADP + Pi but is released in the no-nucleotide image. Rebinding of ATP drives the DNA-binding loop back to the top position to allow binding to a new phosphate (as seen for the blue hairpin—which is one step ahead of the purple hairpin—in the “no nucleotide” panel; the headings of the panels refer to the purple hairpin subunit). Note that the other four DNA-binding loops are also moving through the same intermediates. (From Enemark E.J. and Joshua-Tor L. 2006. *Nature* 442: 270–275. PDB Code: 2GXA. Image prepared with MolScript, BobScript, and Raster3D.)

This is called **cooperative binding** and occurs because SSB molecules bound to immediately adjacent regions of ssDNA also bind to each other. This SSB–SSB interaction strongly stabilizes SSB binding to ssDNA and makes sites already occupied by one or more SSB molecules preferred SSB-binding sites.

Cooperative binding ensures that ssDNA is rapidly coated by SSB as it emerges from the DNA helicase. (Cooperative binding is a property of many DNA-binding proteins. See Chapter 18, Box 18-4, Concentration, Affinity, and Cooperative Binding.) Once coated with SSBs, ssDNA is held in an elongated state that facilitates its use as a template for DNA or RNA primer synthesis.

SSBs interact with ssDNA in a sequence-independent manner. SSBs primarily contact ssDNA through electrostatic interactions with the phosphate backbone and stacking interactions with the DNA bases. In contrast to sequence-specific DNA-binding proteins, SSBs make few, if any, hydrogen bonds to the ssDNA bases.



**FIGURE 9-16** Binding of single-stranded binding protein (SSB) to DNA. (a) A limiting amount of SSBS is bound to four of the nine ssDNA molecules shown. (b) As more SSBS bind to DNA, they preferentially bind adjacent to previously bound SSB molecules. Only after SSBS have completely coated the initially bound ssDNA molecules does binding occur on other molecules. Note that when ssDNA is coated with SSBS, it assumes a more extended conformation that inhibits the formation of intramolecular base pairs.

### Topoisomerases Remove Supercoils Produced by DNA Unwinding at the Replication Fork

As the strands of DNA are separated at the replication fork, the dsDNA in front of the fork becomes increasingly positively supercoiled (Fig. 9-17). This accumulation of supercoils is the result of DNA helicase eliminating the base pairs between the two strands. If the DNA strands remain unbroken, there can be no reduction in linking number (the number of times the two DNA strands are intertwined) to accommodate this unwinding of the DNA duplex (see Chapter 4). Thus, as the DNA helicase proceeds, the DNA must accommodate the same linking number within a smaller and smaller number of base pairs. Indeed, for the DNA in front of the replication fork to remain relaxed, one DNA link must be removed for every  $\sim 10$  bp of DNA unwound. If there were no mechanism to relieve the accumulation of these supercoils, the replication machinery would grind to a halt in the face of mounting strain placed on the DNA in front of the replication fork.

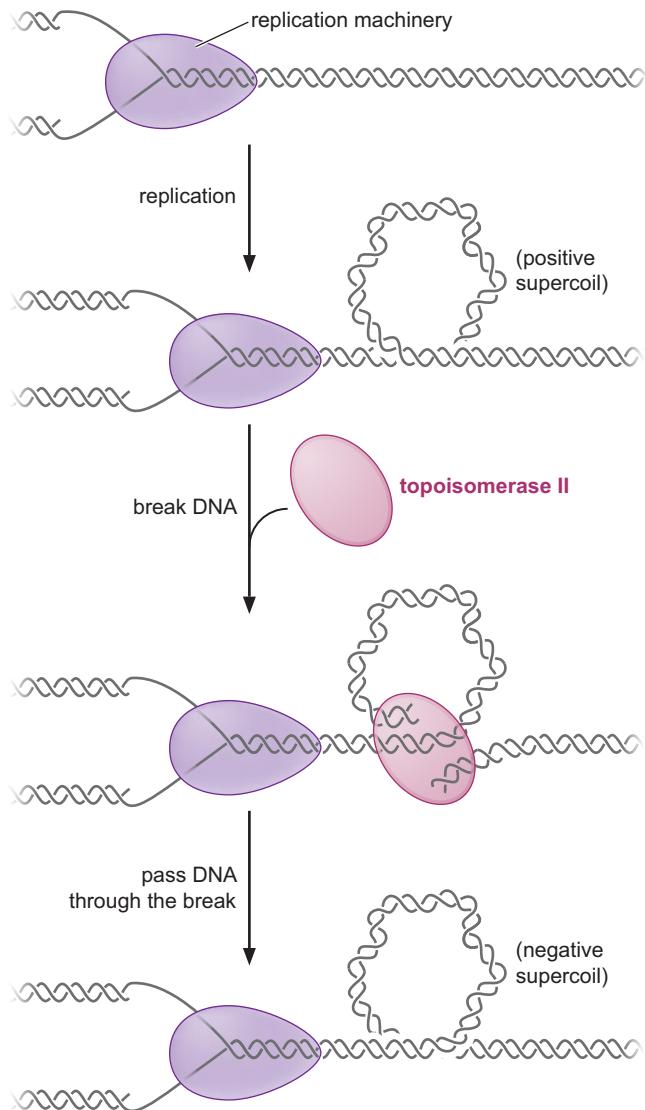
This problem is most clear for the circular chromosomes of bacteria, but it also applies to eukaryotic chromosomes. Because eukaryotic chromosomes are not closed circles, they could in principle rotate along their length to dissipate the introduced supercoils. This is not the case, however: It is simply not possible to rotate a DNA molecule that is millions of base pairs long each time one turn of the helix is unwound.

The supercoils introduced by the action of the DNA helicase are removed by **topoisomerases** that act on the unreplicated dsDNA in front of the replication fork (Fig. 9-17). These enzymes do this by breaking either one or both strands of the DNA without letting go of the DNA and passing the same number of DNA strands through the break (as we discussed in Chapter 4). This action relieves the accumulation of supercoils. In this way, topoisomerases act as a “swivelase” that prevents the accumulation of positive supercoils ahead of the replication fork.

### Replication Fork Enzymes Extend the Range of DNA Polymerase Substrates

On its own, DNA polymerase can only efficiently extend 3'-OH primers annealed to ssDNA templates. The addition of primase, DNA helicase, and topoisomerase dramatically extends the possible substrates for DNA

**FIGURE 9-17 Action of topoisomerase at the replication fork.** As positive supercoils accumulate in front of the replication fork, topoisomerases rapidly remove them. In this diagram, the action of Topo II removes the positive supercoil induced by a replication fork. By passing one part of the unreplicated dsDNA through a double-stranded break in a nearby unreplicated region, the positive supercoils can be removed. It is worth noting that this change would reduce the linking number by two and thus would only have to occur once every 20 bp replicated. Although the action of a type II topoisomerase is illustrated here, type I topoisomerases can also remove the positive supercoils generated by a replication fork. Note that the positive superhelicity in front of the replication fork is shown as right-handed toroidal writhe (one complete turn equals a positive writhe of +1). Passage of one dsDNA molecule through the other at the site of the writhe changes this to one complete left-handed toroidal writhe (equal to a writhe of -1). This illustrates how the linking number is changed by 2 units by a type II topoisomerase (for more information regarding DNA topology and writhe, see Chapter 4, DNA Topology, and Chapter 8, Box 8-2).



polymerase. Primase provides the ability to initiate new DNA strands on any piece of ssDNA. Of course, the use of primase also imposes a requirement for the removal of the RNA primers to complete replication. Similarly, strand separation by DNA helicase and dissipation of positive supercoils by topoisomerase allow DNA polymerase to replicate dsDNA. Although the names of the proteins change from organism to organism (Table 9-1), the same set of enzymatic activities is used by organisms as diverse as bacteria, yeast, and humans to accomplish chromosomal DNA replication.

It is noteworthy that both DNA helicase and topoisomerase perform their functions without permanently altering the chemical structure of DNA or synthesizing any new molecule. DNA helicase breaks only the hydrogen bonds that hold the two strands of DNA together without breaking any covalent bonds. Although topoisomerases break one or two of the targeted DNA's covalent bonds, each bond broken is precisely re-formed before the topoisomerase releases the DNA (see Chapter 4, Fig. 4-25). Instead of altering the chemical structure of DNA, the action of these enzymes results in a DNA molecule with an altered conformation. Importantly, these conformational alterations are essential for the duplication of the large dsDNA molecules that are the foundation of both bacterial and eukaryotic chromosomes.

**TABLE 9-1** Enzymes That Function at the Replication Fork

	<i>Escherichia coli</i>	<i>Saccharomyces cerevisiae</i>	Human
Primase	DnaG	Primase (PRI I/PRI 2)	Primase
DNA helicase	DnaB	Mcm2-7 complex	Mcm2-7 complex
SSB	SSB	RPA	RPA
Topoisomerases	Topo I, Gyrase	Topo I, II	Topo I, II

The proteins that act at the replication fork interact tightly but in a sequence-independent manner with the DNA. These interactions exploit the features of DNA that are the same regardless of the particular base pair: the negative charge and structure of the phosphate backbone (e.g., the thumb domain of DNA polymerase), hydrogen-bonding residues in the minor groove (e.g., the palm domain of the DNA polymerase), and the hydrophobic stacking interactions between the bases (e.g., SSBs). In addition, the structures of some of these proteins are specialized to encourage processive action by either fully (e.g., DNA helicase) or partially (e.g., DNA polymerase) encircling the DNA.

## THE SPECIALIZATION OF DNA POLYMERASES

### DNA Polymerases Are Specialized for Different Roles in the Cell

The central role of DNA polymerases in the efficient and accurate replication of the genome has resulted in the evolution of multiple specialized DNA polymerases. For example, *E. coli* has at least five DNA polymerases that are distinguished by their enzymatic properties, subunit composition, and abundance (Table 9-2). **DNA polymerase III** (DNA Pol III) is the primary enzyme involved in the replication of the chromosome. Because the entire 4.6-Mb *E. coli* genome is replicated by two replication forks, DNA Pol III must be highly processive. Consistent with these requirements, DNA Pol III is generally found to be part of a larger complex that confers very high processivity—a complex known as the **DNA Pol III holoenzyme**.

In contrast, **DNA polymerase I** (DNA Pol I) is specialized for the removal of the RNA primers that are used to initiate DNA synthesis. For this reason, this DNA polymerase has a 5' exonuclease that allows DNA Pol I to remove RNA or DNA immediately *upstream* of the site of DNA synthesis. Unlike DNA Pol III holoenzyme, DNA Pol I is not highly processive, adding only 20–100 nucleotides per binding event. These properties are ideal for RNA primer removal and DNA synthesis across the resulting ssDNA gap. The 5' exonuclease of DNA Pol I can remove the RNA–DNA linkage that is resistant to RNase H (see Fig. 9-13). The low processivity of DNA Pol I readily synthesizes across the short region previously occupied by an RNA primer (<10 nucleotides) but is released before degrading and resynthesizing large amounts of DNA that was primed by the RNA. Finally, when DNA Pol I completes its function, only a nick is present in the DNA.

Because both DNA Pol I and DNA Pol III are involved in DNA replication, both of these enzymes must be highly accurate. Thus, both proteins include an associated proofreading exonuclease. The remaining three DNA polymerases in *E. coli* are specialized for DNA repair and lack proofreading activities. These enzymes are discussed in Chapter 10.

Eukaryotic cells also have multiple DNA polymerases, with a typical cell having more than 15. Of these, three are essential to duplicate the genome:

**TABLE 9-2** Activities and Functions of DNA Polymerases

	Number of subunits	Function
<b>Prokaryotic (<i>E. coli</i>)</b>		
Pol I	1	RNA primer removal, DNA repair
Pol II (Din A)	1	DNA repair
Pol III core	3	Chromosome replication
Pol III holoenzyme	9	Chromosome replication
Pol IV (Din B)	1	DNA repair, translesion synthesis (TLS)
Pol V (UmuC, UmuD <sub>2</sub> 'C)	3	TLS
<b>Eukaryotic</b>		
Pol $\alpha$	4	Primer synthesis during DNA replication
Pol $\beta$	1	Base excision repair
Pol $\gamma$	3	Mitochondrial DNA replication and repair
Pol $\delta$	2–3	Lagging-strand DNA synthesis; nucleotide and base excision repair
Pol $\epsilon$	4	Leading-strand DNA synthesis; nucleotide and base excision repair
Pol $\theta$	1	DNA repair of cross-links
Pol $\zeta$	1	TLS
Pol $\lambda$	1	Meiosis-associated DNA repair
Pol $\mu$	1	Somatic hypermutation
Pol $\kappa$	1	TLS
Pol $\eta$	1	Relatively accurate TLS past <i>cis</i> – <i>syn</i> cyclobutane dimers
Pol $\iota$	1	TLS, somatic hypermutation
Rev1	1	TLS

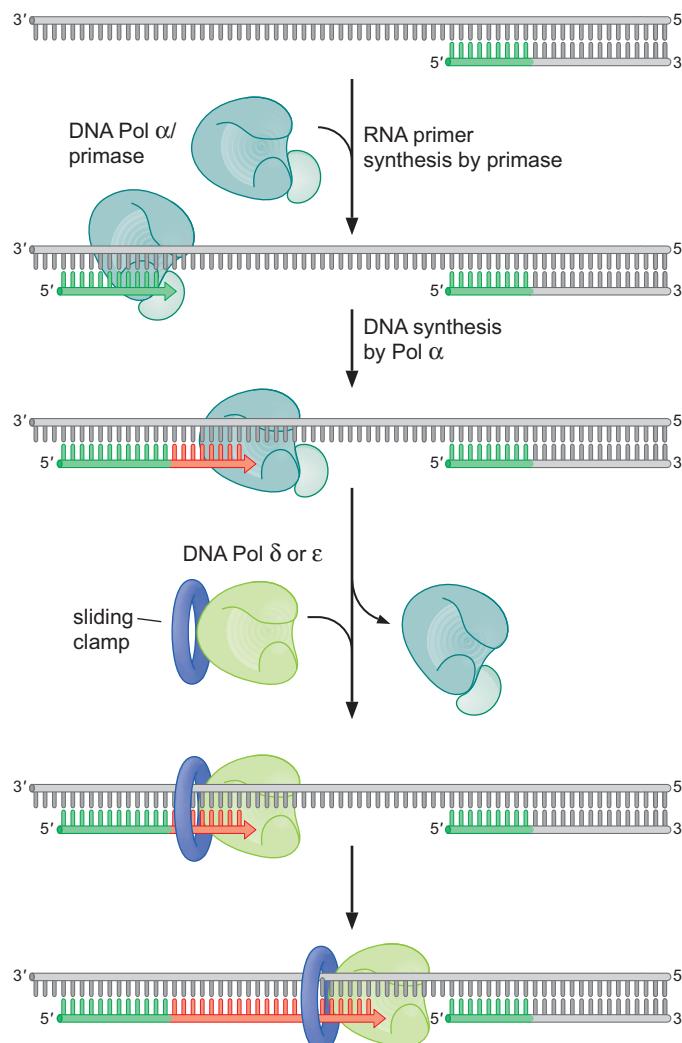
Data from Sutton M.D. and Walker G.C. 2001. *Proc. Natl. Acad. Sci.* **98**: 8342–8349, and references therein.

DNA Pol  $\delta$ , DNA Pol  $\epsilon$ , and DNA Pol  $\alpha$ /primase. Each of these eukaryotic DNA polymerases is composed of multiple subunits (see Table 9-2). DNA Pol  $\alpha$ /primase is specifically involved in initiating new DNA strands. This four-subunit protein complex consists of a two-subunit DNA Pol  $\alpha$  and a two-subunit primase. After the primase synthesizes an RNA primer, the resulting RNA primer:template junction is immediately handed off to the associated DNA Pol  $\alpha$  to initiate DNA synthesis.

Because of its relatively low processivity, DNA Pol  $\alpha$ /primase is rapidly replaced by the highly processive DNA Pol  $\delta$  and Pol  $\epsilon$ . The process of replacing DNA Pol  $\alpha$ /primase with DNA Pol  $\delta$  or Pol  $\epsilon$  is called **polymerase switching** (Fig. 9-18) and results in three different DNA polymerases functioning at the eukaryotic replication fork. DNA Pol  $\delta$  and  $\epsilon$  are specialized to synthesize different strands at the replication fork, with DNA Pol  $\epsilon$  synthesizing the leading strand and DNA Pol  $\delta$  the lagging strand. As in bacterial cells, the majority of the remaining eukaryotic DNA polymerases are involved in DNA repair.

### Sliding Clamps Dramatically Increase DNA Polymerase Processivity

High processivity at the replication fork ensures rapid chromosome duplication. As we have discussed, DNA polymerases at the replication fork synthesize thousands to millions of base pairs without releasing from the template. Despite this, when looked at in the absence of other proteins, the DNA polymerases that act at the replication fork are only able to synthesize 20–100 bp



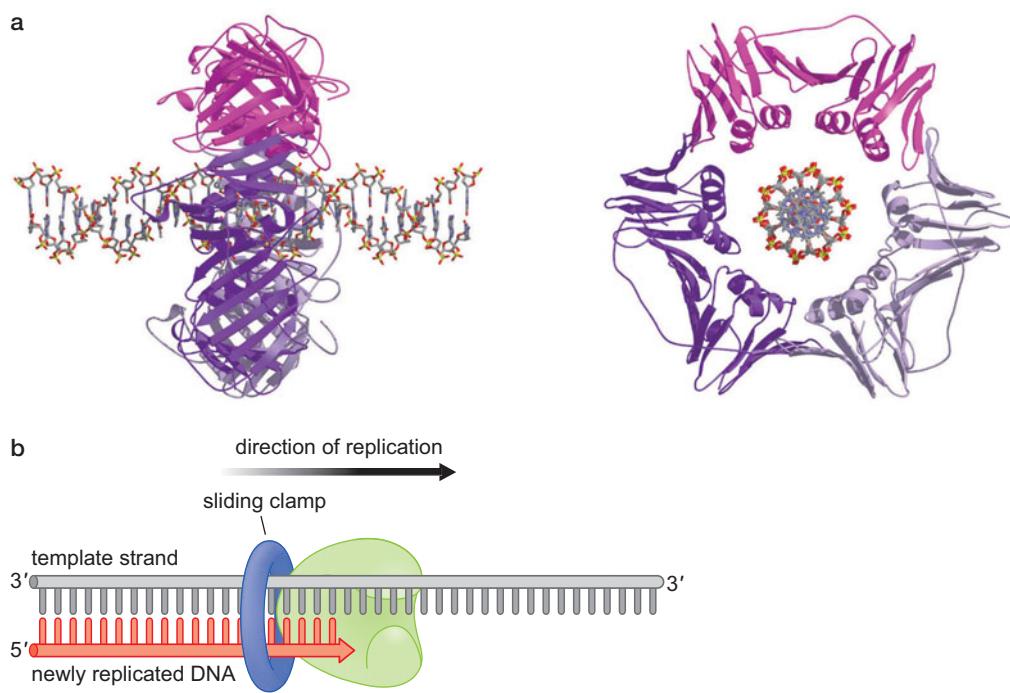
**FIGURE 9-18** DNA polymerase switching during eukaryotic DNA replication.

The order of eukaryotic DNA polymerase function is illustrated. The length of the DNA synthesized is shorter than in reality for illustrative purposes. Typically, the combined DNA Pol  $\alpha$ /primase product is between 50 and 100 bp, and the further extension by Pol  $\delta$  or Pol  $\epsilon$  is between 100 and 10,000 nucleotides. Although both DNA Pol  $\delta$  and  $\epsilon$  can substitute for DNA Pol  $\alpha$ /primase, recent studies indicate that DNA Pol  $\epsilon$  substitutes on the leading-strand template and DNA Pol  $\delta$  substitutes on the lagging-strand template.

before releasing from the template. How is the processivity of these enzymes increased so dramatically at the replication fork?

One key to the high processivity of the DNA polymerases that act at replication forks is their association with proteins called **sliding DNA clamps**. These proteins are composed of multiple identical subunits that assemble in the shape of a “doughnut.” The hole in the center of the clamp is large enough to encircle the DNA double helix and leave room for a layer of one or two water molecules between the DNA and the protein (Fig. 9-19a) (see Structural Tutorial 9-2). These properties allow the clamp proteins to slide along the DNA without dissociating from it. Importantly, sliding DNA clamps also bind tightly to DNA polymerases bound to primer:template junctions (Fig. 9-19b). The resulting complex between the polymerase and the sliding clamp moves efficiently along the DNA template during DNA synthesis.

How does the association with the sliding clamp change the processivity of the DNA polymerase? In the absence of the sliding clamp, a DNA polymerase dissociates and diffuses away from the template DNA on average once every 20–100 bp synthesized. In the presence of the sliding clamp, the DNA polymerase still disengages its active site from the 3'-OH end of the DNA frequently, but the association with the sliding clamp prevents the polymerase from diffusing away from the DNA (Fig. 9-20). By keeping the DNA polymerase in close proximity to the DNA, the sliding clamp

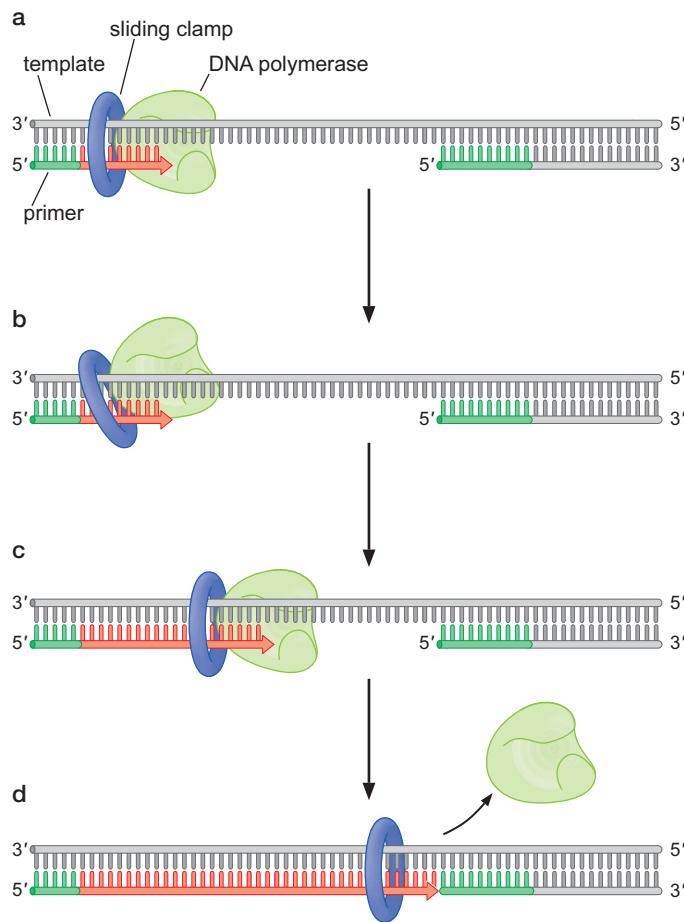


**FIGURE 9-19** Structure of a sliding DNA clamp. (a) 3D structure of a sliding DNA clamp associated with DNA. The opening through the center of the sliding clamp is  $\sim 35 \text{ \AA}$ , and the width of the DNA helix is  $\sim 20 \text{ \AA}$ . This provides enough space to allow a thin layer of one or two water molecules between the sliding clamp and the DNA. This is thought to allow the clamp to slide along the DNA easily. (Adapted from Krishna T.S. et al. 1994. *Cell* **79**: 1233–1243. Image prepared with MolScript, BobScript, and Raster3D. DNA modeled by Leemor Joshua-Tor.) (b) Sliding DNA clamps encircle the newly replicated DNA produced by an associated DNA polymerase. The sliding clamp interacts with the part of the DNA polymerase that is closest to the newly synthesized DNA as it emerges from the DNA polymerase.

ensures that the DNA polymerase rapidly rebinds *the same* primer:template junction, vastly increasing the processivity of the DNA polymerase.

Once an ssDNA template has directed synthesis of its complementary DNA strand, the DNA polymerase must release from the completed dsDNA and the sliding clamp to act at a new primer:template junction. This release is accomplished by a change in the affinity between the DNA polymerase and the sliding clamp that depends on the bound DNA. DNA polymerase bound to a primer:template junction has a high affinity for the clamp. In contrast, when a DNA polymerase reaches the end of an ssDNA template (e.g., at the end of an Okazaki fragment), the presence of dsDNA in its active site results in a change in conformation that reduces the polymerase's affinity for the sliding clamp and the DNA (see Fig. 9-20). Thus, when a polymerase completes the replication of a stretch of DNA, it is released from the sliding clamp so that it can act at a new primer:template junction.

Once released from a DNA polymerase, sliding clamps are not immediately removed from the replicated DNA. Instead, other proteins that function at the site of recent DNA synthesis interact with the clamp proteins. As described in Chapter 8, enzymes that assemble chromatin in eukaryotic cells are recruited to the sites of DNA replication by an interaction with the eukaryotic sliding DNA clamp (called “PCNA”). Similarly, eukaryotic proteins involved in Okazaki fragment repair also interact with sliding clamp proteins. In each case, by interacting with sliding clamps, these proteins accumulate at sites of new DNA synthesis where they are needed most.

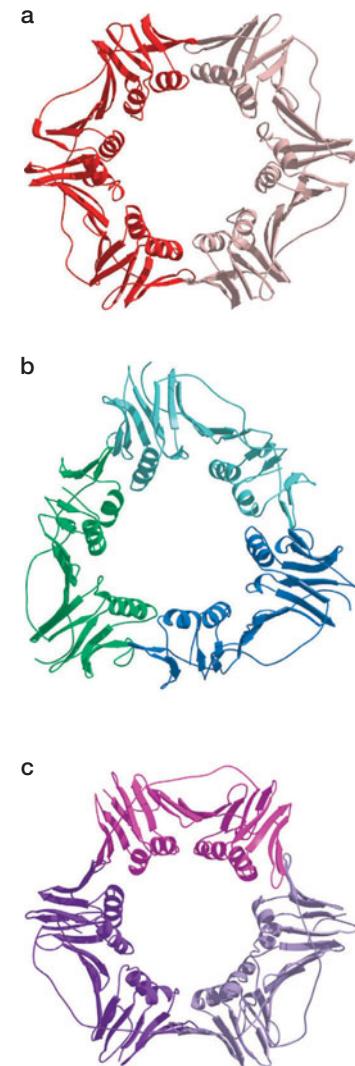


**FIGURE 9-20 Sliding DNA clamps increase the processivity of associated DNA polymerases.** (a) The sliding DNA clamp encircles the DNA and simultaneously binds the DNA polymerase. (b) The relatively low processivity of DNA polymerases leads to frequent release from the primer:template junction, but the association of the polymerase with the sliding clamp prevents diffusion away from the DNA. (c) The association of DNA polymerase with the sliding clamp ensures that the DNA polymerase rebinds the same primer:template junction and resumes DNA synthesis. (d) After DNA polymerase has completed synthesis of the template, the absence of a primer:template junction causes a change in the DNA polymerase that releases it from the sliding clamp.

Sliding clamp proteins are a conserved part of the DNA replication apparatus derived from organisms as diverse as viruses, bacteria, yeast, and humans. Consistent with their conserved function, the structure of sliding clamps derived from these different organisms is also conserved (Fig. 9-21). In each case, the clamp has the same sixfold symmetry and the same diameter. Despite the similarity in overall structure, the number of subunits that come together to form the clamp differs.

### Sliding Clamps Are Opened and Placed on DNA by Clamp Loaders

The sliding clamp is a closed ring in solution but must open to encircle the DNA double helix. A special class of protein complexes, called **sliding clamp loaders**, catalyze the opening and placement of sliding clamps on the DNA. These enzymes couple ATP binding and hydrolysis to the placement of the sliding clamp around primer:template junctions on the DNA (see Box 9-3, ATP Control of Protein Function: Loading a Sliding Clamp). The clamp loader also removes sliding clamps from the DNA when they



**FIGURE 9-21 3D structure of sliding DNA clamps isolated from different organisms.** Sliding DNA clamps are found across all organisms and share a similar structure. (a) The sliding DNA clamp from *E. coli* is composed of two copies of the  $\beta$  protein. (Adapted from Kong X.P. et al. 1992. *Cell* **69**: 425–437.) (b) The T4 phage sliding DNA clamp is a trimer of the gp45 protein. (Adapted from Moarefi I. et al. 2000. *J. Mol. Biol.* **296**: 1215–1223.) (c) The eukaryotic sliding DNA clamp is a trimer of the PCNA protein. (Adapted from Krishna T.S. et al. 1994. *Cell* **79**: 1233–1243. Images prepared with MolScript, BobScript, and Raster3D.)

## ► ADVANCED CONCEPTS

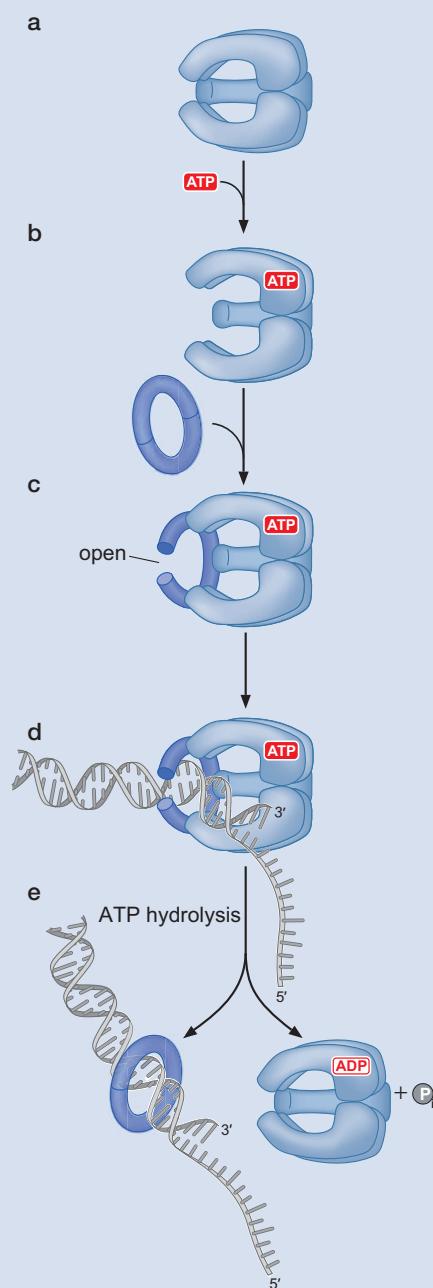
**Box 9-3 ATP Control of Protein Function: Loading a Sliding Clamp**

The five subunits that come together to form a sliding DNA clamp loader are all members of a large class of proteins that have a related ATP-binding and hydrolysis site called AAA<sup>+</sup> proteins. These proteins assemble into multi-AAA<sup>+</sup> protein assemblies that use the energy of ATP binding and hydrolysis to alter the structure of target proteins or DNA. Sliding DNA clamp loaders are formed from several different AAA<sup>+</sup> proteins, but other AAA<sup>+</sup> protein complexes are composed of multiple copies of the same AAA<sup>+</sup> protein. Although AAA<sup>+</sup> proteins have related ATP-binding and hydrolysis sites, they perform diverse functions. In addition to loading sliding clamps, AAA<sup>+</sup> protein complexes unwind DNA, load DNA helicase around substrate DNA (e.g., ORC and DnaC, as we shall see when we discuss the initiation of DNA replication), unfold proteins, and disassemble protein complexes. Indeed, it is the diversity of their functions that led to their AAA name: ATPases—associated with various activities.

How are ATP binding and hydrolysis coupled to sliding clamp loading? The initial events of clamp loading require ATP binding to the loader. When bound to ATP, the clamp loader can bind and open the sliding clamp ring by causing one of the subunit:subunit interfaces to come apart (Box 9-3 Fig. 1). The now open sliding clamp is brought to the DNA through a high-affinity DNA-binding site on the clamp loader. Like sliding clamp binding, DNA binding requires that the clamp loader be bound to ATP. Consistent with the need for sliding clamps at the sites of DNA synthesis, the clamp loader's DNA-binding site specifically recognizes primer:template junctions. The DNA is bound in such a way that the open sliding clamp is placed around the double-stranded region of the primer:template junction.

The final steps in sliding clamp loading are stimulated by ATP hydrolysis. Binding of the clamp loader to the primer:template

junction activates ATP hydrolysis (by the clamp loader). Because the clamp loader can only bind the sliding clamp and DNA when it is bound to ATP (but not ADP), hydrolysis causes the clamp loader to release the sliding clamp and disassociate from the DNA. Once released from the clamp loader, the sliding clamp spontaneously closes around the DNA. The net result of this process is the loading of the sliding clamp at the site of DNA polymerase action—the primer:template junction. Release of ADP and P<sub>i</sub> and binding to a new ATP molecule allow the clamp loader to initiate a new cycle of loading.



**BOX 9-3 FIGURE 1** ATP control of sliding DNA clamp loading. (a) Sliding clamp loaders are five-subunit protein complexes whose activity is controlled by ATP binding and hydrolysis. In *E. coli*, the clamp loader is called the  $\gamma$  complex, and in eukaryotic cells, it is called replication factor C (RF-C). (b) To catalyze the sliding clamp opening, the clamp loader must be bound to ATP. (c) Once bound to ATP, the clamp loader binds the clamp and opens the ring at one of the subunit:subunit interfaces. (d) The resulting complex can now bind to DNA. DNA binding is mediated by the clamp loader, which preferentially binds to primer:template junctions. Correct binding to the DNA has two consequences. First, the opened sliding clamp is positioned so that dsDNA is in what will be the “hole” of the clamp. Second, DNA binding stimulates ATP hydrolysis by the clamp loader. (e) Because only an ATP-bound clamp loader can bind to the clamp and to DNA, the ADP form of the clamp loader rapidly disassociates from the clamp and the DNA, leaving behind a closed clamp positioned around the dsDNA portion of the primer:template junction. (Adapted, with permission, from O'Donnell M. et al. 2001. *Curr. Biol.* **11**: R935–R946, Fig. 5. © Elsevier.)

**Box 9-3** (Continued)

The function of the clamp loader illustrates several general features of the coupling of ATP binding and hydrolysis to a molecular event. ATP binding to a protein typically is involved in the **assembly stage** of the event: the association of the protein with the target molecule. For example, the clamp loader has two target molecules—the sliding clamp and the primer:template junction. ATP is required for the clamp loader to bind to either target. Similarly, ATP binding stimulates the ability of DNA helicases to bind to ssDNA. In each case, the events coupled to ATP binding could be considered the action part of the cycle. For the clamp loader, ATP binding but not ATP hydrolysis is required to open the sliding clamp ring. For the DNA helicase, binding ssDNA is likely to be the key event unwinding DNA. In these cases, binding to ATP stabilizes a conformation of the enzyme that favors interaction with the substrate in a particular conformation.

What is the role of ATP hydrolysis? ATP hydrolysis typically leads to the **disassembly stage** of the event: releasing the bound targets from the enzyme. Once the ATP-stabilized complex is formed, it must be disassembled. This could occur by simple disassociation; however, more often than not, this process would return the components to their starting situation (e.g., the sliding clamp free in solution), and this process would be slow if the ATP-stabilized complex were tightly associated. To

ensure that disassembly occurs at the appropriate time, place, and rate, ATP hydrolysis is used to initiate disassembly. For example, ATP hydrolysis causes the clamp loader to revert back to a state in which it cannot bind either the sliding clamp or DNA. Reversion to this ground state may occur while the enzyme is still bound to the products of ATP hydrolysis (ADP and P<sub>i</sub>) or may require their release.

The final key mechanism to couple ATP hydrolysis to a reaction pertains to the **trigger for ATP hydrolysis**. It is critical that the factor not hydrolyze ATP until a desired complex is assembled. Typically, formation of a particular complex triggers ATP hydrolysis. In the case of the clamp loader, this complex is the tertiary complex of the sliding clamp, the clamp loader, and the primer:template junction.

ATP control of these molecular events is thus most directly related to controlling the timing of conformational changes by the enzyme. By requiring the enzyme to alternate between two conformational states in order and requiring the formation of a key intermediate to trigger ATP hydrolysis, the enzyme can accomplish work. In contrast, if the enzyme merely bound and released ATP (without hydrolysis), the reaction would return to the initial state as often as it would proceed forward, and little, if any, work would be accomplished.

are no longer in use, although this does not require ATP hydrolysis. Like DNA helicases and topoisomerases, these enzymes alter the conformation of their target (the sliding clamp) but not its chemical composition.

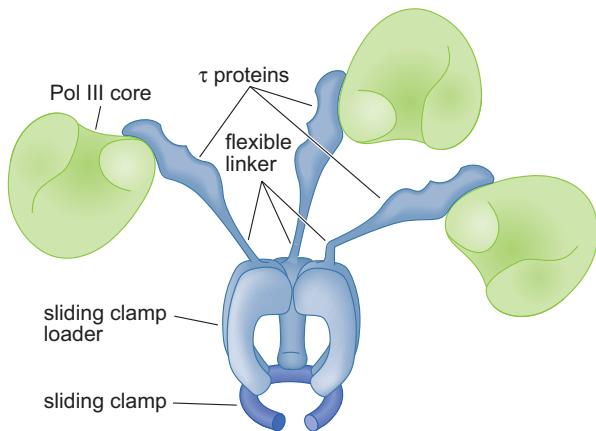
What controls when sliding clamps are loaded and removed from the DNA? Loading of a sliding clamp occurs anytime a primer: template junction is present in the cell. These DNA structures are formed not only during DNA replication, but also during several DNA-repair events (see Chapter 10). A sliding clamp can only be removed from the DNA if it is not bound by another protein. Sliding clamp loaders and DNA polymerases cannot interact with a sliding clamp at the same time because they have overlapping binding sites on the sliding clamp. Thus, a sliding clamp that is bound to a DNA polymerase is not subject to removal from the DNA. Similarly, nucleosome assembly factors, Okazaki fragment repair proteins, and other DNA-repair proteins all interact with the same region of the sliding clamp as the clamp loader. Thus, sliding clamps are only removed from the DNA once all of the enzymes that interact with them have completed their function.

## DNA SYNTHESIS AT THE REPLICATION FORK

At the replication fork, the leading and lagging strands are synthesized simultaneously. This has the important benefit of limiting the amount of ssDNA present in the cell during DNA replication. When an ssDNA region of DNA is broken, there is a complete break in the chromosome that is much more difficult to repair than an ssDNA break in a dsDNA region. Moreover, repair of this type of lesion frequently leads to mutation of the DNA (see Chapter 10). Thus, limiting the time DNA is single-stranded is crucial. To

**FIGURE 9-22 Composition of the DNA Pol III holoenzyme.**

**Pol III holoenzyme.** There are four enzymes in each copy of the DNA Pol III holoenzyme: three copies of the DNA Pol III core enzyme and one copy of the sliding clamp holder. The sliding clamp holder includes three copies of the  $\tau$  protein, each of which interacts with one DNA Pol III core. Analysis of the amino acid sequence of the  $\tau$  protein indicates that the DNA Pol III-binding region of the protein is separated from the part of the protein involved in clamp loading by an extended flexible linker. This linker is proposed to allow the associated polymerases to move in a relatively independent manner that would be necessary for one polymerase to replicate the leading strand and the other two polymerases to replicate the lagging strand. (Adapted, with permission, from O'Donnell M. et al. 2001. *Curr. Biol.* 11: R935–R946, Fig. 6. © Elsevier.)



coordinate the replication of both DNA strands, multiple DNA polymerases function at the replication fork.

In *E. coli*, the coordinate action of these polymerases is facilitated by physically linking them together in a large multiprotein complex called the “DNA Pol III holoenzyme” (Fig. 9-22). *Holoenzyme* is a general name for a multiprotein complex in which a core enzyme activity is associated with additional components that enhance function. The DNA Pol III holoenzyme includes three copies of the “core” DNA Pol III enzyme and one copy of the five-subunit sliding clamp loader. Although present in only one copy in the holoenzyme, the sliding clamp loader includes three copies of  $\tau$  protein, each of which binds one DNA Pol III core enzyme (see Fig. 9-22).

How do the multiple DNA polymerases remain linked at the replication fork while synthesizing DNA on both the leading and lagging template strands? A model that explains this coupling proposes that the replication machinery exploits the flexibility of DNA and the  $\tau$  protein (Fig. 9-23). As the helicase unwinds the DNA at the replication fork, the leading-strand template is exposed and acted on immediately by one DNA Pol III core enzyme, which synthesizes a continuous strand of complementary DNA. In contrast, the lagging-strand template is not immediately acted on by DNA polymerase. Instead, it is spooled out as ssDNA that is rapidly bound by SSBs. Intermittently, primase interacts with the DNA helicase and is

**FIGURE 9-23 “Trombone” model for coordinating replication by two DNA polymerases at the *E. coli* replication fork.** (a) The DNA helicase at the *E. coli* DNA replication fork travels on the lagging-strand template in a  $5' \rightarrow 3'$  direction. The DNA Pol III holoenzyme interacts with the DNA helicase through the  $\tau$  subunits, which also bind the DNA polymerase III core proteins. One DNA Pol III core is replicating the leading strand while the other two DNA Pol III core enzymes are dedicated to replication of the lagging strand. SSBs coat the ssDNA regions of the DNA (for simplicity, SSBs on the lagging strand are only shown in part a). (b) Periodically, DNA primase associates with the DNA helicase and synthesizes a new RNA primer on the lagging-strand template. (c) Immediately after a new RNA primer is synthesized, the sliding DNA clamp loader assembles a sliding DNA clamp at the resulting primer:template junction. (d) The unengaged “second” lagging-strand DNA polymerase rapidly recognizes the loaded sliding DNA clamp at the primer:template junction and initiates a new Okazaki fragment. (e) When the “first” lagging-strand DNA polymerase reaches the end of the Okazaki strand template, it is released from the sliding clamp. This “first” lagging-strand DNA polymerase is now ready to recognize the next RNA primer/sliding clamp that assembles on the lagging-strand template. Thus, this model envisions two lagging-strand DNA polymerases alternately initiating synthesis of new Okazaki fragments.

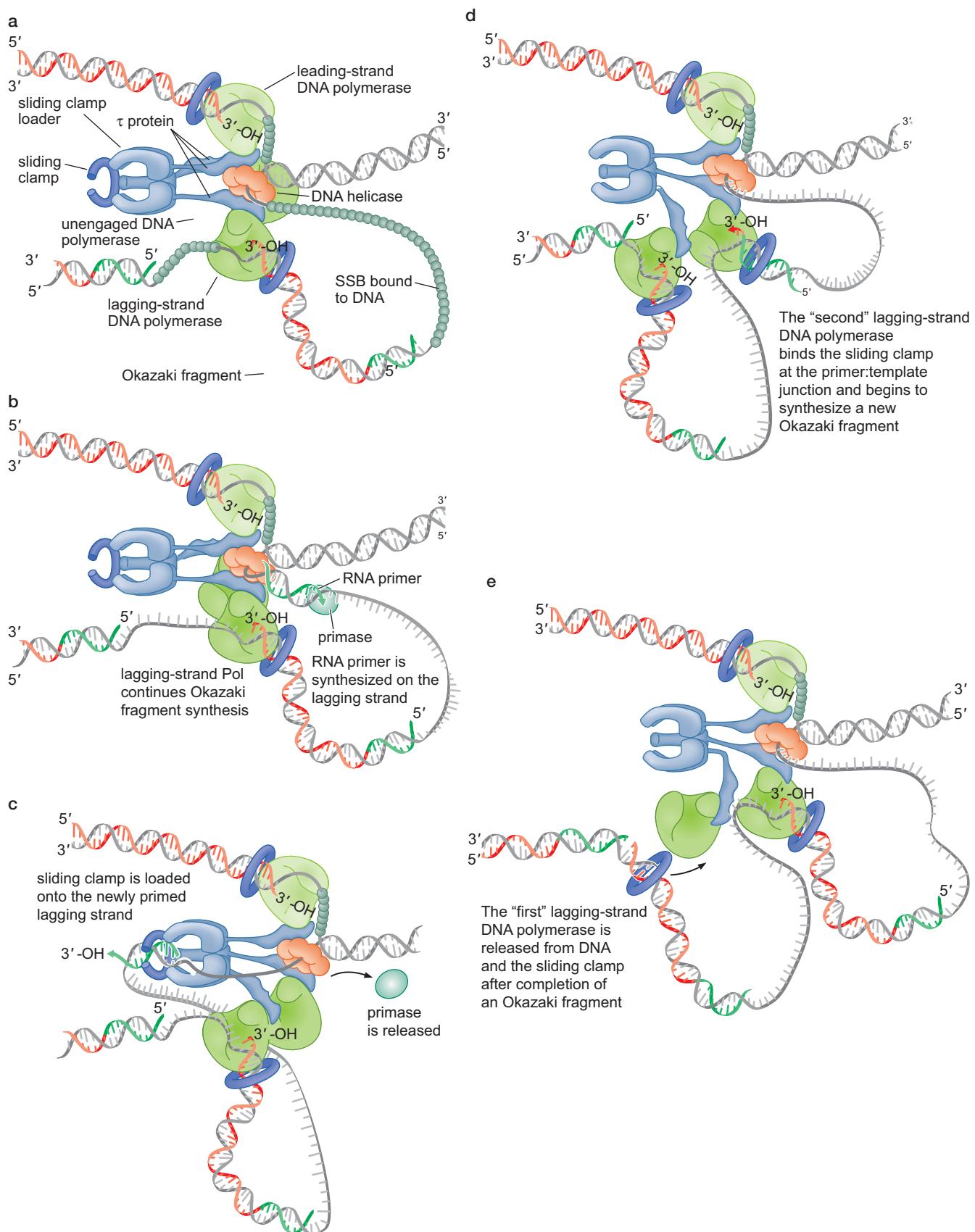


FIGURE 9-23 (See facing page for legend.)

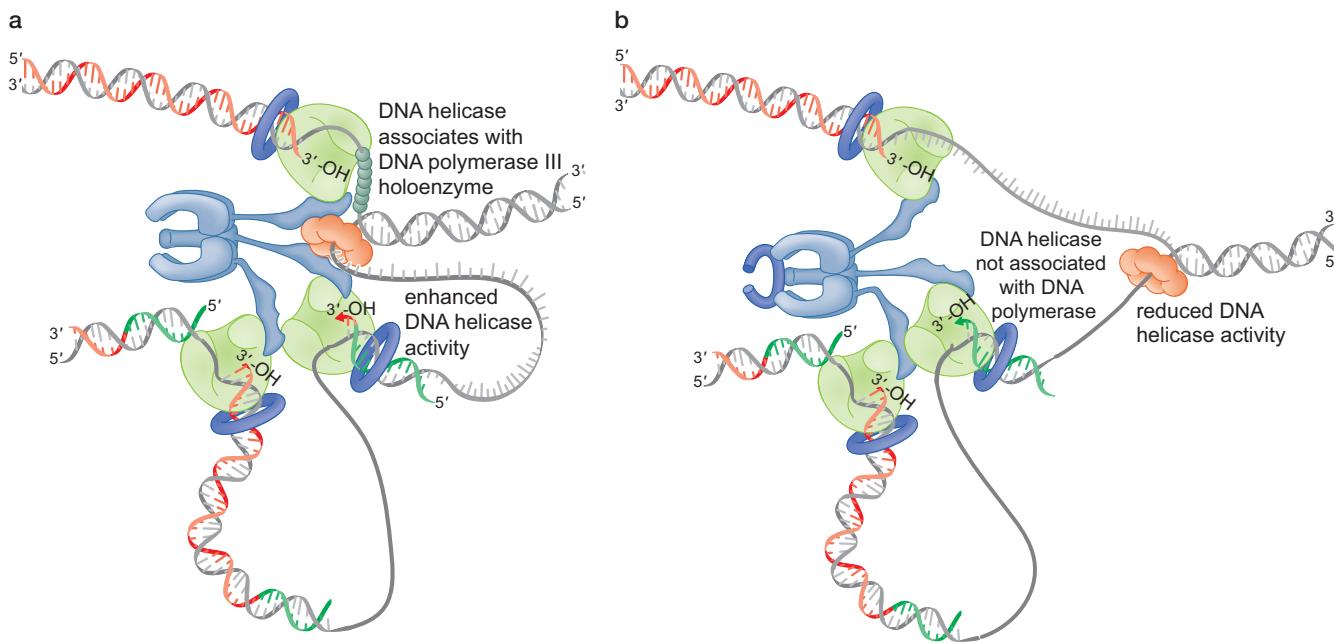
activated to synthesize a new RNA primer on the lagging-strand template. The resulting RNA:DNA hybrid is recognized as a primer:template junction by the sliding DNA clamp loader, a sliding clamp is assembled at this site, and a second DNA Pol III enzyme initiates lagging-strand synthesis.

As one lagging-strand DNA polymerase synthesizes an Okazaki fragment, additional ssDNA is generated by the helicase and a new RNA primer is synthesized on this template. As with the previous lagging-strand primer, the new RNA primer is recognized by the sliding clamp loader. Although it has traditionally been thought that there are only two DNA Pol III core enzymes within the DNA Pol III holoenzyme, recent studies support the presence of a third DNA Pol III. The third DNA Pol III initiates synthesis of a new Okazaki fragment as soon as a sliding DNA clamp is assembled on the RNA primer, likely before the completion of the previous Okazaki fragment. Thus, a second Okazaki fragment is thought to be initiated before the release of the polymerase synthesizing the previous Okazaki fragment. When each Okazaki fragment is completed, the responsible DNA polymerase is released from the template. (Recall that once DNA polymerase completes synthesis of an Okazaki fragment, it is released from its associated sliding clamp.) Because release of the DNA polymerase from the sliding clamp is a slower process than DNA synthesis, having a second DNA polymerase dedicated to lagging-strand DNA synthesis ensures that lagging-strand synthesis is continuous even during this slow polymerase release event. Because the released DNA polymerase III core enzyme remains tethered to the helicase via the  $\tau$  subunit of the sliding clamp loader, this polymerase is in an ideal position to bind the next RNA primer:template junction immediately after the addition of a sliding clamp. The model described is known as the “trombone model” in reference to the changing size of the ssDNA loop formed between the DNA polymerase(s) and the DNA helicase on the lagging-strand template.

DNA replication in eukaryotic cells also requires three DNA polymerases: DNA Pol  $\alpha$ /primase, DNA Pol  $\delta$ , and DNA Pol  $\epsilon$  (see Fig. 9-18). DNA Pol  $\alpha$ /primase initiates new strands, and DNA Pol  $\delta$  and Pol  $\epsilon$  extend these strands. As in *E. coli*, one polymerase (DNA Pol  $\epsilon$ ) is dedicated to the leading strand, and two (DNA Pol  $\alpha$ /primase and DNA Pol  $\delta$ ) are dedicated to the lagging strand (although DNA Pol  $\alpha$ /primase also primes the leading strand, it functions many more times on the lagging strand). Several additional proteins are known to be part of the eukaryotic replication fork. The functions of these additional proteins are currently poorly understood; however, it is likely that they act to coordinate the three DNA polymerases and couple their action to the eukaryotic DNA helicase (the Mcm2-7 complex). Unlike the situation in prokaryotic cells, the eukaryotic sliding clamp loader, RF-C, does not appear to perform these functions.

### Interactions between Replication Fork Proteins Form the *E. coli* Replisome

The connections between the components of the DNA Pol III holoenzyme are not the only interactions that occur between the components of the bacterial replication fork. Additional protein–protein interactions between replication fork proteins facilitate rapid replication fork progression. The most important of these is an interaction between the DNA helicase (the hexameric dnaB protein) (see Table 9-1) and the DNA Pol III holoenzyme (Fig. 9-24). This interaction, which is mediated by the  $\tau$  subunit of the clamp loader component of the holoenzyme, holds the helicase and the DNA Pol III holoenzyme together. In addition, this association stimulates the activity of the helicase by increasing the rate of helicase movement 10-fold. Thus, the



**FIGURE 9-24** Binding of the DNA helicase to DNA Pol III holoenzyme stimulates the rate of DNA strand separation. The  $\tau$  subunit of the sliding clamp loader interacts with both the DNA helicase and the DNA polymerase at the replication fork. (a) When this interaction occurs, the DNA helicase unwinds the DNA at approximately the same rate as the DNA polymerases replicate the DNA. (b) If the DNA helicase is not associated with DNA Pol III holoenzyme, DNA unwinding slows by 10-fold. Under these conditions, the DNA polymerases can replicate faster than the DNA helicase can separate the strands of unreplicated DNA. This allows the DNA Pol III holoenzyme to “catch up” to the DNA helicase and re-form the replisome.

DNA helicase slows down if it becomes separated from the DNA polymerase (see Fig. 9-24). The coupling of helicase activity to the presence of DNA Pol III prevents the helicase from “running away” from the DNA Pol III holoenzyme and thus serves to coordinate these two key replication fork enzymes.

A second important protein–protein interaction occurs between the DNA helicase and primase. Unlike most proteins that act at the *E. coli* replication fork, primase is not tightly associated with the fork. Instead, at an interval of about once per second, primase associates with the helicase and SSB-coated ssDNA and synthesizes a new RNA primer. Although the interaction between the DNA helicase and primase is relatively weak, this interaction strongly stimulates primase function (about 1000-fold). After an RNA primer is synthesized, the primase is released from the DNA helicase into solution.

The relatively weak interaction between the *E. coli* primase and DNA helicase is important for regulating the length of Okazaki fragments. A tighter association would result in more frequent primer synthesis on the lagging strand and therefore shorter Okazaki fragments. Similarly, a weaker interaction would result in longer Okazaki fragments.

The combination of all of the proteins that function at the replication fork is referred to as the **replisome**. Together, these proteins form a finely tuned factory for DNA synthesis that contains multiple interacting machines. Individually, these machines perform important specific functions. When brought together, their activities are coordinated by the interactions between them. Although these interactions are particularly well-understood in *E. coli* cells, studies of bacteriophage and eukaryotic DNA replication machinery show that a similar coordination between multiple machines is involved in DNA replication in these organisms. Indeed, there are clear parallels between the

proteins known to be involved in replication in *E. coli* and those functioning in these other organisms. Table 9-1 presents a list of factors performing analogous functions in phage, prokaryotic, and eukaryotic DNA replication.

To fully appreciate the amazing capabilities of the enzymes that replicate DNA, imagine a situation in which a DNA base is the size of this textbook. Under these conditions, dsDNA would be  $\sim 1$  m in diameter and the *E. coli* genome would be a large circle,  $\sim 500$  miles (800 km) in circumference. More importantly, the replisome would be the size of a delivery truck and would be moving at more than 375 mph (600 km/h)! Replicating the *E. coli* genome would be a 40-min, 250-mile (400 km) trip for two such machines, each leaving two 1-m DNA cables in their wake. Impressively, during this trip, the replication machinery would, on average, make only a single error.

## INITIATION OF DNA REPLICATION

### Specific Genomic DNA Sequences Direct the Initiation of DNA Replication

The initial formation of a replication fork requires the separation of the two strands of the DNA duplex to provide the ssDNA necessary for DNA helicase binding and to act as a template for the synthesis of both the RNA primer and new DNA. Although DNA strand separation (also called “DNA unwinding”) is most easily accomplished at chromosome ends, DNA synthesis generally initiates at internal regions. Indeed, for circular chromosomes, the lack of chromosome ends makes internal DNA unwinding essential to replication initiation.

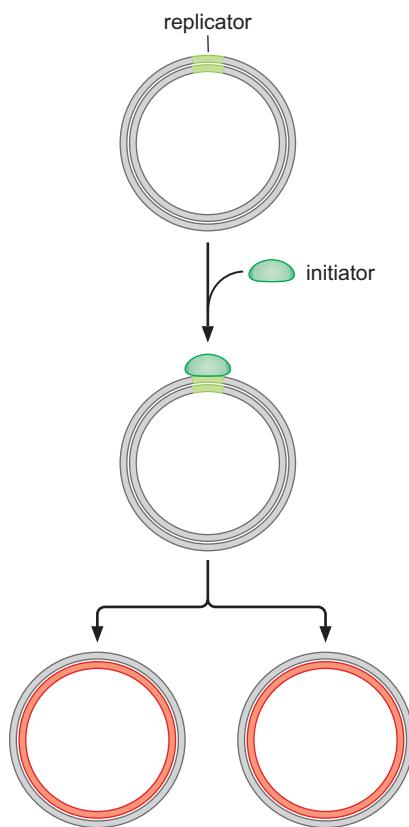
The specific sites at which DNA unwinding and initiation of replication occur are called **origins of replication**. Depending on the organism, there may be as few as one or as many as thousands of origins per chromosome.

### The Replicon Model of Replication Initiation

In 1963, François Jacob, Sydney Brenner, and Jacques Cuzin proposed a model to explain the events controlling the initiation of replication in bacteria. They defined all of the DNA replicated from a particular origin of replication as a **replicon**. For example, because the single chromosome found in *E. coli* cells has only one origin of replication, the entire chromosome is a single replicon. In contrast, the presence of multiple origins of replication divides each eukaryotic chromosome into multiple replicons—one for each origin of replication.

The replicon model proposed two components that controlled the initiation of replication: the replicator and the initiator (Fig. 9-25). The **replicator** is defined as the *cis*-acting DNA sequences that are *sufficient* to direct the initiation of DNA replication. This is in contrast to the origin of replication, which is the physical site on the DNA where the DNA is unwound and DNA synthesis initiates. Although the origin of replication is always part of the replicator, sometimes (particularly in eukaryotic cells) the origin of replication is only a fraction of the DNA sequences required to direct the initiation of replication (the replicator). The same distinction can be made between a transcriptional promoter and the start site of transcription, as we shall see in Chapter 13.

The second component of the replicon model is the **initiator** protein. This protein specifically recognizes a DNA element in the replicator and activates the initiation of replication (see Fig. 9-25). Initiator proteins have been identified in many different organisms, including bacteria, viruses, and



**FIGURE 9-25** The replicon model. Binding of the initiator to the replicator stimulates initiation of replication and the duplication of the associated DNA.

eukaryotic cells. All initiator proteins select the sites that will become origins of replication, although they are recruited to the DNA by different methods. Interestingly, all of the known initiator proteins are regulated by ATP binding and hydrolysis and share a common core  $\text{AAA}^+$  ATP-binding motif related to, but distinct from, that used by sliding DNA clamp loaders.

As we see later, the initiator protein is the only sequence-specific DNA-binding protein involved in the initiation of replication. The remaining proteins required for replication initiation do not bind to a DNA sequence specifically. Instead, these proteins are recruited to the replicator through a combination of protein–protein interactions and affinity for specific DNA structures (e.g., ssDNA or a primer:template junction). Indeed, for many eukaryotic cells even the initiator protein does not show sequence-specific DNA-binding activity.

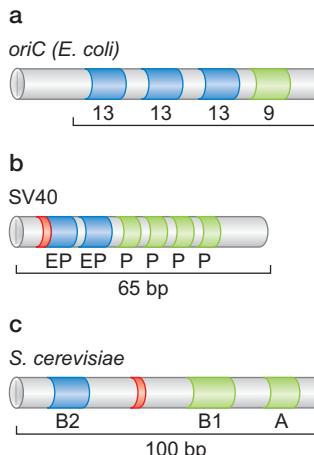
### Replicator Sequences Include Initiator-Binding Sites and Easily Unwound DNA

The DNA sequences of known replicators share two common features (Fig. 9-26). First, they include a binding site for the initiator protein that nucleates the assembly of the replication initiation machinery. Second, they include a stretch of AT-rich DNA that unwinds readily but not spontaneously. Unwinding of DNA at replicators is controlled by the replication initiation proteins, and the action of these proteins is tightly regulated in most organisms.

The single replicator required for *E. coli* chromosomal replication is called *oriC*. Two repeated motifs are critical for *oriC* function (Fig. 9-26a). The 9-mer motif is the binding site for the *E. coli* initiator, DnaA, and is repeated five times at *oriC*. The 13-mer motif, repeated three times, is the initial site of ssDNA formation during initiation.

Although the specific sequences are different, the overall structures of replicators derived from many eukaryotic viruses and the single-cell eukaryote *Saccharomyces cerevisiae* are similar (Fig. 9-26b,c). The methods used to define origins of replication are described in Box 9-4, The Identification of Origins of Replication and Replicators.

Replicators functioning in multicellular eukaryotes are not well-understood. Their identification and characterization have been hampered by the lack of genetic assays for stable propagation of small circular DNA comparable to those used to identify origins in single-cell eukaryotes and bacteria (see Box 9-4). In the few instances in which replicators have been identified, they are found to be larger than the replicators identified in



**FIGURE 9-26** Structure of replicators.

The DNA elements that make up three well-characterized replicators are shown. For each diagram: (green) the initiator DNA-binding site; (blue) DNA elements that facilitate DNA unwinding; (red) the site of initial DNA synthesis. (a) *oriC* is composed of five “9-mer” DnaA-binding sites and three “13-mer” repeated elements that are the site of initial DNA unwinding. (The site for *oriC* is outside the sequence shown.) (b) The origin of the eukaryotic virus SV40 is composed of four pentamer binding sites (P) for the initiator protein called large T antigen and a 20-bp early palindrome (EP) that is the site of DNA unwinding. (c) Three elements are commonly found at *S. cerevisiae* replicators. The A and B1 elements bind to the initiator ORC. The B2 element facilitates binding of the DNA helicase to the origin.

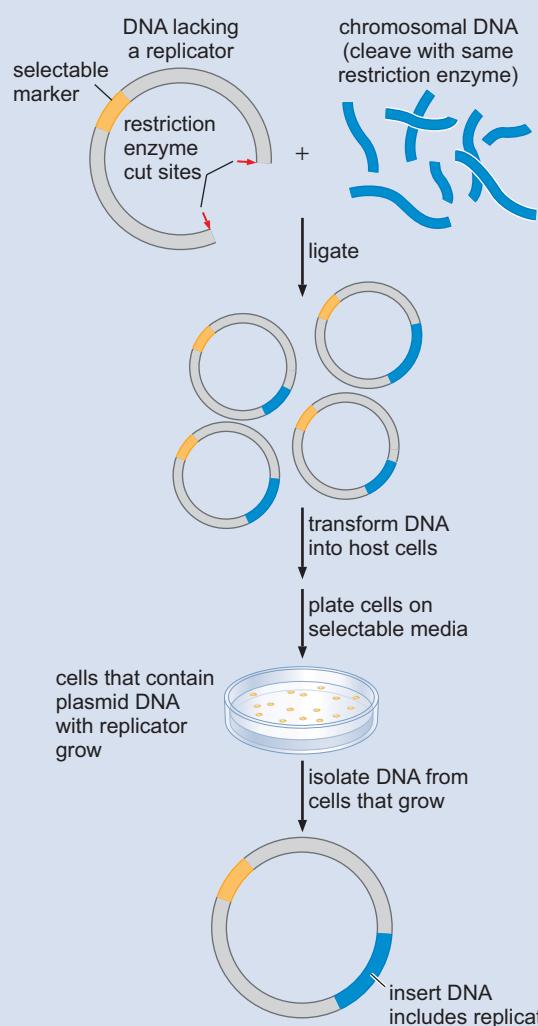
## ► KEY EXPERIMENTS

### Box 9-4 The Identification of Origins of Replication and Replicators

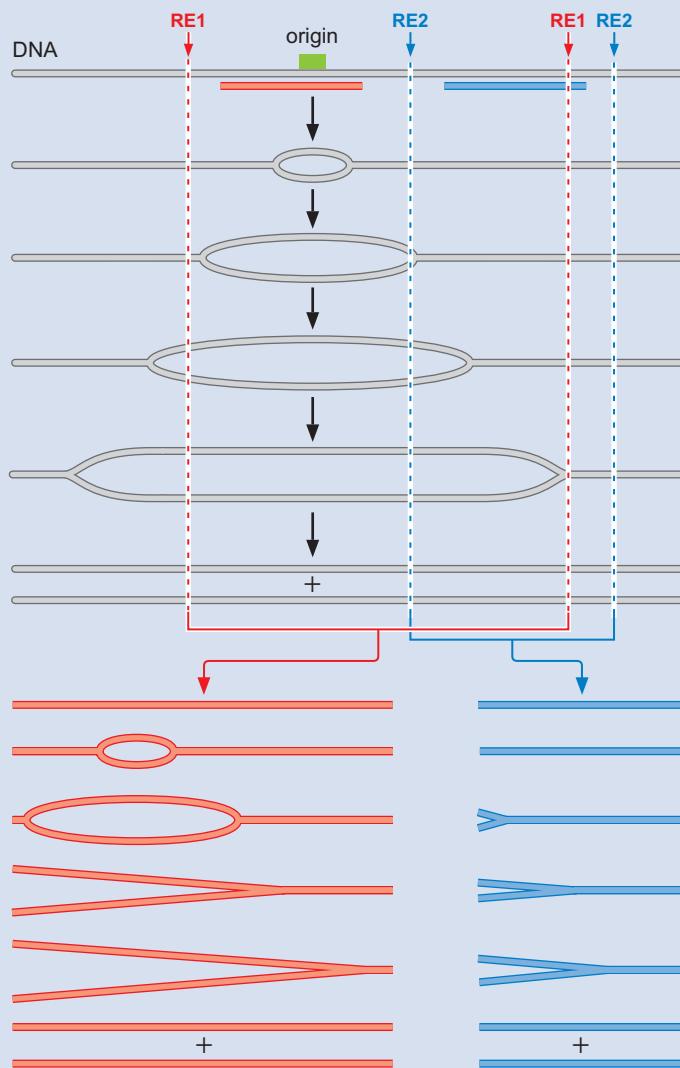
Replicator sequences are typically identified using genetic assays. For example, the first yeast replicators were identified using a DNA transformation assay (Box 9-4 Fig. 1). In these studies, investigators randomly cloned genomic DNA fragments into plasmids lacking a replicator but containing a selectable marker missing in the host cell. For the plasmid to be maintained in the host cell after transformation, the cloned DNA fragment must contain a yeast replicator. The identified DNA fragments were called **autonomously replicating sequences** (ARSs). Although these sequences acted as replicators in the artificial context of a circular plasmid, further evidence was required to show that these sequences were also replicators in their native chromosomal location.

To show that ARSs acted as replicators in the chromosome, it was necessary to develop methods to identify the location of origins of replication in the cell. One approach to identify origins takes advantage of the unusual structure of the DNA replication intermediates formed during replication initiation. Unlike either fully replicated or fully unreplicated DNA, DNA that is in the process of being replicated is not linear. For example, a DNA fragment (generated by cleavage of the DNA with a restriction enzyme) that does not contain an origin of replication will take on a variety of “Y-shaped” conformations as it is replicated (Box 9-4 Fig. 2, blue DNA fragments). Similarly, immediately after the initiation of replication, a DNA fragment containing an origin of replication will take on a “bubble” shape. Finally, if the origin of replication is located asymmetrically within the DNA fragment, the DNA will start out as a bubble shape and then convert to a Y shape (Box 9-4 Fig. 2, red DNA fragments). These unusually shaped DNAs can be distinguished from simple linear DNA using two-dimensional (2D) agarose gel electrophoresis (Box 9-4 Fig. 3).

To identify DNA that is in the process of replicating, DNA derived from dividing cells is first cut with a restriction enzyme and separated on a 2D agarose gel. In the first dimension, the DNA is separated primarily by size, but in the second dimension, the DNA is separated by size and shape. This is accomplished by using different density of agarose and electrophoresis rates for each dimension. To separate by size and shape, the agarose gel pores are small (high agarose density), and the rate of electrophoresis is fast. In contrast, to separate primarily by size, the agarose gel pores are larger (low agarose density), and the rate of electrophoresis is slower. Once electrophoresis is complete, the DNA molecules are transferred to nitrocellulose and detected by Southern blotting (see Chapter 7). The choice of the restriction enzyme and DNA probe used can dramatically affect the outcome of the analysis. In general, this method requires that the investigator already know the approximate location of a potential origin of replication. Recently, newer methods have been developed that use DNA microarrays to identify the location of origins and that do not require any prior knowledge concerning origin locations.



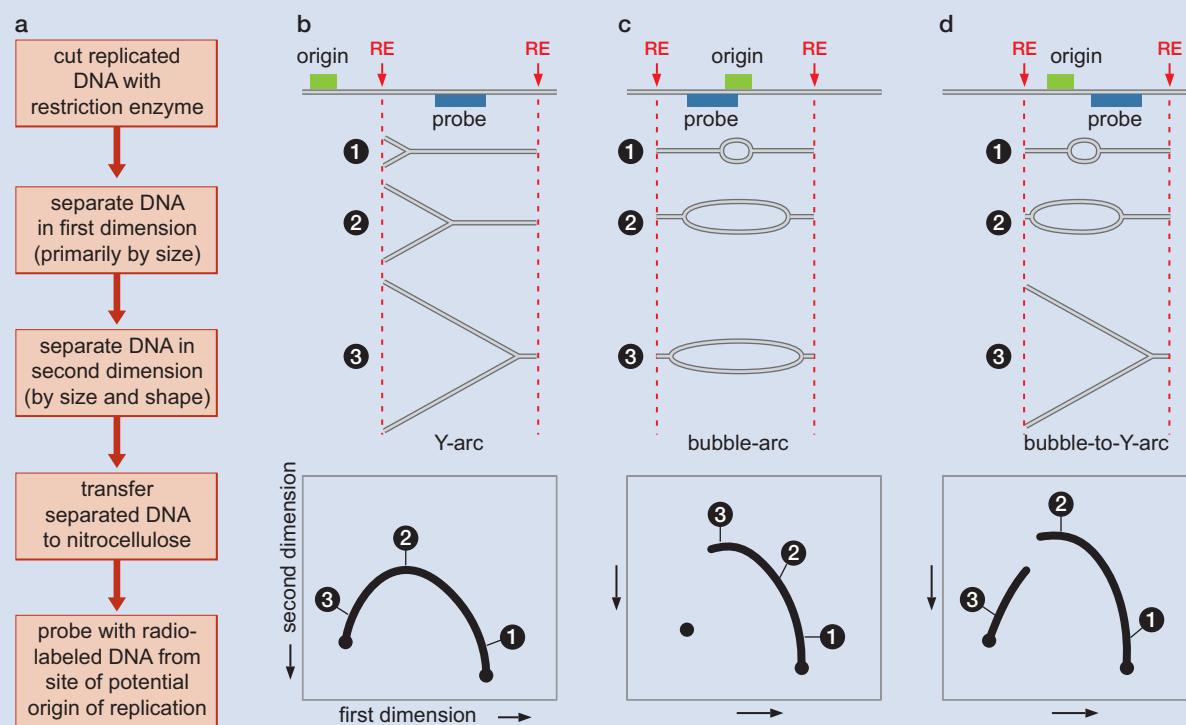
**BOX 9-4 FIGURE 1** Genetic identification of replicators. A plasmid (a small circular DNA molecule) containing a selectable marker is cut with a restriction enzyme that results in the excision of the plasmid's normal replicator. This leaves a DNA fragment that lacks a replicator. To isolate a replicator from a particular organism, the DNA from that organism is cut with the same restriction enzyme and ligated into the cut plasmid to re-create circular plasmids, each including a single fragment derived from the test organism. This DNA is then transformed into the host organism, and the recombinant plasmids are selected using a selectable marker on the plasmid (e.g., if the marker conferred antibiotic resistance, the cells would be grown in the presence of the antibiotic). Cells that grow are able to maintain the plasmid and its selectable marker, indicating that the plasmid can replicate in the cell and must contain a replicator. Isolation of the plasmid from the host cell and sequencing of the inserted DNA allow the identification of the sequence of the fragment that contains the replicator. Further mutagenesis of the inserted DNA (such as deletion of specific regions of the inserted DNA), followed by a repetition of the assay, allows a more precise definition of the replicator.

**Box 9-4** (Continued)

**BOX 9-4 FIGURE 2** DNA that is in the process of replication has an unusual structure. Results of restriction enzyme cleavage of DNA in the process of replication are shown. The illustration shows the growth of a “replication bubble” (created by two replication forks progressing away from an origin of replication). The consequences of cutting these replication intermediates is followed by detection by hybridization with the indicated labeled DNA probe. If the red restriction enzyme is used and only the fragments that hybridize to the red DNA probe are examined, the pattern on the left side will be generated. If the blue restriction enzyme and the blue DNA probe is used to detect the resulting DNA fragments, the pattern on the right will be observed. Note that the left-hand pattern starts with a DNA fragment containing a “bubble” and eventually ends with “Y-shaped” molecules. The right-hand pattern never has a “bubble” but does assume a full variety of “Y-shaped” intermediates. Only a DNA fragment containing an origin of replication can produce the pattern on the left.

How can the 2D gels identify the DNA intermediates associated with a replication origin? The particular pattern of DNA migration can lead to unequivocal evidence of an origin of replication. The most unusual structures migrate most slowly in the first dimension. For example, a Y-shaped molecule that has three equal length arms will migrate the most slowly of all such molecules derived from a particular DNA fragment (Box 9-4 Fig. 3b), and therefore will be at the top of an arc of DNA molecules that are nonlinear. In contrast, a Y-shaped molecule with two very short replicated arms and a large replicated

region will migrate very similarly to the unreplicated version of the same DNA fragment. Finally, the Y-shaped molecule that results from the almost completely replicated fragment is similar in shape to a linear molecule two times the size of the unreplicated fragment. Thus, as a DNA molecule is replicated by a single replication fork, it will migrate in positions that vary from a spot that is close to the unreplicated fragment in an arc that eventually reaches a location to which a linear molecule twice the size of the unreplicated DNA would be expected to migrate. This shape is called a **Y-arc** and indicates that a

**Box 9-4 (Continued)**

**BOX 9-4 FIGURE 3 Molecular identification of an origin of replication.** (a) By electrophoretically separating DNA in two dimensions, DNA in the process of replication can be separated from fully replicated or unreplicated DNA. Total DNA is isolated from dividing cells (and therefore replicating their DNA). The DNA is first separated primarily by size (using low-voltage electrophoresis through relatively large agarose pores), the electric field is rotated by 90°, and the DNA is then separated predominantly by size and shape (electrophoresed with high voltage in smaller-pore agarose). Southern blot analysis is used to detect the DNA of interest. The locations of the origin (green), restriction enzyme cleavage sites (red), and Southern blot DNA probe (blue) are illustrated for the three different patterns that can be observed. (b) The pattern of intermediates observed when the origin is located outside of the DNA fragment detected by Southern blotting. Called a Y-arc, all of the intermediates are Y-shaped. The molecule with three equal length arms (number 2) moves the slowest in the second dimension. (c) A bubble-arc is formed when the origin is near the middle of the DNA fragment detected. The intermediate with the largest bubble migrates the slowest in the second dimension. (d) This pattern arises when the origin is off-center in the restriction fragment. Initially bubble intermediates are detected but later (after one fork passes a restriction site) Y-intermediates are observed. This bubble-to-Y-arc pattern is considered the most indicative of the presence of an origin.

molecule is in the process of being replicated. Because all DNA molecules are replicated during each round of replication, the majority of DNA fragments will show this type of pattern.

Molecules that contain an origin of replication form bubble-shaped replication intermediates that migrate even more slowly in the first dimension than Y-shaped molecules. The larger the bubble, the more these molecules migrate differently from linear DNA (Box 9-4 Fig. 3c). Unfortunately, it is difficult to distinguish the arc of intermediates created by a bubble-containing fragment (called a **bubble arc**) from one created by Y-shaped

intermediates (Box 9-4 Fig. 3b,c). This difficulty can be overcome if the origin is located asymmetrically in the DNA fragment. In this instance, the intermediates will start out as bubbles, but when the replication fork closest to the end of the fragment completes replication, the bubble-shaped intermediates will become Y-shaped. This so-called **bubble-to-Y transition** is easily detected as a discontinuity in the arc and is highly indicative of an origin (Box 9-4 Fig. 3d). Thus, ideally, the restriction enzymes chosen will asymmetrically flank the origin of replication to be detected.

*S. cerevisiae* and bacterial chromosomes. Unlike their smaller counterparts, simple mutations that eliminate the function of these replicators have not been identified. Indeed, it is likely that specific DNA sequences do not play a major role in the definition of these replicators. Instead, recent studies suggest that reduced local nucleosome density and nearby transcription are important replicator determinants.

## BINDING AND UNWINDING: ORIGIN SELECTION AND ACTIVATION BY THE INITIATOR PROTEIN

Initiator proteins perform at least two different functions during the initiation of replication. First, these proteins bind to the replicator DNA, often via a specific binding site. Second, initiator proteins interact with additional factors required for replication initiation, thus recruiting them to the replicator. Some but not all initiator proteins perform a third function: They distort or unwind a region of DNA adjacent to their binding site to facilitate the initial opening of the DNA duplex.

Consider, for example, the *E. coli* initiator protein, DnaA. DnaA includes two DNA-binding domains. One domain binds the repeated 9-mer elements in *oriC* in their double-stranded form (see Fig. 9-26). When bound to ATP (but not ADP), DnaA also interacts with DNA in the region of the repeated 13-mer repeats of *oriC*. These additional interactions involve a distinct single-stranded DNA-binding site in DnaA and result in the separation of the DNA strands over more than 20 bp within the 13-mer repeat region. Intriguingly, once bound to DnaA, the single-stranded DNA is held in a conformation that prevents the formation of more than three continuous base pairs, ensuring that the DNA remains single-stranded. This unwound DNA provides an ssDNA template for additional replication proteins to begin the RNA and DNA synthesis steps of replication (see later discussion).

The formation of ssDNA at a site in the chromosome is not sufficient for the DNA helicase and other replication proteins to assemble. Rather, DnaA recruits additional replication proteins to the ssDNA formed at the replicator including the DNA helicase (see the next section). The regulation of *E. coli* replication is linked to the control of DnaA activity and is discussed later in Box 9-5, *E. coli* DNA Replication Is Regulated by DnaA·ATP Levels and SeqA.

In eukaryotic cells, the initiator is a six-protein complex called the **origin recognition complex** (ORC). The function of ORC is best understood in yeast cells. ORC recognizes a conserved sequence found in yeast replicators, called the “A element,” as well as a second, less-conserved B1 element (see Fig. 9-26). Like DnaA, ORC binds and hydrolyzes ATP. ATP binding is required for sequence-specific DNA binding at the origin, and ATP hydrolysis is required for ORC to participate in the loading of the eukaryotic DNA helicase onto the replicator DNA (see later discussion). Unlike DnaA, binding of ORC to yeast replicators does not lead to strand separation of the adjacent DNA. Nevertheless, ORC is required to recruit, either directly or indirectly, all of the remaining replication proteins to the replicator (see the section Helicase Loading Is the First Step in the Initiation of Replication in Eukaryotes).

### Protein–Protein and Protein–DNA Interactions Direct the Initiation Process

Once the initiator binds to the replicator, the remaining steps in the initiation of replication are largely driven by protein–protein interactions and protein–DNA interactions that are sequence-independent. The end result is the assembly of two replisomes that we described earlier. To explore the events that produce these protein machines, we first turn to *E. coli*, in which they are understood in the most detail.

After the initiator (DnaA) has bound to *oriC* and unwound the 13-mer DNA, the combination of ssDNA and DnaA recruits a complex of two proteins: the DNA helicase (DnaB) and helicase loader (DnaC) (Fig. 9-27a–d). Importantly, binding to the helicase loader inactivates the DNA helicase, preventing it from functioning at inappropriate sites. Once bound to the

## ► ADVANCED CONCEPTS

**Box 9-5** *E. coli* DNA Replication Is Regulated by DnaA-ATP Levels and SeqA

In all organisms, it is critical that replication initiation be tightly controlled to ensure that chromosome number and cell number remain appropriately balanced. Although this balance is most tightly regulated in eukaryotic cells, *E. coli* also prevents runaway chromosome duplication by inhibiting recently initiated origins from reinitiating. Several different mechanisms perform this function.

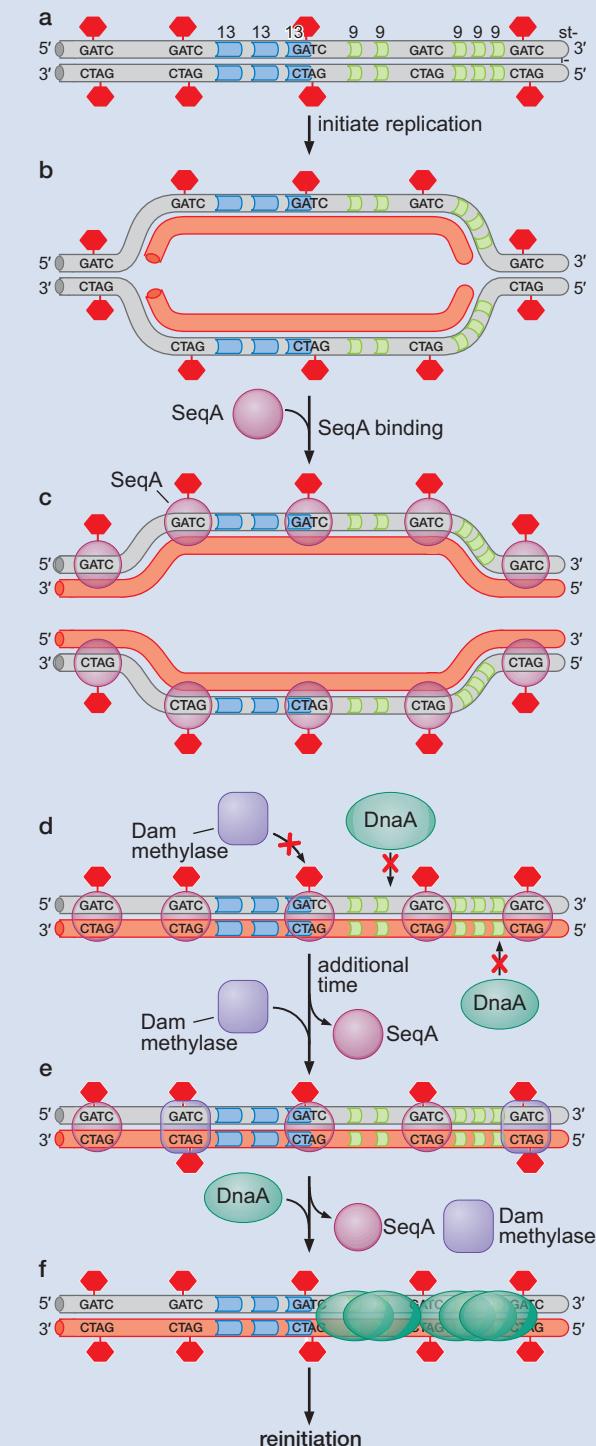
One method exploits changes in the methylated state of the DNA before and after DNA replication (Box 9-5 Fig. 1). In *E. coli* cells, an enzyme called Dam methyltransferase adds a methyl group to the A within every GATC sequence (note that the sequence is a palindrome). Typically, the genome is fully methylated at GATC sequences. This situation is changed after each GATC sequence is replicated. Because the A residues in the newly synthesized DNA strands are unmethylated, recently replicated sites will be methylated on only one strand (referred to as hemimethylated).

The hemimethylated state of newly replicated *oriC* DNA is detected by a protein called SeqA. SeqA binds tightly to the GATC sequence, but only when it is hemimethylated. There is an abundance of GATC sequences within and near *oriC*. Once replication has initiated, SeqA binds to these sites before they can become fully methylated by the Dam methyl transferase.

Binding of SeqA has two consequences. First, it dramatically reduces the rate at which the bound GATC sites are methylated. Second, when bound to these *oriC* proximal sites, SeqA prevents DnaA from binding to *oriC* and initiating a new round of replication. Thus, the conversion of the *oriC*-proximal GATC sites from methylated to hemimethylated (an event that is a direct consequence of initiation of replication from *oriC*) inhibits DnaA binding and, therefore, prevents rapid reinitiation of replication from the two newly synthesized daughter copies of *oriC*.

DnaA is targeted by other mechanisms that inhibit rapid reinitiation at newly synthesized copies of *oriC*. As described above, only DnaA bound to ATP can direct initiation of replication; however, this bound ATP is converted to ADP during the initia-

tion process. In addition, the sliding clamps that are loaded as a consequence of replication initiation recruit a protein (Hda) that stimulates ATP hydrolysis by DnaA. Thus, the process of directing a round of replication initiation inactivates DnaA, preventing its reuse. The process of exchanging the bound ADP for an ATP is

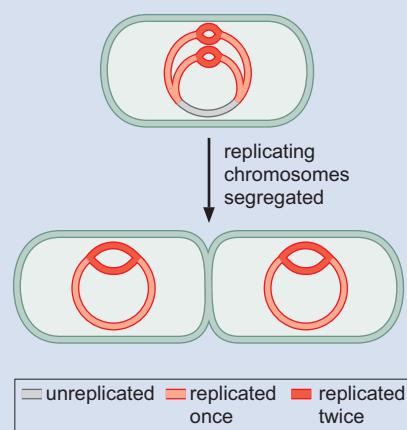


**BOX 9-5 FIGURE 1** SeqA bound to hemimethylated DNA inhibits reinitiation from recently replicated daughter origins. (a) Before DNA replication, GATC sequences throughout the *E. coli* genome are methylated on both strands ("fully" methylated). Note that throughout the figure, the methyl groups are represented by red hexagons. (b) DNA replication converts these sites to the hemimethylated state (only one strand of the DNA is methylated). (c) Hemimethylated GATC sequences are rapidly bound by SeqA. (d) Bound SeqA protein inhibits the full methylation of these sequences and the binding of *oriC* by DnaA protein (for simplicity, only one of the two daughter molecules is illustrated in parts d–f). (e) When SeqA infrequently dissociates from the GATC sites, the sequences can become fully methylated by Dam DNA methyltransferase, preventing rebinding by SeqA. (f) When the GATC sites become fully methylated, DnaA can bind the 9-mer sequences and direct a new round of replication from the daughter *oriC* replicators.

**Box 9-5 (Continued)**

a slow one, further delaying the accumulation of replication-competent ATP-bound DnaA. The process of replicating nearby sequences also acts to reduce the amount of DnaA available to bind at *oriC*. There are more than 300 DnaA 9-mer binding sites outside of *oriC* (DnaA also acts as a transcriptional regulator at several promoters), and as they are replicated, this number doubles. The increase in DnaA-binding sites acts to reduce the level of available DnaA.

Together, these methods rapidly and dramatically reduce the ability of *E. coli* to initiate replication from new copies of *oriC*. Although these mechanisms prevent rapid reinitiation, this inhibition does not necessarily last until cell division is complete. Indeed, for *E. coli* cells to divide at the maximum rate, the daughter copies of *oriC* must initiate replication before the completion of the previous round of replication. This is because *E. coli* cells can divide every 20 min, but it takes more than 40 min to replicate the *E. coli* genome. Thus, under rapid growth conditions, *E. coli* cells reinitiate replication once and sometimes twice before the completion of previous rounds of replication (Box 9-5 Fig. 2). Even under such rapid growth conditions, initiation does not occur more than once per round of cell division. Thus, for each round of cell division, there is only one round of replication initiation from *oriC*.

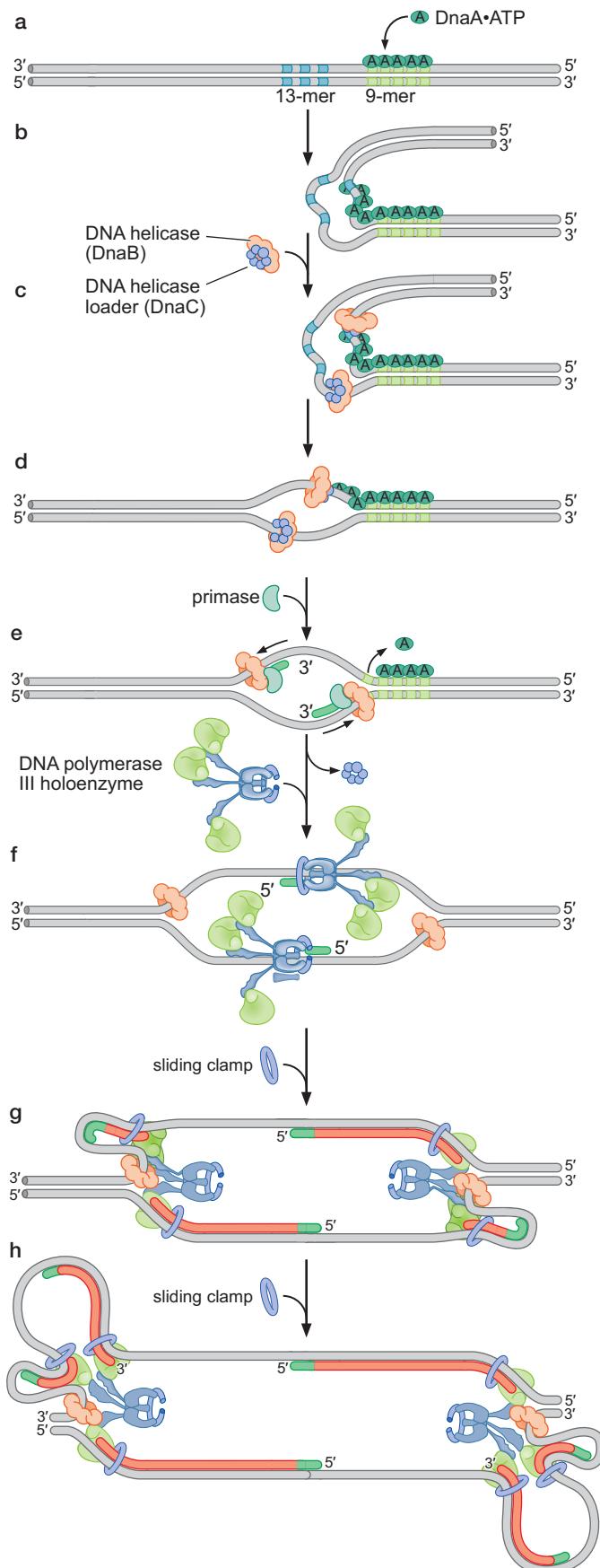


**BOX 9-5 FIGURE 2** Origins of replication reinitiate replication before cell division in rapidly growing cells. To allow the genome to be fully replicated before each round of cell division, bacterial cells frequently have to initiate DNA replication from their single origin before the completion of cell division. This means that the chromosomes that are segregated into the daughter cells are being actively replicated. This is in contrast to eukaryotic cells, which do not start chromosome segregation until all of the chromosomes are completely replicated.

ssDNA at the origin, the helicase loader directs the assembly of its associated DNA helicase around the ssDNA (recall that ssDNA passes through the middle of the DnaB helicase's hexameric protein ring). Although the mechanism of loading is not understood in detail, the process is analogous to the assembly of sliding DNA clamps around a primer: template junction, requiring the opening of the DNA helicase hexameric ring to allow it to encircle the targeted ssDNA (see Box 9-3). Interestingly, like the subunits of the sliding clamp loader, DnaC is a ATP-utilizing AAA+ protein.

The protein–protein interactions between the helicase and other components of the replication fork described above direct the assembly of the rest of the replication machinery (see Fig. 9-27e,f). Helicase recruits DNA primase to the origin DNA, resulting in the synthesis of an RNA primer on each strand of the origin. In addition to generating the primers for the leading DNA strands, this event also causes the release of the helicase loader and, therefore, the activation of the helicase. The DNA Pol III holoenzyme is brought to the origins through interactions with the primer:template junction and the helicase. Once the holoenzyme is present, sliding clamps are assembled on the RNA primers, and the leading-strand polymerases are engaged. As new ssDNA is exposed by the action of the helicase, it is bound by SSBs, and DNA primase synthesizes the first lagging-strand primers. These new primer:template junctions are targeted by the clamp loaders at each fork, which place two additional sliding clamps on the lagging strands. These clamps are recognized by a second core DNA Pol III enzyme, resulting in the initiation of lagging-strand DNA synthesis. As the first Okazaki fragments are extended and more ssDNA lagging-strand DNA template is generated, a new RNA primer is synthesized. After a sliding DNA clamp is assembled, the second Okazaki fragment is initiated by the third DNA Pol III enzyme (see Fig. 9-27g,h). At this point, two replication forks have been assembled, and initiation of replication is complete.

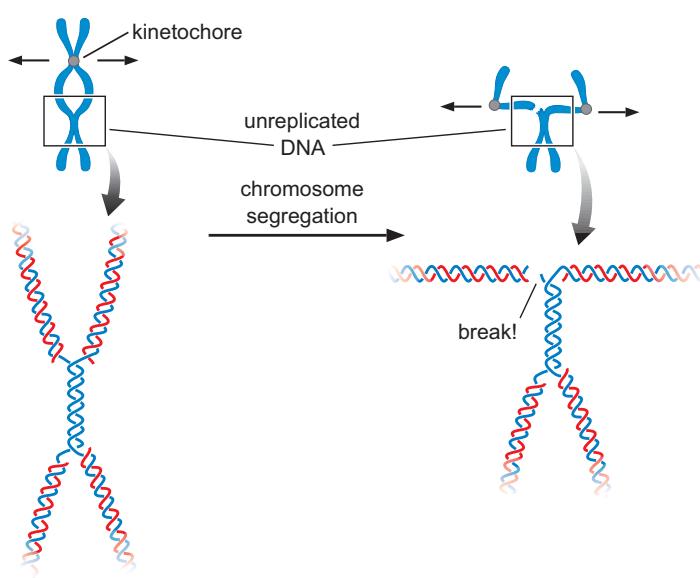
**FIGURE 9-27** Model for *E. coli* initiation of DNA replication. The major events in the initiation of *E. coli* DNA replication are illustrated. (a) Multiple DnaA-ATP proteins bind to the repeated 9-mer sequences within *oriC*. (b) Binding of DnaA-ATP to these sequences leads to strand separation within the 13-mer repeats. This is mediated by an ssDNA-binding domain in DnaA-ATP that elongates and changes the structure of the associated ssDNA such that it cannot hybridize to the complementary ssDNA. (c) A complex between DNA helicase (DnaB) and the DNA helicase loader (DnaC) associates with the DnaA-bound origin. An ssDNA-binding domain in the helicase loader and protein–protein interactions between DnaA and the helicase/helicase loader mediate these interactions. (d) The DNA helicase loader catalyzes the opening of the DNA helicase protein ring and placement of the ring around the ssDNA at the origin. (e) The DNA helicases each recruit a primase that synthesizes an RNA primer on each template. The RNA primer causes the helicase loader to release from the helicase, resulting in the activation of the DNA helicase. The movement of the DNA helicases also removes any remaining DnaA bound to the replicator. (f) The newly synthesized primers and the helicases are recognized by the clamp loader components of DNA Pol III holoenzymes. Sliding clamps are assembled on each RNA primer, and leading-strand synthesis is initiated by one of the three core DNA Pol III enzymes of each holoenzyme. (g) After each DNA helicase has moved ~1000 bases, a second RNA primer is synthesized on each lagging-strand template, and a sliding clamp is loaded. The resulting primer:template junction is recognized by a second DNA Pol III core enzyme in each holoenzyme, resulting in the initiation of lagging-strand synthesis. (h) Leading-strand synthesis and lagging-strand synthesis are now initiated at each replication fork. As shown in Figure 9-23, the third DNA Pol III core enzyme also participates in lagging-strand DNA synthesis. Each replication fork will continue to the end of the template or until it meets another replication fork moving in the opposite direction.



## Eukaryotic Chromosomes Are Replicated Exactly Once per Cell Cycle

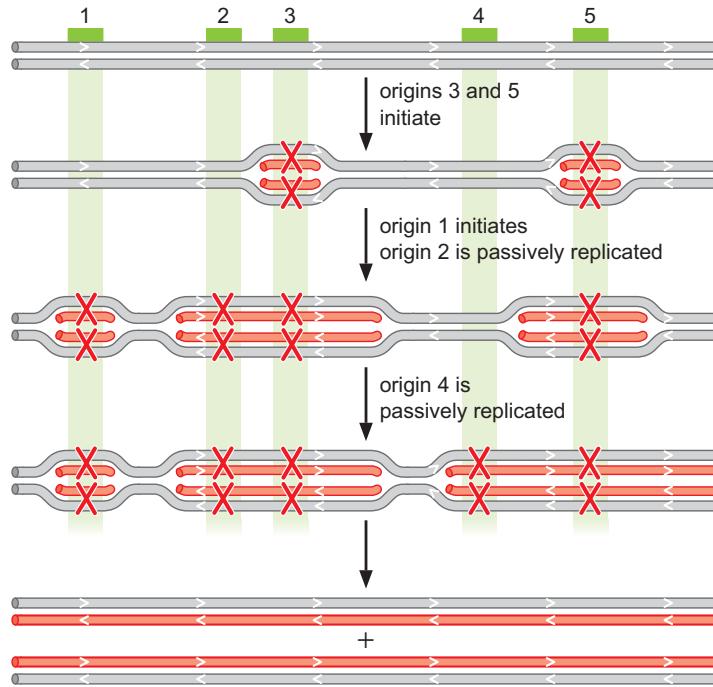
As discussed in Chapter 7, the events required for eukaryotic cell division occur at distinct times during the cell cycle. Chromosomal DNA replication occurs only during the S phase of the cell cycle. During this time, all of the DNA in the cell must be duplicated exactly once. Incomplete replication of any part of a chromosome causes inappropriate links between daughter chromosomes. Segregation of linked chromosomes causes chromosome breakage or loss (Fig. 9-28). Re-replication of even limited amounts of eukaryotic DNA leads to DNA lesions that are difficult for the cell to repair. Attempts to repair such lesions frequently result in amplification of the associated DNA, which can inappropriately increase the expression of the associated genes. Addition of even one or two more copies of critical regulatory genes can lead to catastrophic defects in gene expression, cell division, or the response to environmental signals. Thus, it is critical that every base pair in each chromosome be replicated *once and only once* each time a eukaryotic cell divides.

The need to replicate the DNA once and only once is a particular challenge for eukaryotic chromosomes because they each have many origins of replication. Origins are typically separated by ~30 kb, thus even a small eukaryotic chromosome may have more than 10 origins and a large human chromosome may have thousands. Enough of these origins must be activated to ensure that each chromosome is fully replicated during each S phase. Typically, not all replicators need to be activated to complete replication, but if too few are activated, regions of the genome will escape replication. On the other hand, although some potential origins may not be used in any given round of cell division, *no* replicator can initiate after it has been replicated. Thus, whether a replicator is activated to cause its own replication or replicated by a replication fork derived from an adjacent replicator, it *must be inactivated* until the next round of cell division (Fig. 9-29). If these conditions were not true, the DNA associated with an origin could be replicated twice in the same cell cycle, breaking the “once and only once” rule of eukaryotic DNA replication.



**FIGURE 9-28** Chromosome breakage as a result of incomplete DNA replication. This illustration shows the consequences of incomplete replication followed by chromosome segregation. The top of each illustration shows the entire chromosome. The bottom shows the details of the chromosome breakage at the DNA level. (For details of chromosome segregation, see Chapter 7.) As the chromosomes are pulled apart, stress is placed on the unreplicated DNA, resulting in the breakage of the chromosome.

**FIGURE 9-29** Replicators are inactivated by DNA replication. A eukaryotic chromosome with five replicators is shown. The replicators labeled 3 and 5 are the first to be activated, leading to the formation of two pairs of bidirectional replication forks. Activation of the parental replicator results in the inactivation of the copies of the replicator on both daughter DNA molecules until the next cell cycle (indicated by a red X). Elongation of the resulting replication forks replicates the DNA overlapping with the number 2 and 4 replicators before they initiate replication. When a replicator is copied by a fork derived from an adjacent origin before initiation, it is said to have been passively replicated. Although these replicators have not initiated, they are nevertheless inactivated by the act of replicating their DNA (as we discuss later, this is because helicases loaded at the origin are removed by the passing replication fork). In contrast, replicator 1 is not reached by an adjacent fork before initiation and is able to initiate normally. The presence of more replicators than needed to complete DNA replication is a form of redundancy to ensure the complete replication of each chromosome.

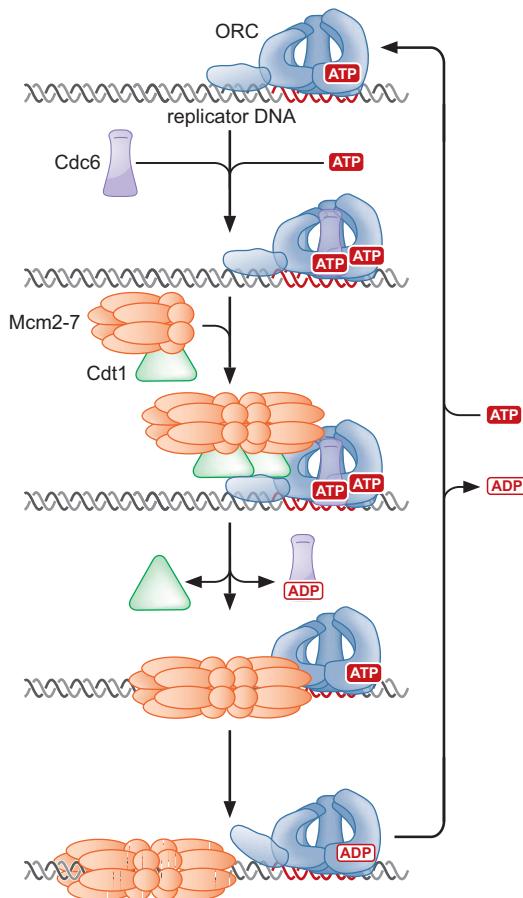


### Helicase Loading Is the First Step in the Initiation of Replication in Eukaryotes

The events of eukaryotic replication initiation occur at distinct times in the cell cycle (see Chapter 8). Helicase loading occurs at all replicators during G<sub>1</sub> (before S phase). Replicator or origin activation, including helicase activation and replisome assembly, only occurs after cells enter S phase.

The separation of helicase loading and origin activation is different from the situation in prokaryotic cells, where binding of the initiator to the replicator DNA directly leads to DNA unwinding, helicase loading, and replisome assembly. As we see later, the temporal separation of helicase loading from helicase activation and replisome assembly during the eukaryotic cell cycle ensures that each chromosome is replicated only once during each cell cycle (bacterial cells solve this problem differently; see Box 9-5).

Eukaryotic helicase loading requires four separate proteins to act at each replicator (Fig. 9-30). The first step in helicase loading is the recognition of the replicator by the eukaryotic initiator, ORC, bound to ATP. As cells enter the G<sub>1</sub> phase of the cell cycle, ORC bound to the origin recruits two helicase loading proteins (Cdc6 and Cdt1) and two copies of the Mcm2-7 helicase to the origin. Interestingly, several ORC subunits and the Cdc6 protein are members of the AAA+ family of proteins like DnaC and the subunits of the sliding clamp loader. Like the sliding clamp loader, ATP binding by ORC and Cdc6 is required for ORC DNA binding and the stable recruitment of the helicase and helicase loading proteins. ATP hydrolysis by Cdc6 results in the loading of a head-to-head dimer of the Mcm2-7 complex such that they encircle the double-stranded origin DNA. During this event, Cdt1 and Cdc6 are released from the origin. ORC ATP hydrolysis is thought to reset the process and allow a new round of Mcm2-7 loading to be initiated upon ATP binding to the ORC. Consistent with the Mcm2-7 complex encircling dsDNA instead of ssDNA, eukaryotic helicase loading does not lead to the immediate unwinding of origin DNA. Instead, helicases that are loaded



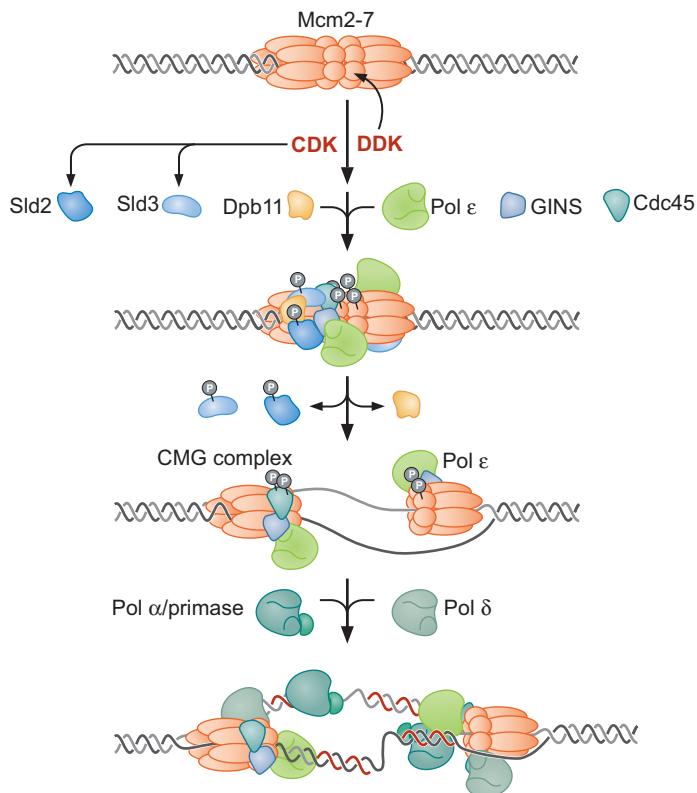
**FIGURE 9-30 Eukaryotic helicase loading.** Loading of the eukaryotic replicative DNA helicase is an ordered process that is initiated by the association of the ATP-bound origin recognition complex (ORC) with the replicator. Once bound to the replicator, ORC recruits ATP-bound Cdc6 and two copies of the Mcm2-7 helicase bound to a second helicase loading protein, Cdt1. This assembly of proteins triggers ATP hydrolysis by Cdc6, resulting in the loading of a head-to-head dimer of the Mcm2-7 complex encircling double-stranded origin DNA and the release of Cdc6 and Cdt1 from the origin. Subsequent ATP hydrolysis by ORC is required to reset the process (illustrated as release from Mcm2-7). Exchange of ATP for ADP allows a new round of helicase loading.

during G<sub>1</sub> are only activated to unwind DNA and initiate replication after cells pass from the G<sub>1</sub> to the S phase of the cell cycle.

Loaded helicases are activated by two protein kinases: CDK (cyclin-dependent kinase) and DDK (Dbf4-dependent kinase) (Fig. 9-31). **Protein kinases** are proteins that covalently attach phosphate groups to target proteins (see Chapter 13). These kinases are activated when cells enter S phase. Once activated, DDK targets the loaded helicase, and CDK targets two other replication proteins. Phosphorylation of these proteins results in the Cdc45 and GINS proteins binding to the Mcm2-7 helicase (see Fig. 9-31). Importantly, Cdc45 and GINS strongly stimulate the Mcm2-7 ATPase and helicase activities and together form the Cdc45–Mcm2-7–GINS (CMG) complex, which is the active form of the Mcm2-7 DNA helicase. Although the helicase is initially loaded around dsDNA as a head-to-head dimer, at the replication fork it is thought to act as a single Mcm2-7 hexamer encircling ssDNA. Thus, during the activation events, one strand of DNA must be ejected from the central channel of each helicase, and the interactions between the two Mcm2-7 complexes must be disrupted (Fig. 9-32). The three eukaryotic DNA polymerases assemble at the origin in a defined order. DNA Pol ε associates with the origin at the same time as Cdc45 and the GINS before DNA unwinding. In contrast, DNA Pol δ and DNA Pol α/primase both require DNA unwinding before their recruitment to the origin. This order ensures that all three DNA polymerases are present at the origin before the synthesis of the first RNA primer (by DNA Pol α/primase).

Only a subset of the proteins that assemble at the origin goes on to function as part of the eukaryotic replisome. The CMG complex and the three DNA polymerases become part of the replication fork machinery. Similar

**FIGURE 9-31** Activation of loaded helicases leads to the assembly of the eukaryotic replisome. As cells enter into the S phase of the cell cycle, two kinases, CDK and DDK, are activated. DDK phosphorylates loaded Mcm2-7 helicase, and CDK phosphorylates Sld2 and Sld3. Phosphorylated Sld2 and Sld3 bind to Dpb11, and together these proteins facilitate binding of the helicase-activating proteins, Cdc45 and GINS, to the helicase. Cdc45 and GINS form a stable complex with the Mcm2-7 helicase (called the Cdc45/Mcm2-7/GINS, or CMG, complex) and dramatically activate Mcm2-7 helicase activity. The leading-strand DNA polymerase ( $\epsilon$ ) is recruited to the helicase at this stage (before DNA unwinding). After formation of the CMG complex, Sld2, Sld3, and Dpb11 are released from the origin. DNA Pol  $\alpha$ /primase and DNA Pol  $\delta$  (which primarily act on the lagging strand) are only recruited after DNA unwinding. The protein–protein interactions that hold the DNA polymerase at the replication fork remain poorly understood.

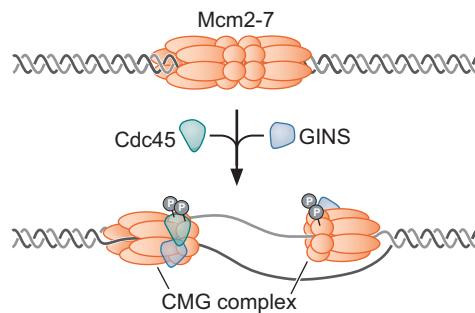


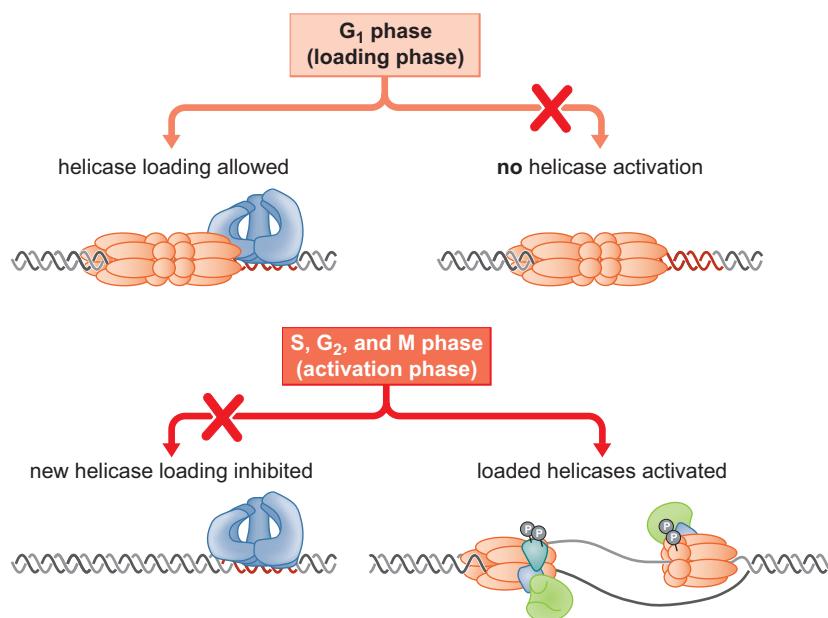
to the *E. coli* DNA helicase loader (DnaC), other factors are only required to assemble the replication fork proteins (such as Cdc6 and Cdt1) and are released or destroyed after their role is complete (see Fig. 9-31).

### Helicase Loading and Activation Are Regulated to Allow Only a Single Round of Replication during Each Cell Cycle

How do eukaryotic cells control the activity of hundreds or even thousands of origins of replication such that *not even one* is activated more than once during a cell cycle? The answer lies in the oscillation between two replication states that occurs once per cell cycle. During G<sub>1</sub>, cells are in the helicase loading phase and are competent for helicase loading but unable to activate the loaded helicases. Upon entry into S phase and continuing throughout G<sub>2</sub> and M phase, helicases loaded during G<sub>1</sub> can be activated, but new helicase loading is strictly inhibited (Fig. 9-33). Importantly, the conditions for helicase loading and activation are incompatible with one another. Although the exact mechanisms vary between different organisms, this same regulation is seen in all actively dividing eukaryotic cells. Thus, during each

**FIGURE 9-32** Helicase activation alters helicase interactions. Before helicase activation, loaded helicases encircle double-stranded DNA and are in the form of a head-to-head double hexamer (mediated by interactions between the Mcm2-7 amino termini). After helicase activation, the Mcm2-7 protein in the CMG complex is proposed to encircle single-stranded DNA, and the interaction between the two Mcm2-7 complexes has been broken.





**FIGURE 9-33** Eukaryotic helicase loading and activation occur during different cell cycle stages. During the G<sub>1</sub> phase of the cell cycle, helicase loading is permitted, but helicase activation is not allowed. During the remainder of the cell cycle (S, G<sub>2</sub>, and M phases), helicase loading is inhibited, but loaded helicases can be activated (this will only occur during S phase because after S phase all loaded Mcm2-7 complexes will be removed from the DNA; see Fig. 9-29).

cell cycle, there is *only one* opportunity for helicases to load onto origins (during G<sub>1</sub>) and *only one* opportunity for those loaded helicases to be activated (during S, G<sub>2</sub>, and M—although in practice, all loaded helicases are activated or disrupted by replication forks during S phase). Only after the cells segregate their replicated chromosomes and divide are they able to re-enter G<sub>1</sub> and load a new set of helicases at their origins.

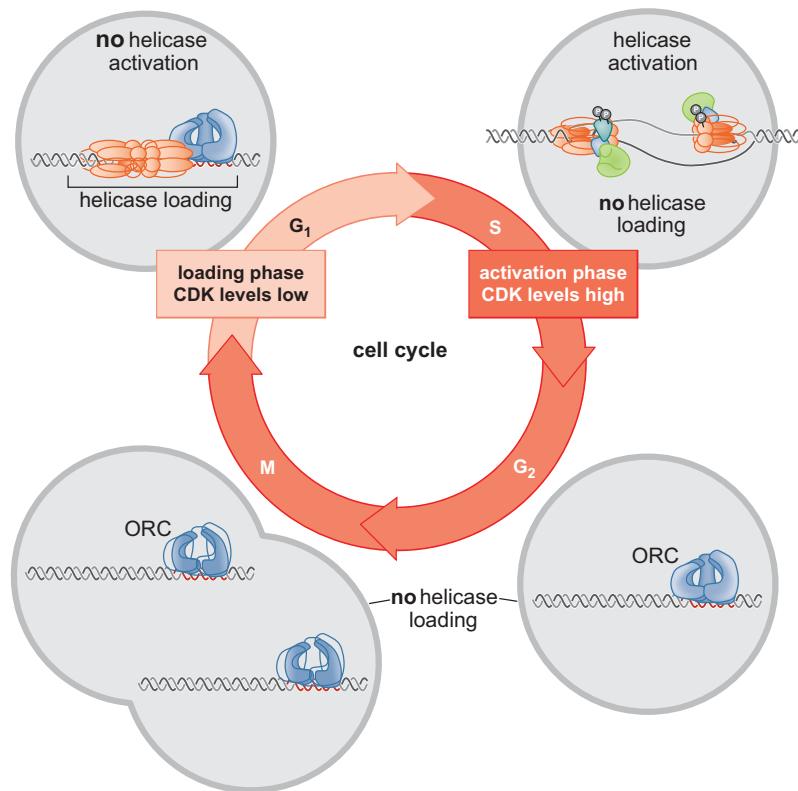
How is this regulation achieved? In the budding yeast *S. cerevisiae*, the regulation is tightly coupled to the function of CDKs (Fig. 9-34). These enzymes play seemingly contradictory roles in regulating replication. First, as described above, they are required to activate loaded helicases to initiate DNA replication. Second, CDK activity *inhibits* helicase loading. When considered in the light of the regulation described above, these different roles allow one enzyme to control the oscillation between the two states of replication initiation. CDK levels are low during G<sub>1</sub>, allowing helicase loading but preventing helicase activation. Entry into the S phase of the cell cycle is coupled with a rapid increase in CDK activity, driving activation of loaded helicases but simultaneously preventing new helicase loading. Importantly, CDK levels remain elevated during the remainder of the cell cycle (S, G<sub>2</sub>, and M phases).

Loaded helicases are released from the DNA after the replication fork they are part of completes DNA synthesis or after the DNA to which they are bound is replicated (by a replication fork derived from an adjacent origin; see Fig. 9-29). These exposed replicators are potentially available for new helicase loading and rapidly bind to ORC. Despite the presence of the initiator at these sites, however, the high levels of CDK activity present during the S, G<sub>2</sub>, and M phases inhibit the function of ORC, Cdc6, and Cdt1. It is only when cells segregate their chromosomes and complete cell division that CDK activity is eliminated, allowing a new round of helicase loading to commence.

### Similarities between Eukaryotic and Prokaryotic DNA Replication Initiation

Now that we have described initiation in eukaryotes and prokaryotes, it is clear that the general principles of replication initiation are the same in

**FIGURE 9-34** Cell cycle regulation of CDK activity controls replication. In *S. cerevisiae* cells, CDK levels tightly regulate helicase loading and activation. During G<sub>1</sub>, CDK levels are low, allowing helicases to be loaded, but the loaded helicases cannot be activated (because of the requirement of CDK for this event). During S phase, elevated CDK activity inhibits new helicase loading and activates previously loaded helicases. When a loaded helicase is used for the initiation of replication, it is incorporated into the replication fork and leaves the origin. Similarly, passive replication of origin DNA also removes the helicase from the origin DNA (not shown). Because CDK levels remain high until the end of mitosis, no new helicase loading can occur until chromosome segregation is complete and the daughter cells have returned to G<sub>1</sub>. Without a new round of helicase loading, reinitiation is impossible.



both cases. The first step is the recognition of the replicator by the initiator protein. The initiator protein in combination with one or more helicase loading proteins assembles the DNA helicase on the replicator. The helicase (and potentially other proteins at the origin in eukaryotes) generates a region of ssDNA that can act as a template for RNA primer synthesis. Once primers are synthesized, the remaining components of the replisome assemble through interactions with the resulting primer:template junction.

Although the events of initiation are similar, the regulation of replication in bacteria and eukaryotic cells is distinctly different. For example, unlike eukaryotic cells, rapidly dividing bacterial cells initiate replication more than once per cell cycle. The step that is most tightly regulated is also different. Eukaryotic cells focus regulation on the initial loading of the MCM helicase onto the origin DNA, whereas bacterial cells focus regulation on the binding of the DnaA initiator protein to the DNA (Box 9-5, *E. coli* DNA Replication Is Regulated by DnaA·ATP Levels and SeqA).

## FINISHING REPLICATION

Completion of DNA replication requires a set of specific events. These events are different for circular versus linear chromosomes. For a circular chromosome, the conventional replication fork machinery replicates the entire molecule, but the resulting daughter molecules are topologically linked to each other. In contrast, the replication fork machinery we have discussed so far cannot complete replication of the very ends of linear chromosomes. Therefore, organisms containing linear chromosomes have developed novel strategies to replicate their chromosome ends.

## Type II Topoisomerases Are Required to Separate Daughter DNA Molecules

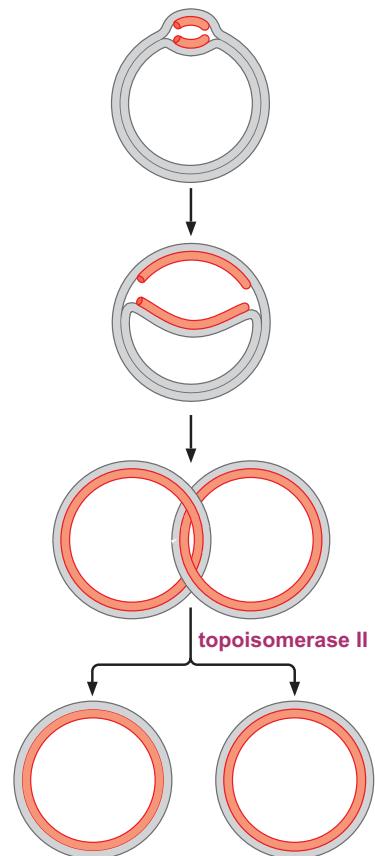
After replication of a circular chromosome is complete, the resulting daughter DNA molecules remain linked together as catenanes (Fig. 9-35; Chapter 4, Fig. 4-23). *Catenane* is the general term for two circles that are linked (similar to links in a chain). To segregate these chromosomes into separate daughter cells, the two circular DNA molecules must be disengaged from each other or “decatenated.” This separation is accomplished by the action of **type II topoisomerases**. As discussed in Chapter 4, these enzymes have the ability to break a dsDNA molecule and pass a second dsDNA molecule through this break. This reaction can easily decatenate the two circular daughter chromosomes by breaking one DNA circle and passing the second through the break, allowing their segregation into separate cells.

Although the importance of this activity for the separation of circular chromosomes is most clear, the activity of type II topoisomerases is also critical to the segregation of large linear molecules. Although there is no inherent topological linkage after the replication of a linear molecule, the large size of eukaryotic chromosomes necessitates the intricate folding of the DNA into loops attached to a protein scaffold (see Chapter 8, Fig. 8-32b). These attachments lead to many of the same problems that circular chromosomes have after replication when the two daughter linear chromosomes must be separated. As in the case of circular chromosomes, type II topoisomerases allow these linked DNAs to be separated.

## Lagging-Strand Synthesis Is Unable to Copy the Extreme Ends of Linear Chromosomes

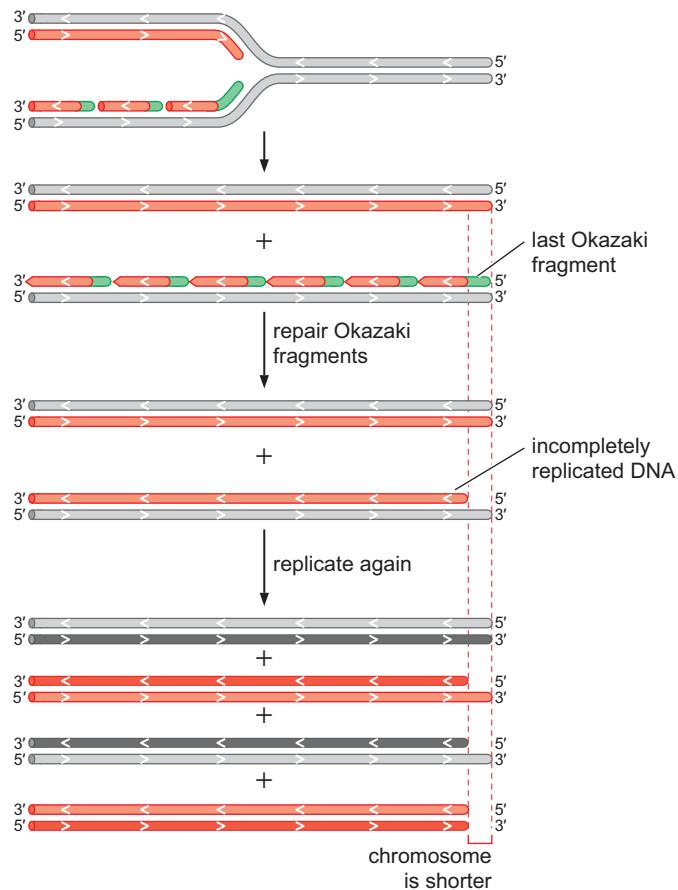
The requirement for an RNA primer to initiate all new DNA synthesis creates a dilemma for the replication of the ends of linear chromosomes, called the **end replication problem** (Fig. 9-36). This difficulty is not observed during the duplication of the leading-strand template. In that case, a single internal RNA primer can direct the initiation of a DNA strand that can be extended to the extreme 5' terminus of its template. In contrast, the requirement for multiple primers to complete lagging-strand synthesis means that a complete copy of its template cannot be made. Even if the end of the last RNA primer for Okazaki fragment synthesis anneals to the final base of the lagging-strand template, once this RNA molecule is removed, there will remain a short region (the size of the RNA primer) of unreplicated ssDNA at the end of the chromosome. Although this shortening would only occur on one of the two strands of the daughter molecule, after the next round of replication occurs both strands of the daughter molecule would be shorter. This means that each round of DNA replication would result in the shortening of one of the two daughter DNA molecules. Obviously, this scenario would disrupt the complete propagation of the genetic material from generation to generation. Slowly, but surely, genes at the end of the chromosomes would be lost.

Organisms solve the end replication problem in a variety of ways. One solution is to use a protein instead of an RNA as the primer for the last Okazaki fragment at each end of the chromosome (Fig. 9-37). In this situation, the “priming protein” binds to the lagging-strand template and uses an amino acid to provide an OH (typically a tyrosine) that replaces the 3'-OH normally provided by an RNA primer. By priming the last lagging strand, the priming protein becomes covalently linked to the 5' end of the chromosome. Terminaly attached replication proteins of this kind are found at the end of the linear chromosomes of certain species of bacteria (most bacteria have circular



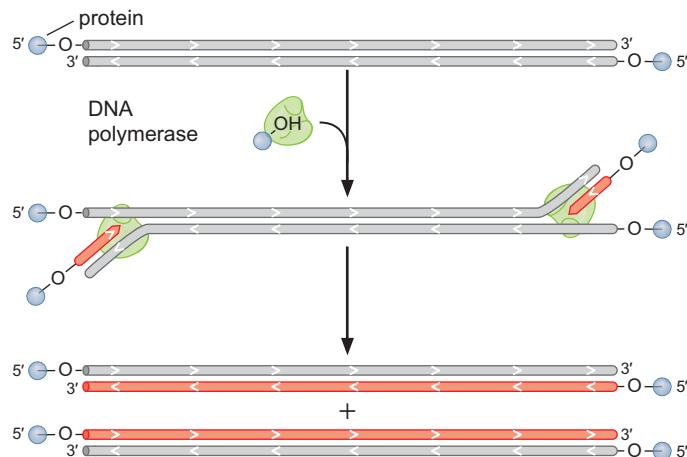
**FIGURE 9-35** Topoisomerase II catalyzes the decatenation of replication products. After a circular DNA molecule is replicated, the resulting complete daughter DNA molecules remain linked to each other. Type II DNA topoisomerases can efficiently separate (or decatenate) these DNA circles.

**FIGURE 9-36** The end replication problem. As the lagging-strand replication machinery reaches the end of the chromosome, at some point, primase no longer has sufficient space to synthesize a new RNA primer. This results in incomplete replication and a short ssDNA region at the 3' end of the lagging-strand DNA product. When this DNA product is replicated in the next round, one of the two products will be shortened and will lack the region that was not fully copied in the previous round of replication.



chromosomes) and at the ends of the linear chromosomes of certain bacterial and animal viruses.

But most eukaryotic cells use an entirely different solution to replicate their chromosome ends. As we learned in Chapter 8, the ends of eukaryotic chromosomes are called **telomeres**, and they are generally composed of head-to-tail repeats of a TG-rich DNA sequence. For example, human telomeres consist of many head-to-tail repeats of the sequence 5'-TTAGGG-3'. Although many of these repeats are double-stranded, the 3' end of each chromosome extends beyond the 5' end as ssDNA. This unique structure acts as a novel origin of replication that compensates for the end replication



**FIGURE 9-37** Protein priming as a solution to the end replication problem. By binding to the DNA polymerase and to the 3' end of the template, a protein provides the priming hydroxyl group to initiate DNA synthesis. In the example shown, the protein primes all DNA synthesis as is seen for many viruses. For longer DNA molecules, this method combines with conventional origin function to replicate the chromosomes.

problem. This origin does not interact with the same proteins as other eukaryotic origins, but it instead recruits a specialized DNA polymerase called **telomerase**.

### Telomerase Is a Novel DNA Polymerase That Does Not Require an Exogenous Template

**Telomerase** is a remarkable enzyme that includes multiple protein subunits and an RNA component (and is therefore an example of a ribonucleoprotein; see Chapter 5). Like all other DNA polymerases, telomerase acts to extend the 3' end of its DNA substrate. But unlike most DNA polymerases, telomerase does not need an exogenous DNA template to direct the addition of new dNTPs. Instead, the RNA component of telomerase serves as the template for adding the telomeric sequence to the 3' terminus at the end of the chromosome (see Interactive Animation 9-3). Telomerase specifically elongates the 3'-OH of telomeric ssDNA sequences using its own RNA as a template. As a result of this unusual mechanism, the newly synthesized DNA is single-stranded.

The key to telomerase's unusual functions is revealed by the RNA component of the enzyme, called "telomerase RNA" (TER). Depending on the organism, TER varies in size from 150 to 1300 bases. In all organisms, the sequence of the RNA includes a short region that encodes about 1.5 copies of the complement of the telomere sequence (for humans, this sequence is 5'-AAUCCCAAUC-3'). This region of the RNA can anneal to the ssDNA at the 3' end of the telomere (Fig. 9-38). Annealing occurs in such a way that a part of the RNA template remains single-stranded, creating a primer:template junction that can be acted on by telomerase. Interestingly, one of the protein subunits of telomerase is a member of a class of DNA polymerases that use RNA templates called "reverse transcriptases" (this subunit is called "telomerase reverse transcriptase," or TERT). As we shall see in Chapter 12, these enzymes "reverse-transcribe" RNA into DNA instead of the more conventional transcription of DNA into RNA. Using the associated RNA template, TERT synthesizes DNA to the end of the TER template region but cannot continue to copy the RNA beyond that point. At this point, the RNA template disengages from the DNA product, reanneals to the last four nucleotides of the telomere, and then repeats this process.

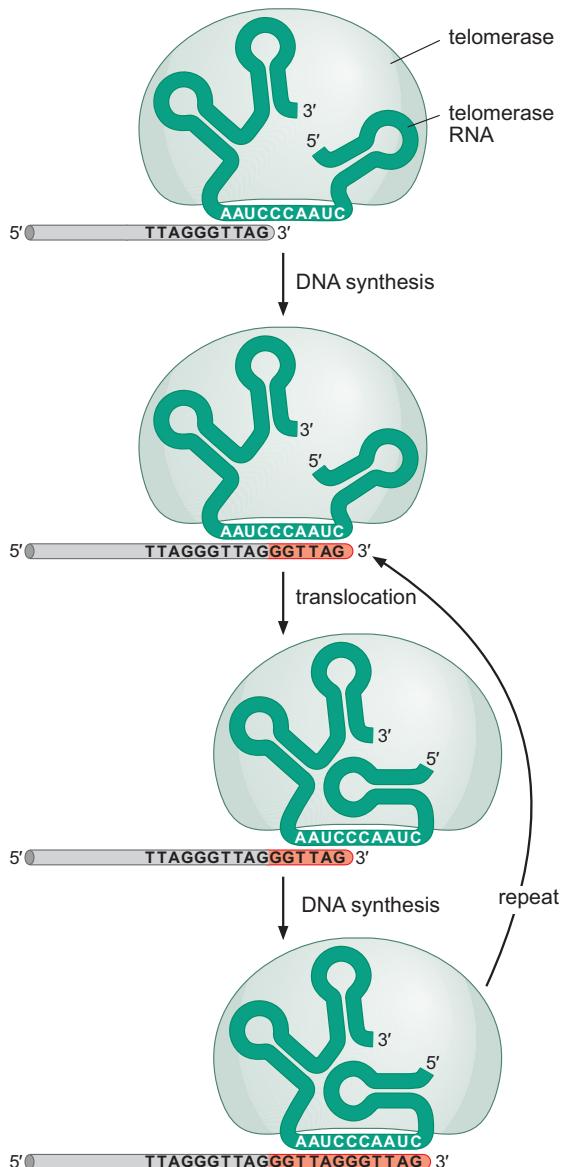
The characteristics of telomerase are in some ways distinct and in other ways similar to those of other DNA polymerases. The inclusion of an RNA component, the lack of a requirement for an exogenous template, and the ability to use an entirely ssDNA substrate to produce an ssDNA product sets telomerase apart from other DNA polymerases. In addition, telomerase must have the ability to displace its RNA template from the DNA product to allow repeated rounds of template-directed synthesis. Formally, this means that telomerase includes an RNA-DNA helicase activity. On the other hand, like all other DNA polymerases, telomerase requires a template to direct nucleotide addition, can only extend a 3'-OH end of DNA, uses the same nucleotide precursors, and acts in a processive manner, adding many sequence repeats each time it binds to a DNA substrate. Intriguing implications of the role of telomerase in regulating cell growth and cellular aging are discussed in Box 9-6, Aging, Cancer, and the Telomere Hypothesis.



### Telomerase Solves the End Replication Problem by Extending the 3' End of the Chromosome

When telomerase acts on the 3' end of the telomere, it extends only one of the two strands of the chromosome. How is the 5' end extended? This is

**FIGURE 9-38** Replication of telomeres by telomerase. Telomerase uses its RNA component to anneal to the 3' end of the ssDNA region of the telomere. Telomerase uses its reverse transcription activity to synthesize DNA to the end of the RNA template. Telomerase then displaces the RNA from the DNA product and rebinds at the end of the telomere and repeats the process.



accomplished by the lagging-strand DNA replication machinery (Fig. 9-39). By providing an extended 3' end, telomerase provides additional template for the lagging-strand replication machinery. By synthesizing and extending RNA primers using the telomerase extended 3' end as a template, the cell can effectively increase the length of the 5' end of the chromosome as well.

Even after the action of the lagging-strand machinery, there remains a short ssDNA region at the end of the chromosome. Indeed, the presence of a 3' overhang may be important for the end protection function of the telomere (as we discuss later). Nevertheless, the action of telomerase and the lagging-strand replication machinery ensures that the telomere is maintained at sufficient length to protect the end of the chromosome from shortening. Because of the repetitive and non-protein-coding nature of the telomeric DNA, variations in the length of the telomere are easily tolerated by the cell.

## MEDICAL CONNECTIONS

### Box 9-6 Aging, Cancer, and the Telomere Hypothesis

All organisms are mortal. Whether it is the days or weeks lived by many smaller organisms or the many years that the average human lives, organisms cannot escape their intrinsic mortality. Not surprisingly, researchers (and others) have long studied these limitations, hoping to understand them and, perhaps, overcome them and find the mythical “fountain of youth.”

When researchers developed ways to grow individual cells outside of the body, they thought that the cells were immortal. This suggested that mortality was a problem of whole organisms, not of cells. This hypothesis was eliminated when Leonard Hayflick studied cell division in culture more carefully. He found that, even in isolation, cells could divide only a limited number of times. Interestingly, Hayflick's studies found that the number of divisions a cell can pass through is characteristic of the source of cells, now known as the “Hayflick limit.”

Hayflick's studies led to the idea that cells contain an intrinsic countdown clock that limits the number of divisions in which a cell can participate. When the clock reaches zero, a cell would be prevented from dividing again. For years the molecular identity of such a clock was unknown; however, as the nature of telomeres and their role in DNA replication were better understood, it became clear that the telomere could be the long-sought-after divisional clock. Consistent with this idea, telomere DNA isolated from young people is longer than that isolated from older

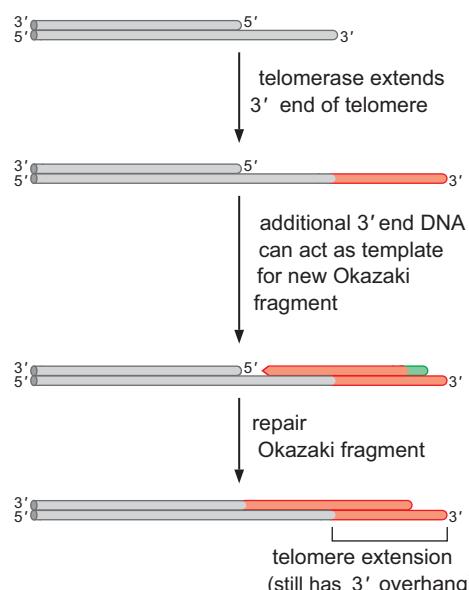
people. This observation led to the hypothesis that the length of telomeric DNA limited the number of times a cell could divide.

Although the concept is still very much a hypothesis, experimental support for the idea that telomeres are connected to cellular aging has accumulated. For example, for the hypothesis to be viable, normal cells should have little or no telomerase activity. Otherwise these cells would simply continue to extend their telomeres as they shortened. Indeed, many normal cells have limited telomerase activity. In contrast, cells that have increased proliferative capacity, such as stem cells and cells derived from tumors, have higher levels of telomerase activity. Indeed, studies of cancer cells in culture indicate that they can divide indefinitely. A second important experiment in support of the model showed that expression of telomerase in normal cells effectively immortalized the cells.

The finding of elevated telomerase activity in cancer cells has led to the hypothesis that telomeres may represent a method to limit the growth capacity of cells that have lost normal growth control. If true, this may explain why multicellular organisms have not allowed telomerase activity to be present in all cells. Indeed, there are numerous efforts seeking telomerase inhibitors as chemotherapeutic agents. The elevation of telomerase activity in cancer cells also suggests that globally activating telomerase would not be a wise method to seek immortality!

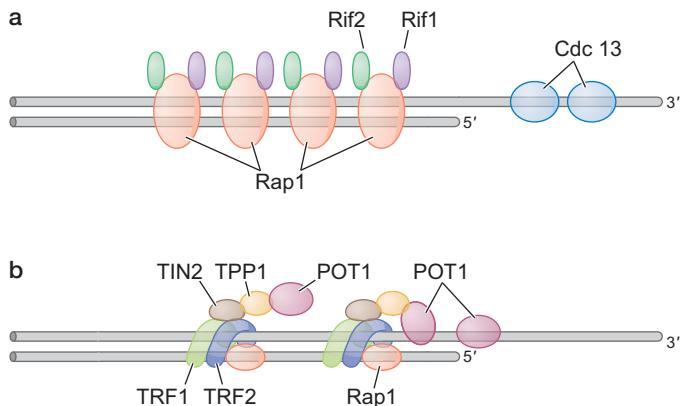
### Telomere-Binding Proteins Regulate Telomerase Activity and Telomere Length

Although extension of telomeres by telomerase could theoretically go on indefinitely, proteins bound to the double-strand regions of the telomere regulate telomere length (Fig. 9-40). In *S. cerevisiae* cells, proteins bound



**FIGURE 9-39** Extension of the 3' end of the telomere by telomerase solves the end replication problem. Although telomerase only directly extends the 3' end of the telomere, by providing an additional template for lagging-strand DNA synthesis, both ends of the chromosome are extended.

**FIGURE 9-40 Telomere-binding proteins.** Telomere-binding proteins that regulate telomerase activity are illustrated for *S. cerevisiae* and human cells. (a) *S. cerevisiae* cells. Rap1 directly binds to the double-stranded telomere repeat DNA, whereas Rif1 and Rif2 associate with the telomere indirectly by binding to Rap1. All three proteins have been implicated in the inhibition of telomerase activity. Cdc13 binds to the single-stranded telomere repeat DNA and is involved in telomerase recruitment. (b) Human cells. TRF1 and TRF2 bind directly to the double-stranded telomere repeat DNA. The human homolog of Rap1 as well as TIN2, TPP1, and POT1 all associate with either TRF1 or TRF2. Together these proteins form a complex that is called Shelterin for its ability to “shelter” the telomeres from the action of DNA repair enzymes. POT1 also binds directly to the single-stranded telomere repeat DNA and inhibits telomerase activity.



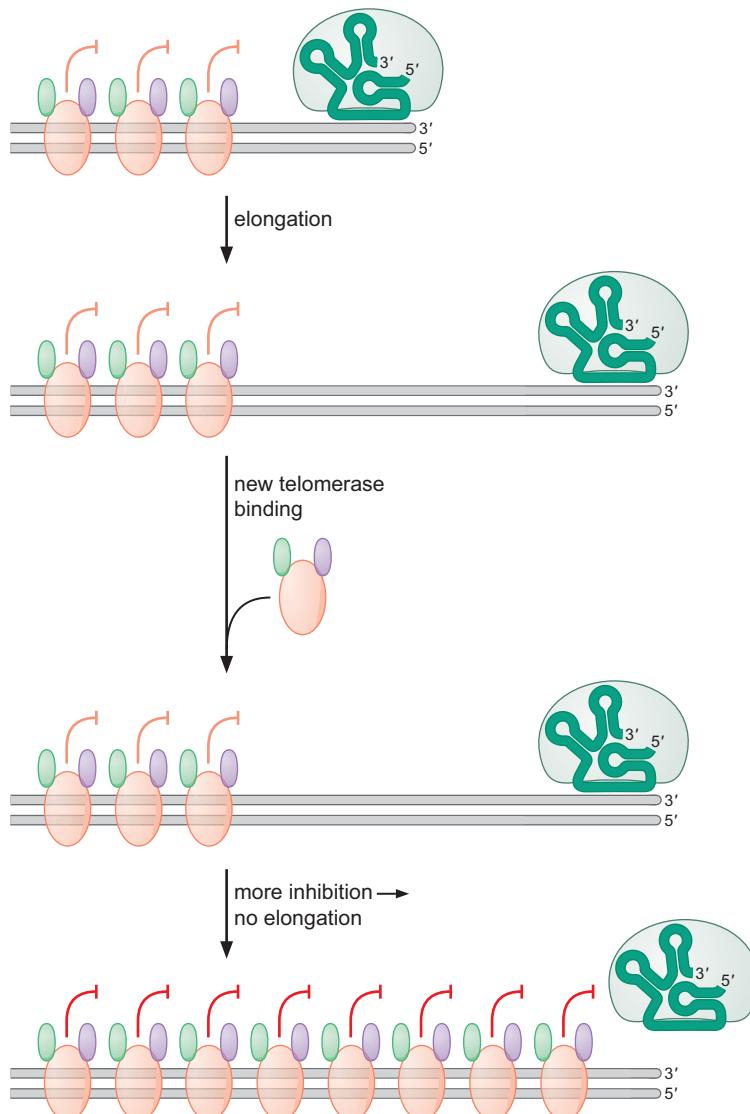
to the telomere act as weak inhibitors of telomerase activity (Fig. 9-41). When there are relatively few copies of the telomere sequence repeat, few of these proteins are bound to the telomere, and telomerase can extend the 3'-OH end of the telomere. As the telomere becomes longer, more of the telomere-binding proteins accumulate and inhibit telomerase extension of the 3'-OH end of the telomere. This simple negative-feedback loop mechanism (longer telomeres inhibit telomerase) is a robust method to maintain a similar telomere length at the ends of all chromosomes.

Proteins that recognize the single-stranded form of the telomere can also modulate telomerase activity. In *S. cerevisiae* cells, the Cdc13 protein binds to single-stranded regions of the telomere. Studies of this protein indicate that it recruits telomerase to the telomeres. Thus, Cdc13 is a positive activator of telomerase. In contrast, the human protein that binds to single-stranded telomeric DNA, POT1, acts in the opposite manner—that is, as an inhibitor of telomerase activity. *In vitro* studies show that POT1 binding to single-stranded telomere DNA inhibits telomere activity. Cells that lack this protein show dramatically increased telomere DNA length. Interestingly, this protein interacts indirectly with the double-strand telomere-binding proteins in human cells. It has been proposed that as telomeres increase in length, more POT1 is recruited, thereby increasing the likelihood that it binds to the ssDNA ends of the telomere and inhibits telomerase.

### Telomere-Binding Proteins Protect Chromosome Ends

In addition to their role in regulating telomerase function, telomere-binding proteins also play a crucial role in protecting the ends of chromosomes. Ordinarily in a cell, the presence of a DNA end is considered the sign of a double-stranded break in the DNA, which is targeted by the DNA repair machinery (see Chapter 10). The most common outcome of this repair is to initiate recombination with other DNA in the genome. (In a diploid cell this recombination is targeted to the intact copy of the broken chromosome.) Whereas this response is appropriate for random DNA breaks, it would be disastrous for the telomeres to participate in the same events. Attempts to repair telomeres in the same manner as double-stranded DNA breaks would lead to chromosome fusion events, which eventually result in random chromosome breaks.

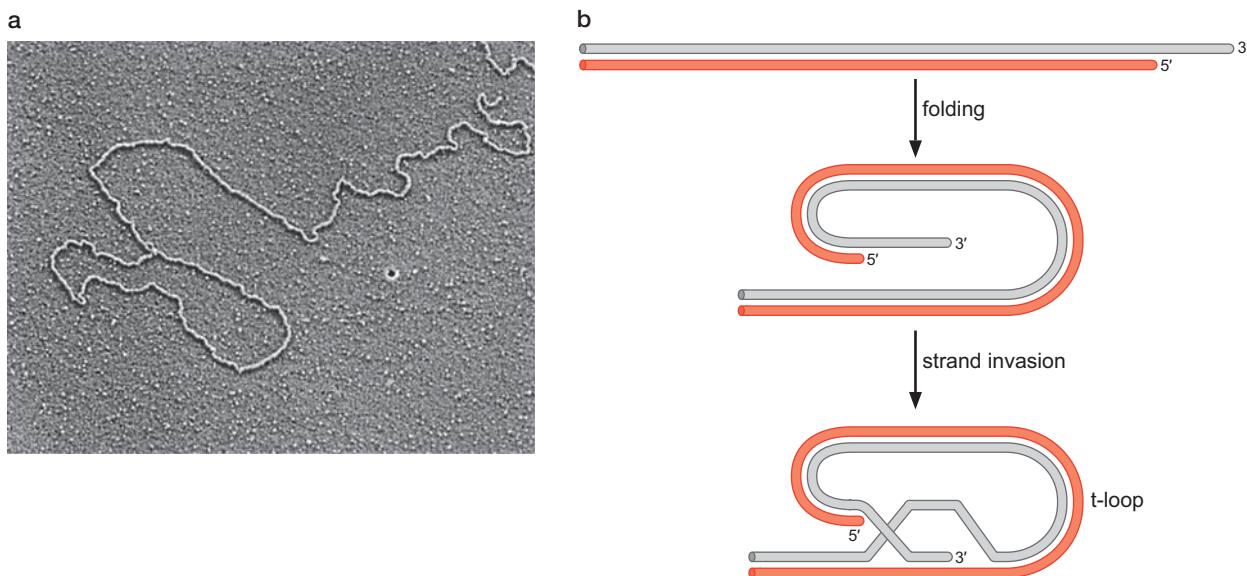
What protects the telomeres from this fate? The simple answer is that the proteins bound at the telomere distinguish telomeres from other DNA ends in the cell. Elimination of these proteins leads to the recognition of the telomeres as normal DNA breaks. It is possible that protection is conferred



**FIGURE 9-41** Telomere length regulation by telomere-binding proteins. When telomeres are relatively short, few telomere-binding proteins will be present, and inhibition of telomerase is weak. Under these conditions, telomerase can extend the 3' end of the telomere. When these regions are made double-stranded by the action of the lagging-strand DNA synthesis machinery, additional telomere-binding proteins can associate with the telomere. Binding of these proteins increases the level of inhibition, preventing further elongation by telomerase. (Adapted, with permission, from Smogorzewska A. and de Lange T. 2004. *Annu. Rev. Biochem.* **73**: 177–208, Fig. 3a. © Annual Reviews.)

simply by coating the telomere with binding proteins. Studies of the structure of the human telomere have led to an alternative possibility. Telomeres isolated from human cells were observed by electron microscopy and found to form a loop rather than a linear structure (Fig. 9-42a). Subsequent analysis indicated that this structure, called a **t-loop**, was formed by the 3'-ssDNA end of the telomere invading the dsDNA region of the telomere (Fig. 9-42b). It has been proposed that by forming a t-loop, the end of the telomere is masked and cannot be recognized as a normal DNA end. Interestingly, purified TRF2 is capable of directing t-loop formation with purified telomere DNA.

The t-loop structure may also be relevant to telomere length control. Just as the loop structure may protect the telomere from DNA repair enzymes, it is also likely that telomerase cannot recognize this form of the telomere, because it lacks an obvious single-strand 3' end. It has been proposed that as telomeres shorten, they would have an increasingly difficult time forming the t-loop, thereby allowing increased access to the 3' end of the telomere.



**FIGURE 9-42** Telomeres form a looped structure in the cell. (a) An electron micrograph of a telomere isolated from a human cell. The loop found at the end of the DNA included the ssDNA at the end of the telomere and is referred to as a t-loop. The end of the DNA in the upper right-hand corner would be attached to the rest of the chromosome. (Reprinted, with permission, from Griffith J.D. et al. 1999. *Cell* 97: 503–514, Fig. 3f. © Elsevier.) (b) An illustration of the proposed mechanism of t-loop formation. The first step folds the telomere such that the ssDNA at the end of the telomere can access the dsDNA telomeric repeats. Once the ssDNA end is positioned properly, it can invade the dsDNA repeats and form a helix with the complementary strand, displacing the other strand of the dsDNA. This is called a displacement loop and is a common intermediate in homologous recombination (see Chapter 11). It is likely that telomere-binding proteins and other cellular proteins (e.g., recombination proteins) facilitate this process. Note how the folding process would be increasingly difficult as the telomere becomes shorter.

## SUMMARY

DNA synthesis is dependent on the presence of two types of substrates: the four deoxynucleoside triphosphates (dATP, dGTP, dCTP, and dTTP) and the template DNA structure, a primer/template junction. The template DNA determines the sequence of incorporated nucleotides. The primer serves as the substrate for deoxynucleotide addition, each being added successively to the OH at its 3' end.

DNA synthesis is catalyzed by an enzyme called DNA polymerase that uses a single active site to add any of the four dNTP precursors. Structural studies of DNA polymerases reveal that these enzymes resemble a hand that grips the DNA and incoming nucleotide in the catalytic site. DNA polymerases are processive: Each time they bind a substrate, they add many nucleotides. Proofreading exonucleases further enhance the accuracy of DNA synthesis by acting like a “delete key” that removes incorrectly added nucleotides.

In the cell, both strands of a DNA template are duplicated simultaneously at a structure called the replication fork. Because the two strands of the DNA are antiparallel, only one of the template DNA strands can be replicated in a continuous fashion (called the leading strand). The other DNA strand (called the lagging strand) must be synthesized first

as a series of short DNA fragments, called Okazaki fragments. Each DNA strand is initiated with an RNA primer that is synthesized by an enzyme called primase. These primers must be removed to complete the replication process. After the replacement of the RNA primers with DNA, all of the separately primed lagging-strand DNA fragments are joined together to form one continuous DNA strand by DNA ligase.

An array of proteins in addition to the DNA polymerases coordinates and facilitates the DNA replication reaction. These additional factors facilitate the unwinding of the dsDNA template (DNA helicase), stabilize the ssDNA template (SSBs), and remove supercoils generated in front of the replication fork (topoisomerases). DNA polymerases are specialized to perform different events during DNA replication. Some are designed to be highly processive and others, only weakly processive. DNA sliding clamps enhance the processivity of the DNA polymerases that replicate large regions of DNA. These clamp proteins are topologically linked to DNA, but they are able to slide along the recently synthesized DNA while bound to the DNA polymerase. This interaction effectively prevents the attached DNA polymerase from dissociating from the primer/template junction.

Special protein complexes called sliding DNA clamp loaders use the energy of ATP binding and hydrolysis to place sliding clamps on the DNA near primer/template junctions.

Interactions between the proteins at the replication fork have an important role in DNA synthesis. In *E. coli*, the three DNA polymerases are part of a large complex called the DNA Pol III holoenzyme. Binding of the DNA Pol III holoenzyme to the DNA helicase stimulates the rate of DNA unwinding. Similarly, binding of primase to the DNA helicase increases its ability to synthesize RNA primers. Thus, the replication reaction works best when the entire array of replication proteins is present at the replication fork. Together, this set of proteins forms a complex called the replisome.

The initiation of DNA replication is directed by specific DNA sequences called replicators. The physical site of replication initiation is called an origin of replication. The replicator is specifically bound by a protein called the initiator, which stimulates the recruitment of other proteins required for the initiation of replication (such as DNA helicase) and, in some but not all cases, the unwinding of the origin DNA. The subsequent events in the initiation of DNA replication are largely driven by either protein–protein or nonspecific protein–DNA interactions.

In eukaryotic cells, the initiation of DNA replication is tightly regulated to ensure that every nucleotide of every

chromosome is replicated once and only once per round of cell division. This tight regulation is accomplished by controlling loading and activation of the replicative helicase during the cell cycle. During the G<sub>1</sub> phase of the cell cycle, helicases can be loaded but not activated. During the remainder of the cell cycle (the S, G<sub>2</sub>, and M phases), loaded helicases can be activated, leading to the initiation of DNA replication, but no new helicase loading can occur. Thus, each replicator can direct only one round of replication initiation per cell cycle, ensuring that the DNA is replicated exactly once.

Finishing DNA replication requires the action of specific enzymes. For circular chromosomes, type II DNA topoisomerases separate the topologically linked circular products from one another. Linear chromosomes also require special proteins to ensure their complete replication. In eukaryotic cells, a specialized DNA polymerase called telomerase allows the ends of the chromosome (called telomeres) to act as a unique origin of replication. By extending the 3' ends of the telomere, telomerase eliminates the progressive loss of chromosome ends that conventional DNA synthesis by the replication fork machinery would cause. Proteins bound to telomeric DNA act to regulate the activity of telomerase and protect the ends of chromosomes from degradation and recombination.

## BIBLIOGRAPHY

### Books

DePamphilis M.L., Bell S., and Méchali M. 2012. *DNA replication*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York. In press.

### The Chemistry of DNA Synthesis

Brautigam C.A. and Steitz T.A. 1998. Structural and functional insights provided by crystal structures of DNA polymerases. *Curr. Opin. Struct. Biol.* **8**: 54–63.

Jäger J. and Pata J.D. 1999. Getting a grip: Polymerases and their substrate complexes. *Curr. Opin. Struct. Biol.* **9**: 21–28.

### The Mechanism of DNA Polymerase

Doublie S. and Ellenberger T. 1998. The mechanism of action of T7 DNA polymerase. *Curr. Opin. Struct. Biol.* **8**: 704–712.

Steitz T.A. 1998. A mechanism for all polymerases. *Nature* **391**: 231–232.

—. 2006. Visualizing polynucleotide polymerase machines at work. *EMBO J.* **25**: 3458–3468.

### The Replication Fork

Corn J.E. and Berger J.M. 2006. Regulation of bacterial priming and daughter strand synthesis through helicase–primase interactions. *Nucleic Acids Res.* **34**: 4082–4088.

McHenry C.S. 2011. DNA replicases from a bacterial perspective. *Annu. Rev. Biochem.* **80**: 403–436.

O'Donnell M. and Kuriyan J. 2006. Clamp loaders and replication initiation. *Curr. Opin. Struct. Biol.* **16**: 405–415.

Vos S.M., Tretter E.M., Schmidt B.H., and Berger J.M. 2011. All tangled up: How cells direct, manage and exploit topoisomerase function. *Nat. Rev. Mol. Cell Biol.* **12**: 827–841.

### The Specialization of DNA Polymerases

Lovett S.T. 2007. Polymerase switching in DNA replication. *Mol. Cell.* **27**: 523–526.

### Initiation of DNA Replication

Arias E.E. and Walter J.C. 2007. Strength in numbers: Preventing rereplication via multiple mechanisms in eukaryotic cells. *Genes Dev.* **21**: 497–518.

Duderstadt K.E. and Berger J.M. 2008. AAA+ ATPases in the initiation of DNA replication. *Crit. Rev. Biochem. Mol. Biol.* **43**: 163–187.

Remus D. and Diffley J.F.X. 2009. Eukaryotic DNA replication control: Lock and load, then fire. *Curr. Opin. Cell Biol.* **21**: 771–777.

Robinson N.P. and Bell S.D. 2005. Origins DNA replication in the three domains of life. *FEBS J.* **272**: 3757–3766.

### Finishing Replication

Blackburn E.H. and Collins K. 2011. Telomerase: An RNP enzyme synthesizes DNA. *Cold Spring Harb. Perspect. Biol.* **3**: a003558. doi: 10.1101/cshperspect.a003558.

Linger B.R. and Price C.M. 2009. Conservation of telomere protein complexes: Shuffling through evolution. *Crit. Rev. Biochem. Mol. Biol.* **44**: 434–446.

**QUESTIONS**

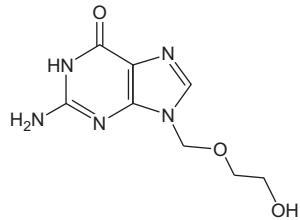
For answers to even-numbered questions, see Appendix 2: Answers.

**Question 1.** Name the two substrates for DNA synthesis. Explain why each is necessary for DNA synthesis.

**Question 2.** List the mechanistic steps of DNA synthesis starting with the primed template and deoxynucleoside triphosphate.

**Question 3.** Explain why DNA synthesis is coupled to the hydrolysis of pyrophosphate.

**Question 4.** The antiviral drug Acyclovir (structure pictured below) is used to treat infections caused by double-stranded DNA viruses such as herpes simplex virus. Acyclovir acts at the level of DNA synthesis.

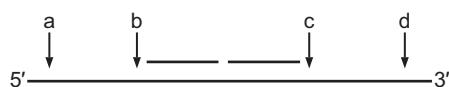


- A. Acyclovir functions as the analog of what deoxynucleoside?
- B. Acyclovir cannot be incorporated into the DNA unless it is modified by a virally encoded kinase. Explain why the activity of a kinase is required for Acyclovir to be incorporated during DNA synthesis.

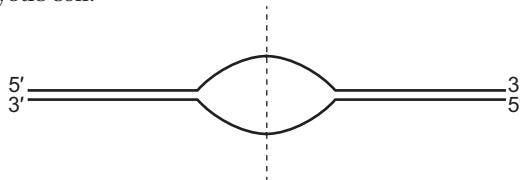
**Question 5.** Explain why magnesium chloride is added to the buffer used for PCR (polymerase chain reaction).

**Question 6.** Hypothesize why some DNA polymerases lack exonuclease activity without significantly contributing to the number of mismatches introduced during DNA replication.

**Question 7.** Shown below is a long template strand of DNA where lagging strand DNA synthesis is occurring. The short horizontal lines represent two Okazaki fragments that have already been made. In the context of the replication fork, select the letter (a–d) that indicates where primase will synthesize the next RNA primer. Why did you choose that location?



**Question 8.** Below is a picture of a single origin of replication in a eukaryotic cell.



- A. With respect to the dotted line, in which direction(s)—right, left, or both—does total replication proceed?

**B.** On the right-hand side of the dotted line, the replication of which template strand (top or bottom) will be continuous by DNA polymerase?

**C.** On the left-hand side of the dotted line, the complete replication of which template strand (top or bottom) will be more affected by a mutation that causes DNA ligase to be partially functional?

**Question 9.** You want to set up an assay starting with a sliding clamp bound to DNA. What special property must the DNA have to establish binding between the sliding clamp and DNA? What other protein components must be in the reaction to ensure binding?

**Question 10.**

**A.** Explain how the time required to complete replication of the *E. coli* genome is 40 min, yet the cells can divide every 20 min.

**B.** Why is telomerase not required in *E. coli* cells?

**Question 11.**

**A.** Describe the role of a DNA helicase at a replication fork.

**B.** As a result of DNA helicase activity, topoisomerases are also required during replication. Explain how topoisomerases help DNA helicases function more efficiently.

**C.** During PCR, you do not have to add DNA helicase to the reaction. Explain why not.

**Question 12.** In *E. coli*, DNA polymerase I possesses 5' exonuclease and 3' exonuclease activities, whereas DNA polymerase III possesses 3' exonuclease activity. Explain the functionality behind the differences in exonuclease activities associated with these two DNA polymerases.

**Question 13.** Researchers have mapped mutations associated with diseases such as dyskeratosis congenital to the DNA encoding the RNA component of telomerase. Describe why defects in the RNA component of telomerase are associated with diseases.

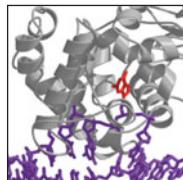
**Question 14.** Review Box 9-1. Incorporation assays measure DNA synthesis using  $^{32}\text{P}$ -labeled dNTPs (where  $^{32}\text{P}$  replaces the  $\alpha$  phosphate of the dNTP).

**A.** If you use dNTPs labeled at the  $\beta$  or  $\gamma$  phosphates, you do not detect any radioactivity in the newly synthesized DNA. Explain why.

**B.** Following incorporation of  $^{32}\text{P}$ -labeled dNTPs, you must separate the unincorporated  $^{32}\text{P}$ -labeled dNTPs from the newly synthesized DNA strand before measuring the amount of  $^{32}\text{P}$  incorporation. Explain how gel electrophoresis serves as a means to separate unincorporated  $^{32}\text{P}$ -labeled dNTPs from the newly synthesized DNA strand.

**C.** For the filter binding shown in Box 8-1 Figure 2, describe a negative control that would ensure that your filter is separating the unincorporated  $^{32}\text{P}$ -labeled dNTPs from the DNA.

CHAPTER 10



# The Mutability and Repair of DNA

THE PERPETUATION OF THE GENETIC MATERIAL from generation to generation depends on maintaining a low mutation rate. A high mutation rate in the germline would destroy the species, and high mutation rates in the soma would destroy the individual. Living cells require the correct functioning of thousands of genes, each of which could be damaged by a mutation at many sites in its protein-coding sequence or in flanking sequences that govern its expression or the processing of its messenger RNA (mRNA).

If progeny are to have a good chance at survival, DNA sequences must be passed on largely unchanged in the germline. Likewise, the specialized cells of the adult organism could not perform their mission if mutation rates in the soma were high. Cancer, for example, arises from cells that have lost the capacity to grow and divide in a controlled manner as a consequence of damage to genes that encode proteins that govern the cell cycle. If the mutation rates in the soma were high, the incidence of cancer would be catastrophic and unsustainable.

At the same time, if the genetic material were perpetuated with perfect fidelity, the genetic variation needed to drive evolution would be lacking, and new species, including humans, would not have arisen. Thus, life and biodiversity depend on a happy balance between mutation and its repair. In this chapter, we consider the causes of mutation and the systems that are responsible for reversing or correcting, and thereby minimizing, damage to the genetic material.

Two important sources of mutations are inaccuracy in DNA replication and chemical damage to the genetic material. Replication errors arise from tautomerization, which, as we have seen in Chapter 9, imposes an upper limit on the accuracy of base pairing during DNA replication. The enzymatic machinery for replicating DNA attempts to cope with the misincorporation of incorrect nucleotides through a proofreading mechanism, but some errors escape detection. In addition, DNA is a complex and fragile organic molecule of finite chemical stability. Not only does it undergo spontaneous damage such as the loss of bases, but also natural and unnatural chemicals and radiation break its backbone and chemically alter its bases. Simply put, errors in replication and damage to the genetic material from the environment are unavoidable. A third important source of mutation is the class of insertions generated by DNA elements known as **transposons**. Transposition is a major topic in its own right, which we shall consider in detail in Chapter 12.

## OUTLINE

- Replication Errors and Their Repair, 314
  - DNA Damage, 320
  - Repair and Tolerance of DNA Damage, 324
- Visit Web Content for Structural Tutorials and Interactive Animations

Errors in replication and damage to DNA have two consequences. One is, of course, the introduction of permanent changes to the DNA (**mutations**), which can alter the coding sequence of a gene or its regulatory sequences. The second consequence is that some chemical alterations to the DNA prevent its use as a template for replication and transcription. The effects of mutations generally become manifest only in the progeny of the cell in which the sequence alteration has occurred, but DNA **lesions** or structural changes to the DNA that impede replication or transcription can have immediate effects on cell function and survival.

The challenge for the cell is twofold. First, it must scan the genome to detect errors in synthesis and damage to the DNA. Second, it must mend the lesions and do so in a way that, if possible, restores the original DNA sequence. Here, we discuss errors that are generated during replication, lesions that arise from spontaneous damage to DNA, and damage that is wrought by chemical agents and radiation. In each case, we consider how the alteration to the genetic material is detected and how it is properly repaired or tolerated. Among the questions we address are the following: How is the DNA mended rapidly enough to prevent errors from becoming set in the genetic material as mutations? How does the cell distinguish the parental strand from the daughter strand in repairing replication errors? How does the cell restore the proper DNA sequence when, because of a break or severe lesion, the original sequence can no longer be read? How does the cell cope with lesions that block replication? The answers to these questions depend on the kind of error or lesion that needs to be repaired.

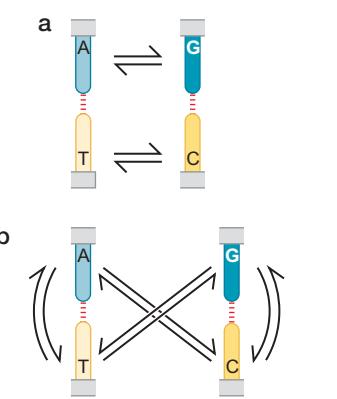
We begin by considering errors that occur during replication. We then consider various kinds of lesions that arise spontaneously or from environmental assaults, before turning to the multiple repair mechanisms that allow the cell to mend replication errors and DNA damage. Lastly, we review the pathways that allow DNA damage to be tolerated during replication so as to prevent cell death and allow the DNA lesion to be repaired subsequently. We will see that multiple overlapping systems enable the cell to cope with a wide range of insults to DNA, underscoring the investment that living organisms make in the preservation of the genetic material.

## REPLICATION ERRORS AND THEIR REPAIR

### The Nature of Mutations

Mutations include almost every conceivable permanent change in DNA sequence. The simplest mutations are switches of one base for another. There are two kinds: **transitions**, which are pyrimidine-to-pyrimidine and purine-to-purine substitutions, such as T to C and A to G; and **transversions**, which are pyrimidine-to-purine and purine-to-pyrimidine substitutions, such as T to G or A and A to C or T (Fig. 10-1). Other simple mutations are insertions or deletions of a nucleotide or a small number of nucleotides. Mutations that alter a single nucleotide are called **point mutations**.

Other kinds of mutations cause more drastic changes in DNA, such as extensive insertions and deletions and gross rearrangements of chromosome structure. Such changes might be caused, for example, by the insertion of a transposon, which typically places many thousands of nucleotides of foreign DNA in the coding or regulatory sequences of a gene (see Chapter 12) or by the aberrant actions of cellular recombination processes. The overall rate at which new mutations arise spontaneously at any given site on the chromosome ranges from  $\sim 10^{-6}$  to  $10^{-11}$  per round of DNA replication,



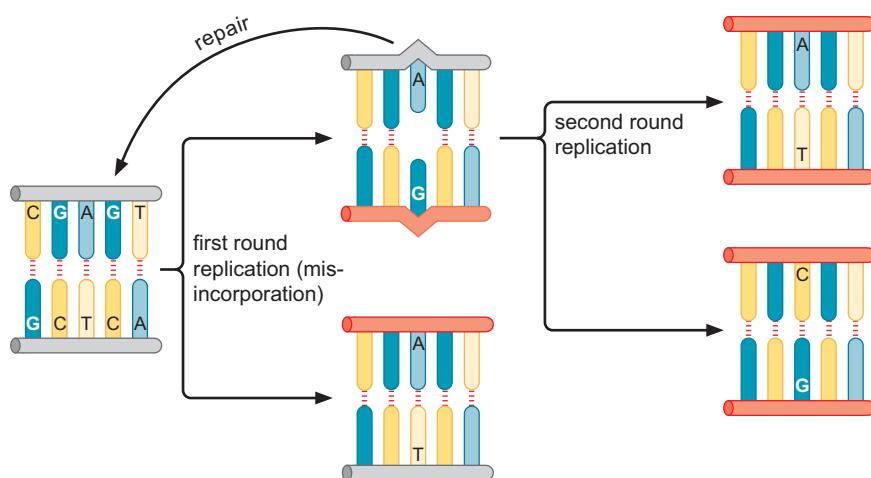
**FIGURE 10-1** Base-change substitutions. (a) Transitions. (b) Transversions.

with some sites on the chromosome being “hot spots,” where mutations arise at high frequency, and other sites undergoing alterations at a comparatively low frequency.

One kind of sequence that is particularly prone to mutation merits special comment because of its importance in human genetics and disease. These mutation-prone sequences are repeats of simple di-, tri-, or tetranucleotide sequences, which are known as **DNA microsatellites**. One well-known example involves repeats of the dinucleotide sequence CA. Stretches of CA repeats are found at many widely scattered sites in the chromosomes of humans and some other eukaryotes. The replication machinery has difficulty copying such repeats accurately, frequently undergoing “slippage.” This slippage increases or reduces the number of copies of the repeated sequence. As a result, the CA repeat length at a particular site on the chromosome is often highly polymorphic in the population. This polymorphism provides a convenient physical marker for mapping inherited mutations, such as mutations that increase the propensity to certain diseases in humans (see Box 10-1, Expansion of Triple Repeats Causes Disease).

### Some Replication Errors Escape Proofreading

As we have seen, the replication machinery achieves a remarkably high degree of accuracy using a proofreading mechanism, the 3' → 5' exonuclease component of the replisome, which removes wrongly incorporated nucleotides (as we discussed in Chapter 9). Proofreading improves the fidelity of DNA replication by a factor of ~100. The proofreading exonuclease is not, however, foolproof. Some misincorporated nucleotides escape detection and become a mismatch between the newly synthesized strand and the template strand. Three different nucleotides can be misincorporated opposite each of the four kinds of nucleotides in the template strand (e.g., T, G, or C opposite a T in the template) for a total of 12 possible mismatches (T:T, T:G, T:C, etc.). If the misincorporated nucleotide is not subsequently detected and replaced, the sequence change will become permanent in the genome: during a second round of replication, the misincorporated nucleotide, now part of the template strand, will direct the incorporation of its complementary nucleotide into the newly synthesized strand (Fig. 10-2). At this point, the mismatch will no longer exist; instead, it will have resulted in a permanent change (a mutation) in the DNA sequence.



**FIGURE 10-2** Replication can change a misincorporated base into a permanent mutation. A potential mutation may be introduced by misincorporation of a base in the first round of replication. In the second round of replication, the misincorporated base becomes permanent in the DNA sequence and is now a mutation.

## ► MEDICAL CONNECTIONS

### Box 10-1 Expansion of Triple Repeats Causes Disease

Another well-known example of error-prone sequences is repeats of the triplet nucleotide sequences CGG and CAG in certain genes. In humans, such triplet repeats are often found to undergo expansion from one generation to the next, resulting in diseases that are progressively more severe in the children and grandchildren of afflicted individuals. Examples of diseases that are caused by triplet expansion are adult muscular (myotonic) dystrophy; fragile X syndrome, which causes mental retardation; and Huntington's disease, which causes neurodegeneration. CAG is the codon for glutamine, and its expansion in the coding sequence for the huntingtin protein results in an extended stretch of glutamine

residues in the mutant protein in patients with Huntington's disease. Recent research indicates that this polyglutamine stretch interferes with the normal interaction between a glutamine-rich patch in a transcription factor called Sp1 and a corresponding glutamine-rich patch in TAFII130, a subunit of a component of the transcription machinery called TFIID (see Chapter 13). This interference impairs transcription in neurons of the brain, including the transcription of the gene for the receptor of a neurotransmitter. Similar polyglutamine stretches from CAG expansions in other genes may also exert their effects by disrupting interactions between transcription factors and TAFII130.

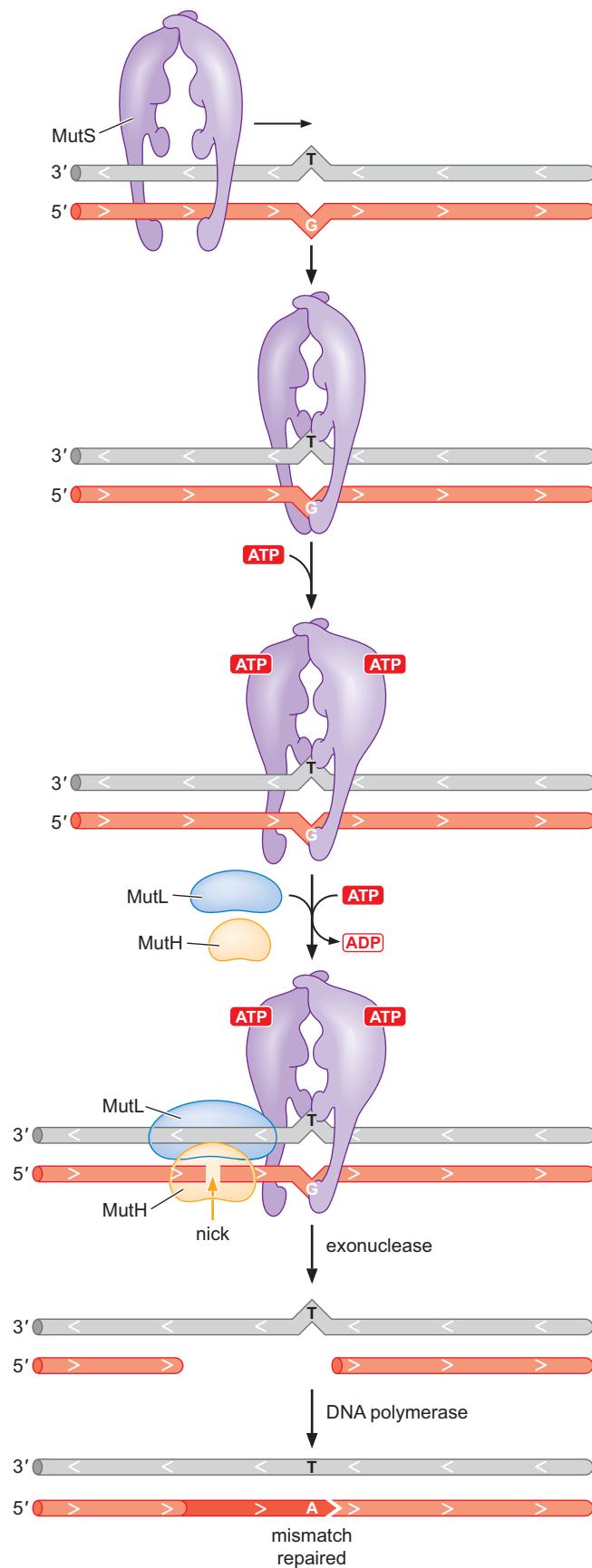
### Mismatch Repair Removes Errors That Escape Proofreading

Fortunately, a mechanism exists for detecting mismatches and repairing them. Final responsibility for the fidelity of DNA replication rests with this **mismatch repair system**, which increases the accuracy of DNA synthesis by an additional two to three orders of magnitude. The mismatch repair system faces two challenges. First, it must scan the genome for mismatches. Because mismatches are transient (they are eliminated following a second round of replication when they result in mutations), the mismatch repair system must rapidly find and repair mismatches. Second, the system must correct the mismatch accurately; that is, it must replace the misincorporated nucleotide in the newly synthesized strand and not the correct nucleotide in the parental strand.

In *Escherichia coli*, mismatches are detected by a dimer of the mismatch repair protein **MutS** (Fig. 10-3) (see Structural Tutorial 10-1). MutS scans the DNA, recognizing mismatches from the distortion they cause in the DNA backbone. MutS embraces the mismatch-containing DNA, inducing a pronounced kink in the DNA and a conformational change in MutS itself (Fig. 10-4). A key to the specificity of MutS is that DNA containing a mismatch is much more readily distorted than properly base-paired DNA. MutS has an ATPase activity that is required for mismatch repair, but its precise role in repair is not understood. The complex of MutS and the mismatch-containing DNA recruits **MutL**, a second protein component of the repair system. MutL, in turn, activates **MutH**, an enzyme that causes an incision or nick on one strand near the site of the mismatch. Nicking is followed by the action of a specific helicase (UvrD) and one of three exonucleases (see later discussion). The helicase unwinds the DNA, starting from the incision and moving in the direction of the site of the mismatch, and the exonuclease progressively digests the displaced single strand, extending to and beyond the site of the mismatched nucleotide. This action produces a single-strand gap, which is then filled in by DNA polymerase III (Pol III) and sealed with DNA ligase. The overall effect is to remove the mismatch and replace it with the correctly base-paired nucleotide.

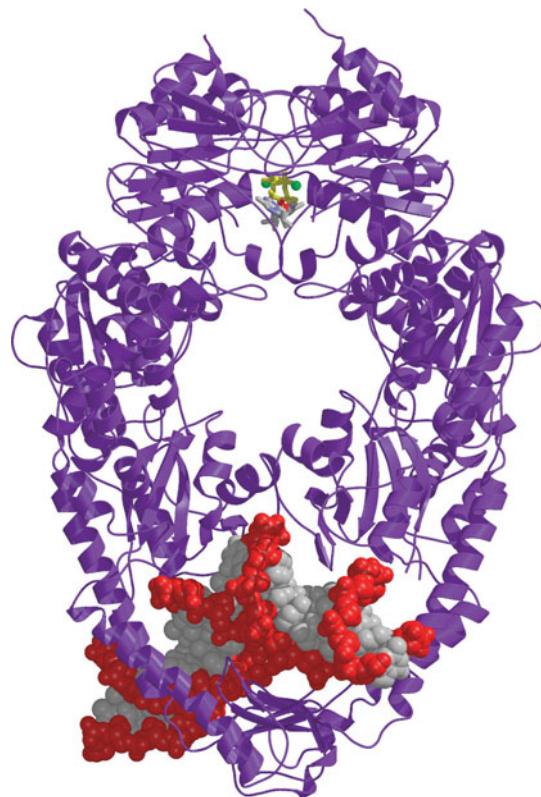
But how does the *E. coli* mismatch repair system know which of the two mismatched nucleotides to replace? If repair occurred randomly, then half the time the error would become permanently established in the DNA. The answer is that *E. coli* tags the parental strand by transient hemimethylation as we now describe.





**FIGURE 10-3** Mismatch repair pathway for the repair of replication errors. MutS embraces mismatch-containing DNA, inducing a kink (not shown, but see Fig. 10-4). In subsequent steps, MutS recruits MutL and MutH, and the ATPase activity of MutS catalyzes the hydrolysis of ATP. MutH is an endonuclease that creates a nick in the DNA near the site of the mismatch. Next, an exonuclease digests the nicked strand moving toward and beyond the mismatch. Finally, the resulting single-strand gap is filled in by DNA polymerase, eliminating the mismatch. (Adapted, with permission, from Junop M.S. et al. 2001. *Mol. Cell* 7: 1–12, Fig. 6b. © Elsevier.)

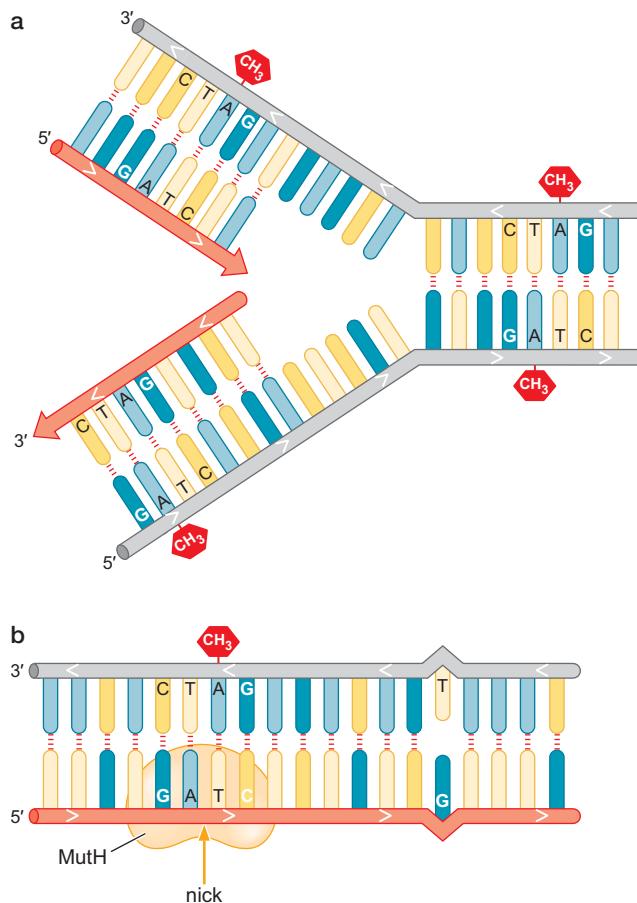
**FIGURE 10-4** Crystal structure of the MutS–DNA complex. Notice the kink in the DNA, present near the bottom of the structure. In addition, near the top of the structure of the enzyme is ATP, shown in yellow, green, and red. The DNA is depicted as a space-filling representation with the backbone in red and bases in gray. (Junop M.S. et al. 2002. *Mol. Cell* 7: 1–12.) Image prepared with MolScript, BobScript, and Raster3D.



The *E. coli* enzyme **Dam methylase** methylates A residues on both strands of the sequence 5'-GATC-3'. The GATC sequence is widely distributed along the entire genome (occurring at about once every 256 bp [ $4^4$ ]), and all of these sites are methylated by the Dam methylase. When a replication fork passes through DNA that is methylated at GATC sites on both strands (fully methylated DNA), the resulting daughter DNA duplexes will be hemimethylated (i.e., methylated on only the parental strand). Thus, for a few minutes, until the Dam methylase catches up and methylates the newly synthesized strand, daughter DNA duplexes will be methylated only on the strand that served as a template (Fig. 10-5a). Thus, the newly synthesized strand is marked (it lacks a methyl group) and hence can be recognized as the strand for repair.

The MutH protein binds at such hemimethylated sites, but its endonuclease activity is normally latent. Only when MutH is contacted by MutL and MutS located at a nearby mismatch (which is likely to be within a distance of a few hundred base pairs) does MutH become activated as we described above. Just how this interaction takes place over distances of up to several hundred base pairs is uncertain, but recent evidence indicates that the MutS–MutL complex leaves the mismatch and moves along the DNA contour to reach MutH at the site of hemimethylation. Once activated, MutH selectively nicks the unmethylated strand, thus only newly synthesized DNA in the vicinity of the mismatch is removed and replaced (Fig. 10-5b). Methylation is therefore a “memory” device that enables the *E. coli* repair system to retrieve the correct sequence from the parental strand if an error has been made during replication.

Different exonucleases are used to remove single-stranded DNA between the nick created by MutH and the mismatch, depending on whether MutH cuts the DNA on the 5' or the 3' side of the misincorporated nucleotide. If the DNA is cleaved on the 5' side of the mismatch, then exonuclease VII or

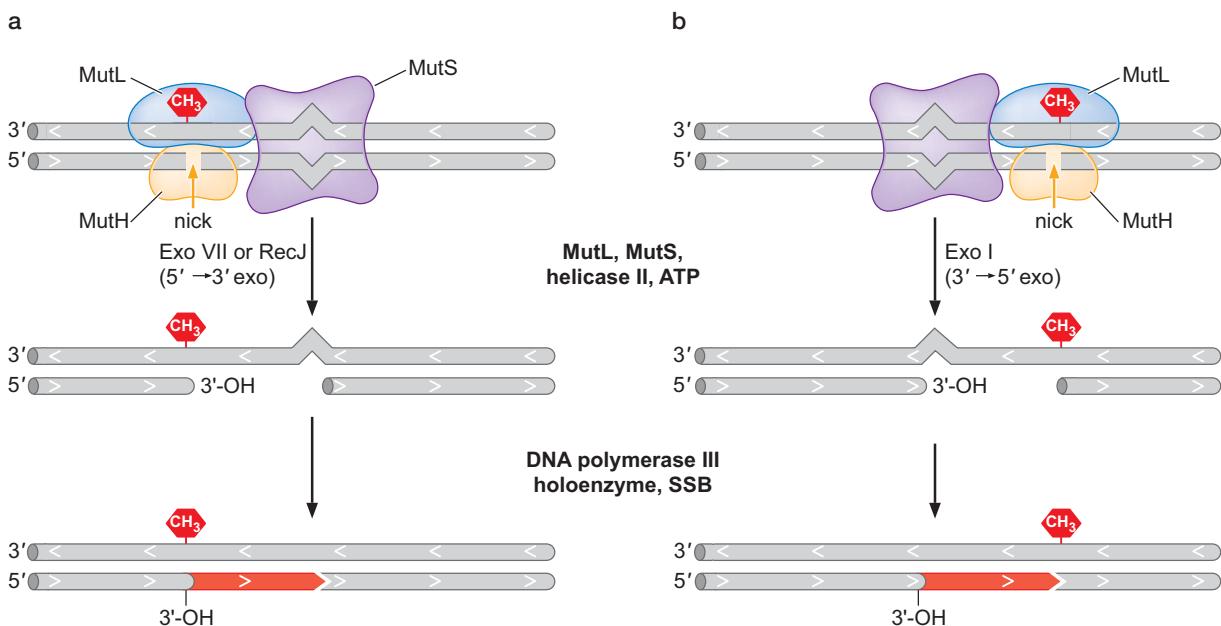


**FIGURE 10-5** Dam methylation at replication fork. (a) Replication generates hemimethylated DNA in *E. coli*. (b) MutH makes incision in unmethylated daughter strand.

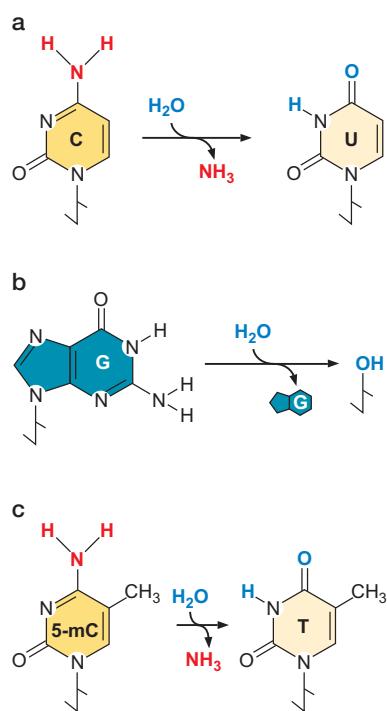
RecJ, which degrades DNA in a 5' → 3' direction, removes the stretch of DNA from the MutH-induced cut through the misincorporated nucleotide. Conversely, if the nick is on the 3' side of the mismatch, then the DNA is removed by exonuclease I, which degrades DNA in a 3' → 5' direction. As we have seen, after removal of the mismatched base, DNA Pol III fills in the missing sequence (Fig. 10-6).

Eukaryotic cells also repair mismatches and do so using homologs to MutS (called MSH proteins for “MutS homologs”) and MutL (called MLH and PMS). Indeed, eukaryotes have multiple MutS-like proteins with different specificities. For example, one is specific for simple mismatches, whereas another recognizes small insertions or deletions resulting from “slippage” during DNA replication. Dramatic evidence that mismatch repair has a critical role in higher organisms came from the discovery that a genetic predisposition to colon cancer (hereditary nonpolyposis colorectal cancer) is due to a mutation in the genes for human homologs of MutS (specifically the MSH2 homolog) and MutL.

Even though eukaryotic cells have mismatch repair systems, they lack MutH and the clever trick of using hemimethylation to tag the parental strand as found in *E. coli*. (Indeed, most bacteria lack Dam methylase and are also unable to use hemimethylation to mark the newly synthesized strand.) How then does the mismatch repair system know which of the two strands to correct? Lagging-strand synthesis, as we saw in Chapter 9, takes place discontinuously with the formation of Okazaki fragments that are joined to previously synthesized DNA by DNA ligase. Before the ligation step, the Okazaki fragment is separated from previously synthesized DNA by



**FIGURE 10-6** Directionality in mismatch repair: exonuclease removal of mismatched DNA. For simplicity, DNA-bound MutH is shown as being immediately adjacent to MutS at the mismatch. (a) Unmethylated GATC is 5' of mutation. (b) Unmethylated GATC is 3' of mutation.



**FIGURE 10-7** Common types of hydrolytic DNA damage. (a) Deamination of cytosine creates uracil. (b) Depurination of guanine by hydrolysis creates apurinic deoxyribose. (c) Deamination of 5-methylcytosine generates a natural base in DNA, thymine.

a nick, which can be thought of as being equivalent to the nick created in *E. coli* by MutH on the newly synthesized strand. Indeed, extracts of eukaryotic cells will repair mismatches in artificial templates that contain a nick and do so selectively on the strand that carries the nick. Recent results indicate that human homologs of MutS (MSH) interact with the sliding-clamp component of the replisome (PCNA, which we discussed in Chapter 9), and would thereby be recruited to the site of discontinuous DNA synthesis on the lagging strand. Interaction with the sliding clamp could also recruit mismatch repair proteins to the 3' (growing) end of the leading strand.

## DNA DAMAGE

### DNA Undergoes Damage Spontaneously from Hydrolysis and Deamination

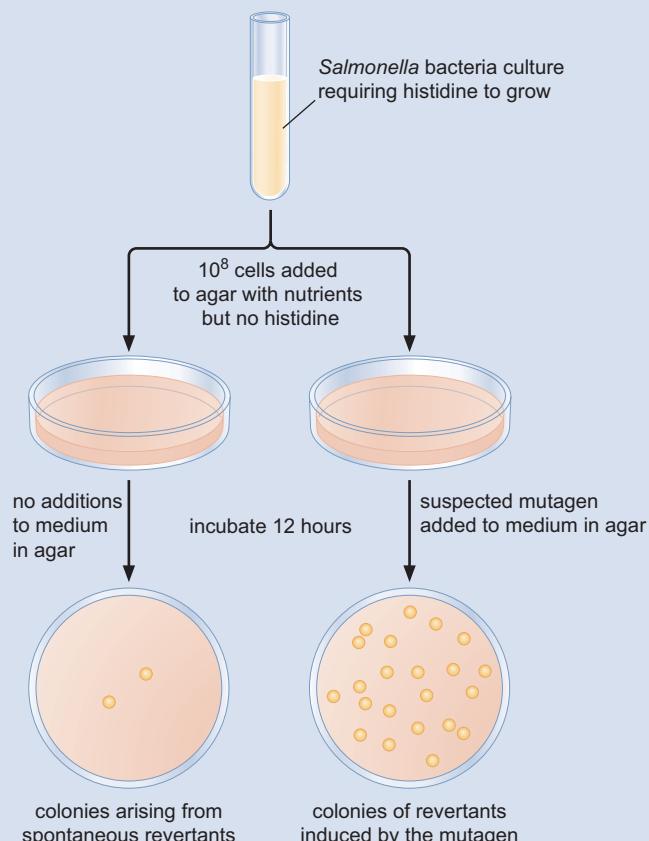
Mutations arise not only from errors in replication, but also from damage to the DNA. Some damage is caused, as we shall see, by environmental factors, such as radiation and so-called **mutagens**, which are chemical agents that increase mutation frequency (see Box 10-2, The Ames Test). But DNA also undergoes spontaneous damage from the action of water. (This is ironic because the proper structure of the double helix depends on an aqueous environment.)

The most frequent and important kind of hydrolytic damage is deamination of the base cytosine (Fig. 10-7a). Under normal physiological conditions, cytosine undergoes spontaneous deamination, thereby generating the unnatural (in DNA) base uracil. Uracil preferentially pairs with adenine and thus introduces that base in the opposite strand upon replication, rather

### MEDICAL CONNECTIONS

#### Box 10-2 The Ames Test

Determining the potential carcinogenic effects of chemicals in animals is time-consuming and expensive. However, because most tumor-causing agents are mutagens, the potential carcinogenic effects of chemicals can be conveniently assessed from their capacity to cause mutations. Bruce Ames of the University of California at Berkeley devised a simple test for the potential carcinogenic effects of chemicals based on their capacity to cause mutations in the bacterium *Salmonella typhimurium*. The Ames test uses a strain of *S. typhimurium* that is mutant for the operon responsible for the biosynthesis of the amino acid histidine. For example, the mutant operon might contain a missense or a frameshift mutation in one of the genes for histidine biosynthesis. As a consequence, the mutant cells fail to grow and form colonies on solid medium lacking histidine (Box 10-2 Fig. 1). However, if the mutant cells are treated with a chemical that is mutagenic (and hence potentially carcinogenic), the missense or frameshift mutation (depending on the nature of the mutagen) reverts in a small number of the mutant cells because of the chemical's action in the cell. This reversal restores the capacity of the cells to grow and form colonies on solid medium lacking histidine. A more potent mutagen translates into a greater number of revertant colonies. Some chemicals that cause cancers are not mutagenic to begin with, but rather are converted into mutagens by the liver, which metabolizes foreign substances. To identify chemicals that are converted into mutagens in the liver, the Ames test treats potential mutagens with a mixture of liver enzymes. Chemicals that are found to be mutagenic in the Ames test can then be tested for their potential carcinogenic effects in animals.



**BOX 10-2 FIGURE 1** The Ames test.

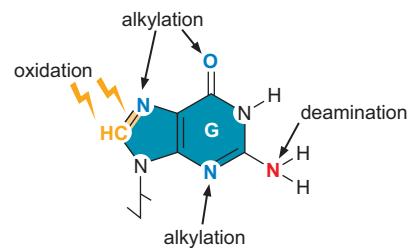
than the G that would have been directed by C. Adenine and guanine are also subject to spontaneous deamination. Deamination converts adenine to hypoxanthine, which hydrogen-bonds to cytosine rather than to thymine; guanine is converted to xanthine, which continues to pair with cytosine, although with only two hydrogen bonds. DNA also undergoes **depurination** by spontaneous hydrolysis of the N-glycosyl linkage, and this produces an abasic site (i.e., deoxyribose lacking a base) in the DNA (Fig. 10-7b).

Notice that, in contrast to the replication errors discussed above, all of these hydrolytic reactions result in unnatural alterations to the DNA. Apurinic sites are, of course, unnatural, and each of the deamination reactions generates an unnatural base. This situation allows changes to be recognized by the repair systems described later. This situation also suggests an explanation for why DNA has thymine instead of uracil. If DNA naturally contained uracil instead of thymine, then deamination of cytosine would generate a natural base, which the repair systems could not easily recognize.

The hazard of having deamination generate a naturally occurring base is illustrated by the problem caused by the presence of 5-methylcytosine. Vertebrate DNA frequently contains 5-methylcytosine in place of cytosine as a result of the action of methyltransferases. This modified base has a role in

transcriptional silencing (see Chapter 19). Deamination of 5-methylcytosine generates thymine (Fig. 10-7c), which obviously will not be recognized as an abnormal base and, following a round of DNA replication, can become fixed as a C-to-T transition. Indeed, methylated Cs are hot spots for spontaneous mutations in vertebrate DNA.

### DNA Is Damaged by Alkylation, Oxidation, and Radiation



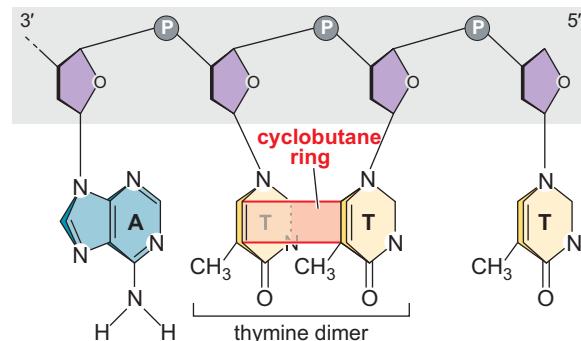
**FIGURE 10-8 G modification.** The figure shows specific sites on guanine that are vulnerable to damage by chemical treatment, such as alkylation or oxidation, and by radiation. The products of these modifications are often highly mutagenic.

DNA is vulnerable to damage from alkylation, oxidation, and radiation. In alkylation, methyl or ethyl groups are transferred to reactive sites on the bases and to phosphates in the DNA backbone. Alkylating chemicals include nitrosamines and the very potent laboratory mutagen *N*-methyl-*N*<sup>1-nitro-*N*-nitrosoguanidine. One of the most vulnerable sites of alkylation is the keto group at carbon atom 6 of guanine (Fig. 10-8). The product of this methylation, *O*<sup>6</sup>-methylguanine, often mispairs with thymine, resulting in the change of a G:C base pair into an A:T base pair when the damaged DNA is replicated.</sup>

DNA is also subject to attack from reactive oxygen species (e.g., O<sub>2</sub><sup>-</sup>, H<sub>2</sub>O<sub>2</sub>, and OH•). These potent oxidizing agents are generated by ionizing radiation and by chemical agents that generate free radicals. Oxidation of guanine, for example, generates 7,8-dihydro-8-oxoguanine, or oxoG. The oxoG adduct is highly mutagenic because it can base-pair with adenine as well as with cytosine. If it base-pairs with adenine during replication, it gives rise to a G:C to T:A transversion, which is one of the most common mutations found in human cancers. Thus, perhaps the carcinogenic effects of ionizing radiation and oxidizing agents are partly caused by free radicals that convert guanine to oxoG.

Yet another type of damage to bases is caused by ultraviolet light. Radiation with a wavelength of ~260 nm is strongly absorbed by the bases, one consequence of which is the photochemical fusion of two pyrimidines that occupy adjacent positions on the same polynucleotide chain. In the case of two thymines, the fusion is called a **thymine dimer** (Fig. 10-9), which comprises a **cyclobutane** ring generated by links between carbon atoms 5 and 6 of adjacent thymines. In the case of a thymine adjacent to a cytosine, the resulting fusion is a thymine–cytosine adduct in which the thymine is linked via its carbon atom 6 to the carbon atom 4 of cytosine. These linked bases are incapable of base pairing and cause the DNA polymerase to stop during replication. Assays exist to measure the amount of DNA damage such as thymine dimers and the effects of DNA damage on the ability of a cell to survive or maintain its genomic fidelity (see Box 10-3, Quantitation of DNA Damage and Its Effects on Cellular Survival and Mutagenesis).

Finally,  $\gamma$ -radiation and X-rays (ionizing radiation) are particularly hazardous because they cause double-strand breaks in the DNA, which are difficult to repair. If left unrepaired, double-strand breaks can be lethal to a cell.



**FIGURE 10-9 Thymine dimer.** Ultraviolet light induces the formation of a cyclobutane ring between adjacent thymines.

**► ADVANCED CONCEPTS****Box 10-3 Quantitation of DNA Damage and Its Effects on Cellular Survival and Mutagenesis**

To study DNA damage, repair, and mutagenesis, researchers use assays to measure DNA damage and its effects on cells. The ability to measure DNA damage seems challenging at first. How can one see what modifications to DNA are present inside a cell? Scientists have developed many techniques to accomplish this. In one assay, researchers use antibodies against a specific type of DNA damage, such as the thymine dimer. They measure the level of thymine dimers in a sample of isolated genomic DNA similarly to the use of antibodies in measuring protein levels in immunoblot analysis (see Chapter 7). Another assay, the comet or single-cell gel electrophoresis assay, detects the presence of single- and double-strand breaks as well as other types of damage in the DNA of individual cells through alterations in migration patterns during gel electrophoresis. Damaged DNA shows a comet-like appearance as observed by fluorescence microscopy. The ongoing development of new technologies promises to provide more accurate and specific methods for detecting DNA damage.

How do we measure the impact of DNA damage on cell viability? For single-celled organisms like bacteria or yeast, a survival assay can be as simple as plating cells on a solid medium and comparing the number of colonies (colony-forming units) that grow for treated versus untreated cells. The relationship of killing to a DNA-damage-inducing agent or condition (the killing curve) is determined by plotting the percent of surviving

cells at each dose of the DNA-damaging agent over a range of doses. A mutant in a pathway required to repair a specific type of damage produced by the treatment will show a lower percent survival than wild-type cells over the same range of treatments. A different approach is used for measuring the effect of DNA-damaging agents on the viability of mammalian cells. In this case, a fluorescent stain (a live–dead stain) is used that distinguishes between living and dead cells. The percent of cell survival as a function of treatment with a DNA-damaging agent is determined by counting cells that have or have not been stained by the dye using a fluorescence microscope.

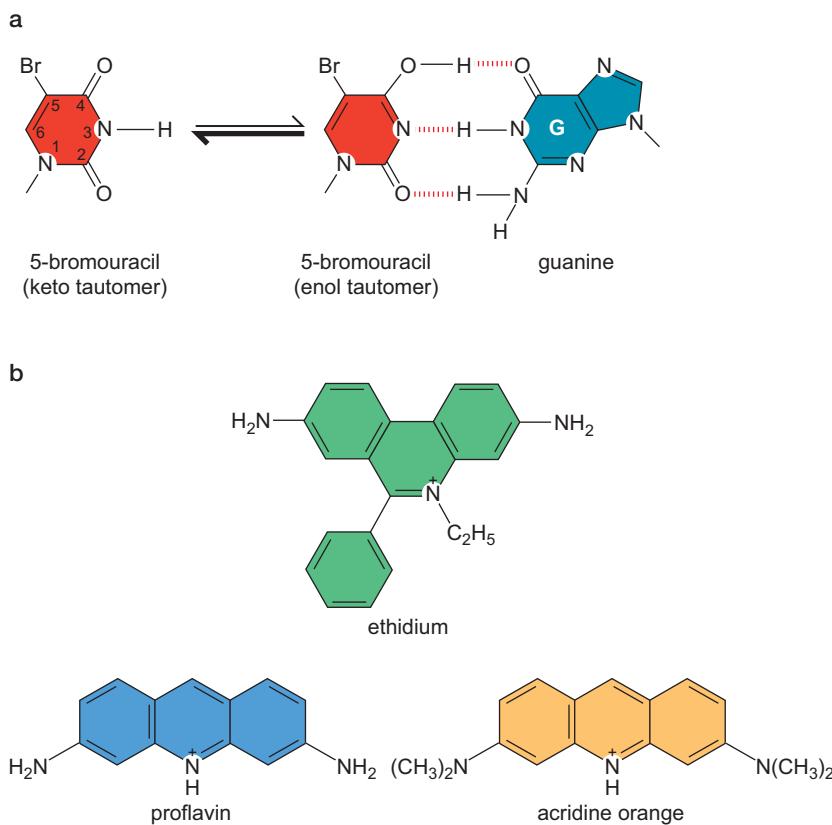
Quantitative studies of DNA-damaging agents or conditions involve measurements of mutagenesis as well as cell survival. Similar to the Ames test (see Box 10-2), mutagenesis assays may use measurements of the reversion of a specific mutation through the ability of mutant cells to grow on a solid medium lacking the required product or downstream product of the mutated gene. Mutagenesis assays may also involve forward mutations (from wild type to mutant) of a specific gene and a selective medium permitting only mutant cells to grow. Similar to survival assays, assays of mutagenesis involve treatment of cells with various doses of a DNA-damaging agent. The frequency of mutagenesis is determined from the percent of revertants or forward mutants as a function of the dose of the agent relative to cell survival.

Ionizing radiation can directly attack (ionize) the deoxyribose in the DNA backbone. Alternatively, this radiation can exert its effect indirectly by generating reactive oxygen species (described above), which, in turn, react with the deoxyribose subunits. Because cells require intact chromosomes to replicate their DNA, ionizing radiation is used therapeutically to kill rapidly proliferating cells in cancer treatment. Certain anticancer drugs, such as bleomycin, also cause breaks in DNA. Ionizing radiation and agents like bleomycin that cause DNA to break are said to be **clastogenic** (from the Greek *klastos*, which means “broken”).

### Mutations Are Also Caused by Base Analogs and Intercalating Agents

Mutations are also caused by compounds that substitute for normal bases (**base analogs**) or slip between the bases (**intercalating agents**) to cause errors in replication (Fig. 10-10). Base analogs are structurally similar to proper bases but differ in ways that make them treacherous to the cell. Thus, base analogs are similar enough to the proper bases to get taken up by cells, converted into nucleoside triphosphates, and incorporated into DNA during replication. But, because of the structural differences between these analogs and the proper bases, the analogs base-pair inaccurately, leading to frequent mistakes during the replication process. One of the most mutagenic base analogs is **5-bromouracil**, an analog of thymine. The presence of the bromo substituent allows the base to mispair with guanine via the enol tautomer (see Fig. 10-10a). As we saw in Chapter 4, the keto tautomer is strongly favored over the enol tautomer, but more so for thymine than for 5-bromouracil.

**FIGURE 10-10** Base analogs and intercalating agents that cause mutations in DNA. (a) Base analog of thymine, 5-bromouracil, can mispair with guanine. (b) Intercalating agents.



As we discussed for ethidium in Chapter 4, **intercalating agents** are flat molecules containing several polycyclic rings that bind to the equally flat purine or pyrimidine bases of DNA, just as the bases bind or stack with each other in the double helix. Intercalating agents, such as proflavin, acridine, and ethidium, cause the deletion or addition of a base pair or even a few base pairs. When such deletions or additions arise in a gene, they can have profound consequences on the translation of its mRNA because they shift the coding sequence out of its proper reading frame, as we shall see when we consider the genetic code in Chapter 16.

How do intercalating agents cause short insertions and deletions? One possibility in the case of insertions is that, by slipping between the bases in the template strand, these mutagens cause the DNA polymerase to insert an extra nucleotide opposite the intercalated molecule. (The intercalation of one of these structures approximately doubles the typical distance between two base pairs.) Conversely, in the case of deletions, the distortion to the template caused by the presence of an intercalated molecule might cause the polymerase to skip a nucleotide.

## REPAIR AND TOLERANCE OF DNA DAMAGE

As we have seen, damage to DNA can have two consequences. Some kinds of damage, such as thymine dimers or nicks and breaks in the DNA backbone, create impediments to replication or transcription. Other kinds of damage create altered bases that have no immediate structural consequence on replication but cause mispairing; these can result in a permanent alteration to the DNA sequence after replication. For example, the conversion of cytosine to

**TABLE 10-1** DNA Damage Repair and Tolerance Systems

Type	Damage	Enzyme
Mismatch repair	Replication errors	MutS, MutL, and MutH in <i>E. coli</i> ; MSH, MLH, and PMS in humans
Photoreactivation	Pyrimidine dimers	DNA photolyase
Base excision repair	Damaged base	DNA glycosylase
Nucleotide excision repair	Pyrimidine dimer; bulky adduct on base	UvrA, UvrB, UvrC, and UvrD in <i>E. coli</i> ; XPC, XPA, XPD, ERCCI-XPF, and XPG in humans
Double-strand break repair	Double-strand breaks	RecA and RecBCD in <i>E. coli</i>
Translesion DNA synthesis	Pyrimidine dimer, apurinic site, or bulky adduct on base	Y-family DNA polymerases, such as UmuC in <i>E. coli</i>

uracil by deamination creates a U:G mismatch, which, after a round of replication, becomes a C:G to T:A transition mutation on one daughter chromosome. These considerations explain why cells have evolved elaborate mechanisms to identify and repair DNA damage before it blocks replication or causes a mutation. Cells would not endure long without such mechanisms.

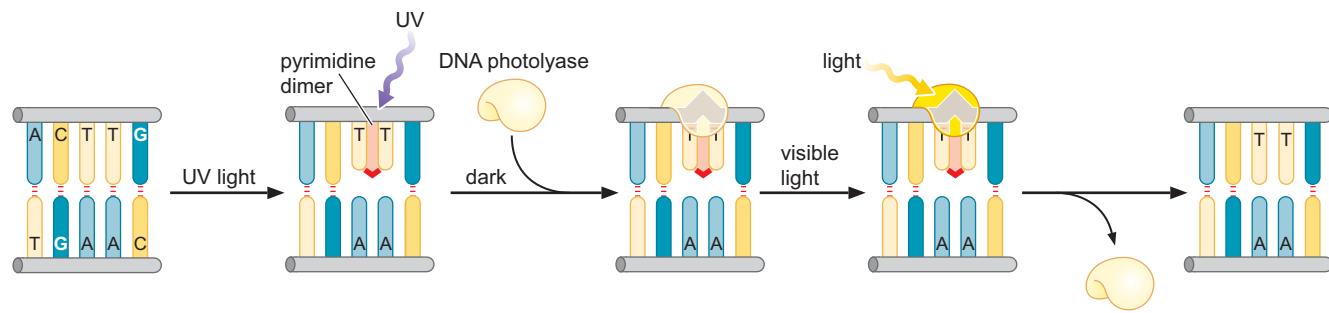
In this section, we consider the systems that repair DNA damage in addition to the mismatch repair system that corrects mismatches during replication (Table 10-1). In the most direct of these systems (representing true repair), a repair enzyme simply reverses (undoes) the damage. A more elaborate system is **excision repair**, in which the damaged nucleotide is not repaired but removed from the DNA. In excision repair systems, the other, undamaged, strand serves as a template for reincorporation of the correct nucleotide by DNA polymerase. As we shall see, two kinds of excision repair systems exist, one involving the removal of only the damaged nucleotide and the other involving the removal of a short stretch of single-stranded DNA that contains the lesion.

Yet more elaborate is **recombinational repair**, which is used when both strands are damaged, as when the DNA is broken. In such situations, one strand cannot serve as a template for the repair of the other. Hence, in recombinational repair (known as **double-strand break repair**), sequence information is retrieved from a second undamaged copy of the chromosome. Finally, when damaged bases block progression of a replicating DNA polymerase, a special **translesion polymerase** copies across the site of the damage in a manner that does not depend on base pairing between the template and newly synthesized DNA strands. This mechanism is an example of DNA damage tolerance, a system of last resort because translesion synthesis is inevitably error-prone (mutagenic).

### Direct Reversal of DNA Damage

An example of repair by simple reversal of damage is **photoreactivation**. Photoreactivation directly reverses the formation of pyrimidine dimers that result from ultraviolet irradiation. In photoreactivation, the enzyme DNA photolyase captures energy from light and uses it to break the covalent bonds linking adjacent pyrimidines (Fig. 10-11). In other words, the damaged bases are mended directly.

Another example of direct reversal is the removal of the methyl group from the methylated base  $O^6$ -methylguanine (see above). In this case, a methyltransferase removes the methyl group from the guanine residue by



**FIGURE 10-11** Photoreactivation. Ultraviolet irradiation causes formation of thymine dimers. Upon exposure to light, DNA photolyase breaks the ring formed between the dimers to restore the two thymine residues.

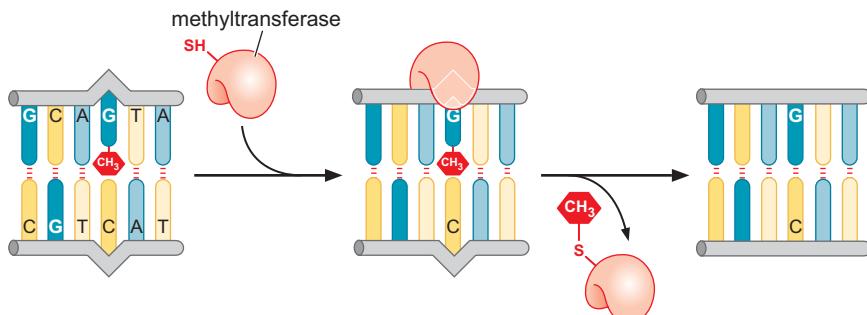
transferring it to one of its own cysteine residues (Fig. 10-12). This is costly to the cell because the methyltransferase is not catalytic; having once accepted a methyl group, it cannot be used again.

### Base Excision Repair Enzymes Remove Damaged Bases by a Base-Flipping Mechanism

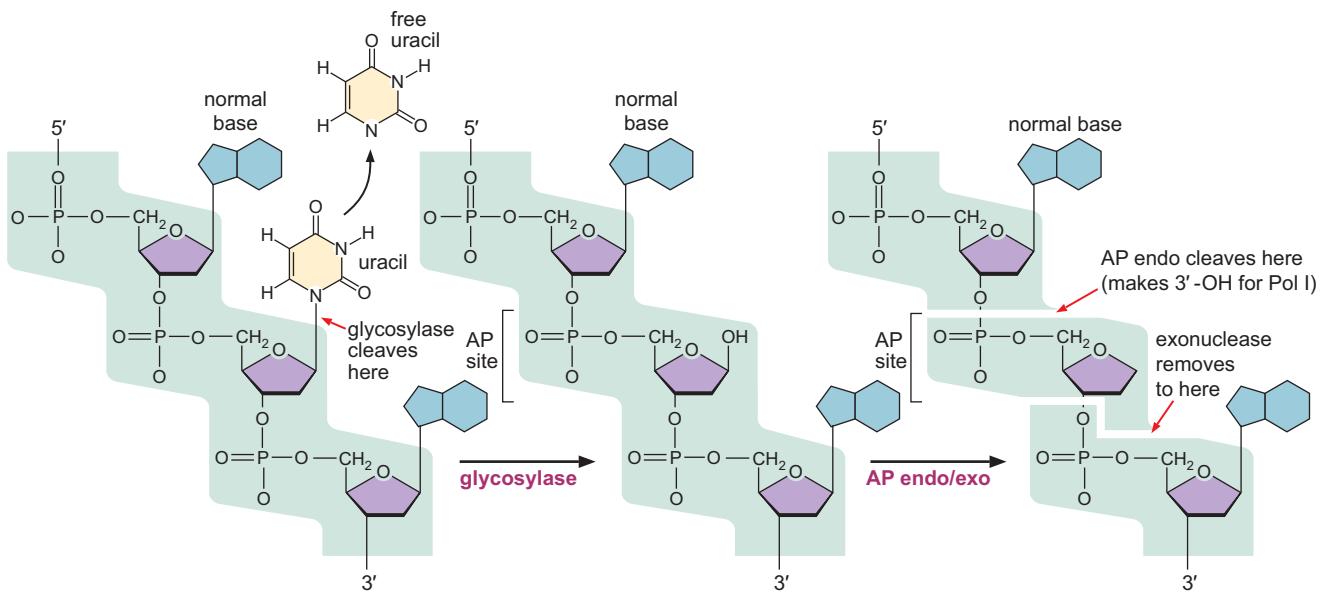
The most prevalent way in which DNA is cleansed of damaged bases is by repair systems that remove and replace the altered bases. The two principal repair systems are **base excision repair** and **nucleotide excision repair**. In base excision repair, an enzyme called a **glycosylase** recognizes and removes the damaged base by hydrolyzing the glycosidic bond (Fig. 10-13). The resulting abasic sugar is removed from the DNA backbone in a further endonucleolytic step. Endonucleolytic cleavage also removes apurinic and apyrimidinic sugars that arise by spontaneous hydrolysis. After the damaged nucleotide has been entirely removed from the backbone, a repair DNA polymerase and DNA ligase restore an intact strand using the undamaged strand as a template.

DNA glycosylases are lesion-specific and cells have multiple DNA glycosylases with different specificities. Thus, a specific glycosylase recognizes uracil (generated as a consequence of deamination of cytosine), and another is responsible for removing oxoG (generated as a consequence of oxidation of guanine). A total of 11 different DNA glycosylases have been identified in human cells.

Cleansing the genome of damaged bases is a formidable problem because each base is buried in the DNA helix. How do DNA glycosylases detect damaged bases while scanning the genome? Evidence indicates that these

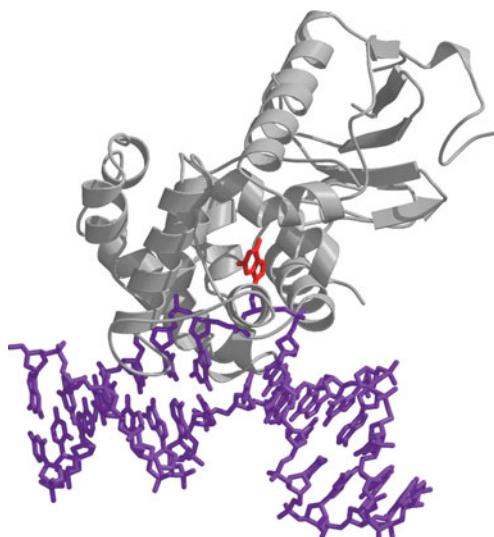


**FIGURE 10-12** Methyl group removal. Methyltransferase catalyzes the transfer of the methyl group on O<sup>6</sup>-methylguanine to a cysteine residue on the enzyme, thereby restoring the normal G in DNA.

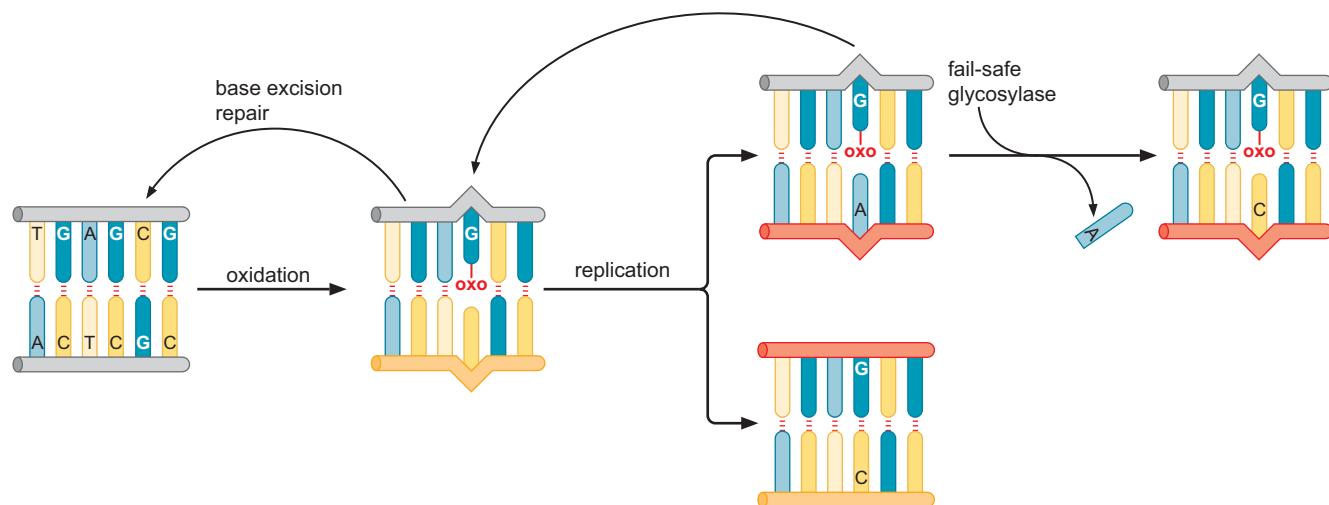


**FIGURE 10-13** Base excision pathway: the uracil glycosylase reaction. Uracil glycosylase hydrolyzes the glycosidic bond to release uracil from the DNA backbone to leave an AP site (apurinic or, in this case, apyrimidinic site). AP endonuclease cuts the DNA backbone at the 5' position of the AP site, leaving a 3'-OH; exonuclease cuts at the 3' position of the AP site, leaving a 5'-phosphate. The resulting gap is filled by DNA Pol I.

enzymes diffuse laterally along the minor groove of the DNA until a specific kind of lesion is detected. But how is the enzyme able to act on the base if it is buried in the helix? The answer to this riddle highlights the remarkable flexibility of DNA. X-ray crystallographic studies reveal that the damaged base is flipped out so that it projects away from the double helix, where it sits in the specificity pocket of the glycosylase (Fig. 10-14). Interestingly, the double helix is able to allow base flipping with only modest distortion to its structure, and hence the energetic cost of base flipping may not be great (see Chapter 4 and Fig. 4-8). Nevertheless, it is unlikely that glycosylases flip out every base to check for abnormalities as they diffuse along DNA.



**FIGURE 10-14** Structure of a DNA-glycosylase complex. The enzyme is shown in gray and the DNA in purple. The damaged base, in this case oxoG (shown in red), is flipped out of the helix and into the catalytic center of the enzyme. (Bruner S.D. et al. 2000. *Nature* **403**: 859–866.) Image prepared with MolScript, BobScript, and Raster3D.



**FIGURE 10-15** **oxoG:A repair.** Oxidation of guanine produces oxoG. The modified base can be repaired before replication by DNA glycosylase via the base excision pathway. If replication occurs before the oxoG is removed, resulting in the misincorporation of an A, then a fail-safe glycosylase can remove the A, allowing it to be replaced by a C. This provides a second opportunity for the DNA glycosylase to remove the modified base.

Thus, the mechanism by which these enzymes scan for damaged bases remains mysterious.

What if a damaged base is not removed by base excision before DNA replication? Does this inevitably mean that the lesion will cause a mutation? In the case of oxoG, which has the tendency to mispair with A, a fail-safe system exists (Fig. 10-15). A dedicated glycosylase recognizes oxoG:A base pairs generated by misincorporation of an A opposite an oxoG on the template strand. In this case, however, the glycosylase removes the A. Thus, the repair enzyme recognizes an A opposite an oxoG as a mutation and removes the undamaged but incorrect base.

Another example of a fail-safe system is a glycosylase that removes a T opposite a G. Such a T:G mismatch can arise, as we have seen, by spontaneous deamination of 5-methylcytosine, which occurs frequently in the DNA of vertebrates. Because both T and G are normal bases, how can the cell recognize which is the incorrect base? The glycosylase system assumes, so to speak, that the T in a T:G mismatch arose from deamination of 5-methylcytosine and selectively removes the T so that it can be replaced with a C.

### Nucleotide Excision Repair Enzymes Cleave Damaged DNA on Either Side of the Lesion

Unlike base excision repair, the nucleotide excision repair enzymes do not recognize any particular lesion. Rather, this system works by recognizing distortions to the shape of the double helix, such as those caused by a thymine dimer or by the presence of a bulky chemical adduct on a base. Such distortions trigger a chain of events that lead to the removal of a short single-strand segment (or patch) that includes the lesion. This removal creates a single-strand gap in the DNA, which is filled in by DNA polymerase using the undamaged strand as a template and thereby restoring the original nucleotide sequence.

Nucleotide excision repair in *E. coli* is largely accomplished by four proteins: UvrA, UvrB, UvrC, and UvrD (Fig. 10-16). A complex of two UvrA and two UvrB molecules scans the DNA, with the two UvrA subunits being

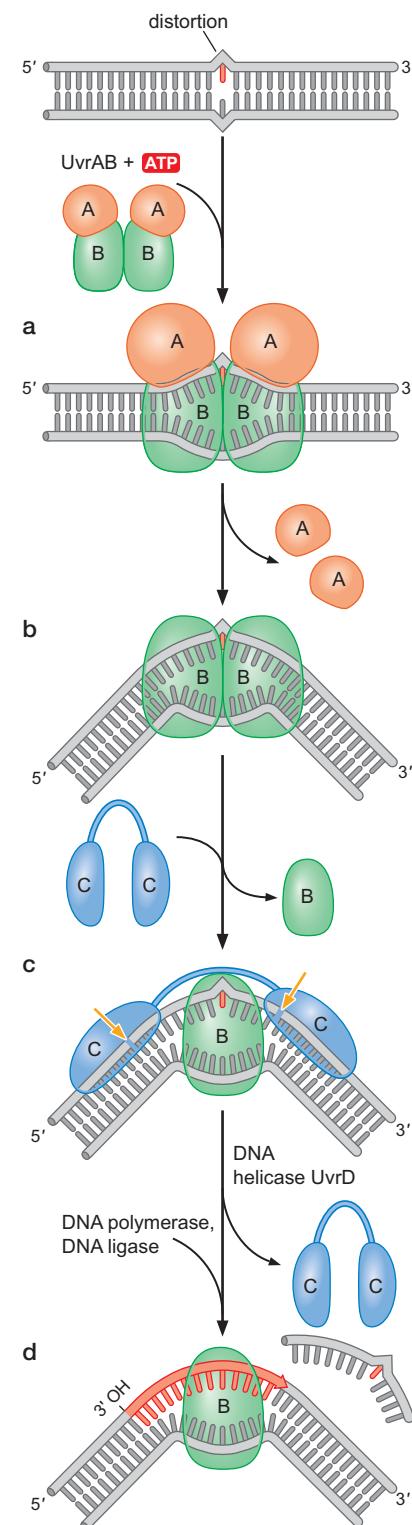
**FIGURE 10-16** Nucleotide excision repair pathway. (a) ATP hydrolysis promotes dimer formation by UvrA, which forms a complex with a dimer of UvrB. The UvrA and UvrB complex scans DNA to identify a distortion. (b) UvrA leaves the complex, and the remaining UvrB dimer melts DNA locally around the distortion. (c) UvrC forms a complex with UvrB and creates nicks 3' to the lesion and 5' to the lesion. (d) DNA helicase UvrD releases the single-strand fragment from the duplex, and DNA Pol I and ligase repair and seal the gap. (Adapted, with permission, from Zou Y. and Van Houten B. 1999. *EMBO J.* **18**: 4898, Fig. 7. © Macmillan.)

responsible for detecting distortions to the helix. Upon encountering a distortion, UvrA exits the complex, and the remaining dimer of UvrB melts the DNA to create a single-stranded bubble around the lesion. Next, the UvrB dimer recruits UvrC, and UvrC creates two incisions: one located 4 or 5 nucleotides 3' to the lesion and the other 8 nucleotides 5' to the lesion. These cleavages create a 12- to 13-residue-long DNA strand that contains the lesion. The lesion-containing strand is removed from the rest of the DNA by the action of the DNA helicase UvrD, resulting in a 12- to 13-nucleotide-long gap. Finally, DNA Pol I and DNA ligase fill in the gap.

The principle of nucleotide excision repair in higher cells is much the same as that in *E. coli*, but the machinery for detecting, excising, and repairing the damage is more complicated, involving 25 or more polypeptides. Among these is XPC, which is responsible for detecting distortions to the helix, a function attributed to UvrA in *E. coli*. As in *E. coli*, the DNA is opened to create a bubble around the lesion. Formation of the bubble involves the helicase activities of the proteins XPA and XPD (the equivalent to UvrB in *E. coli*) and the single-strand-binding protein RPA. The bubble creates cleavage sites 5' to the lesion for a nuclease known as ERCC1-XPF and 3' to the lesion for nuclease XPG (representing the function of UvrC). In higher cells, the resulting DNA strand is 24–32 nucleotides long. As in bacteria, the DNA strand is released to create a gap that is filled in by the action of DNA polymerase and ligase.

As their names imply, the UVR proteins are needed to mend damage from ultraviolet light; mutants of the *uvr* genes are sensitive to ultraviolet light and lack the capacity to remove thymine-thymine and thymine-cytosine adducts. In fact, these proteins broadly recognize and repair bulky adducts of many kinds. Nucleotide excision repair is important in humans, too. Humans can exhibit a genetic disease called xeroderma pigmentosum, which renders afflicted individuals highly sensitive to sunlight and results in skin lesions, including skin cancer (see Box 10-4, Linking Nucleotide Excision Repair and Translesion Synthesis to a Genetic Disorder in Humans).

Not only is nucleotide excision repair capable of mending damage throughout the genome, but it is also capable of rescuing RNA polymerase, the progression of which has been arrested by the presence of a lesion in the transcribed (template) strand of a gene. This phenomenon, known as **transcription-coupled repair**, involves recruitment to the stalled RNA polymerase of nucleotide excision repair proteins (Fig. 10-17). The significance of transcription-coupled repair is that it focuses repair enzymes on DNA (genes) being actively transcribed. In effect, RNA polymerase serves as another damage-sensing protein in the cell. Central to transcription-coupled repair in eukaryotes is the general transcription factor TFIIH. As we shall see in Chapter 13, TFIIH unwinds the DNA template during the initiation of transcription. Subunits of TFIIH include the DNA helix-opening proteins XPA and XPD discussed above. Thus, TFIIH is responsible for two separate functions: its strand-separating helicases melt the DNA around a lesion during nucleotide excision repair (including transcription-coupled repair) and also help to open the DNA template during the process of gene transcription. Systems for coupling repair to transcription also exist in prokaryotes.



## ► MEDICAL CONNECTIONS

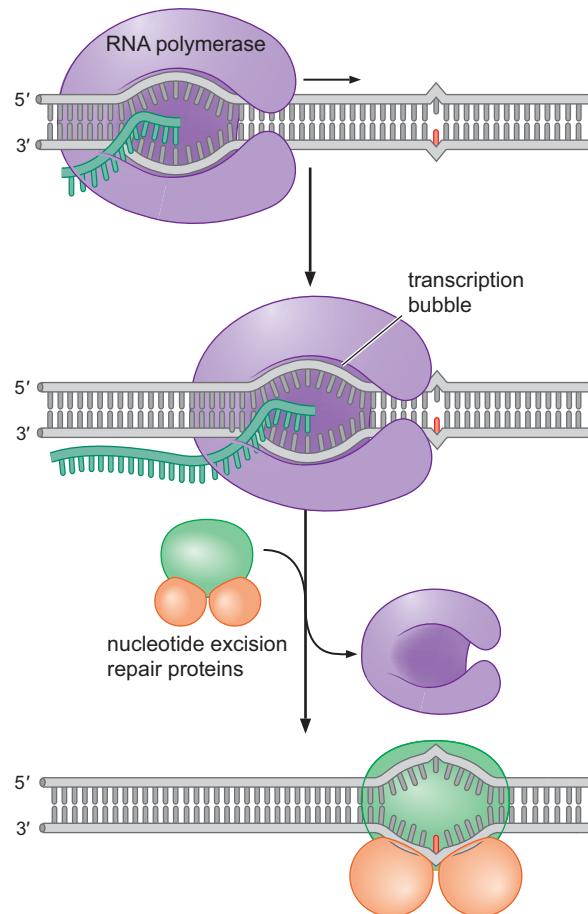
**Box 10-4** Linking Nucleotide Excision Repair and Translesion Synthesis to a Genetic Disorder in Humans

Humans can exhibit a genetic disorder called xeroderma pigmentosum (XP), an autosomal-recessive disease that renders afflicted individuals highly sensitive to sunlight and results in skin lesions, including skin cancer. Seven genes have been identified in which mutations give rise to XP. These genes specify proteins (such as XPA, XPC, XPD, XPF, and XPG; see text) in the human pathway for nucleotide excision repair (NER), underscoring the importance of NER in mending damage from ultraviolet light (UV). In addition to proteins involved in NER, a variant form of XP called XP-V is caused by a defect in the translesion DNA polymerase, Pol η (see later discussion on translesion polymerases). The gene encoding Pol η is sometimes called *XPV*. Individuals with XP-V have a milder form of XP.

What happens at the cellular level in individuals with XP? In the presence of defective NER, cells are limited in their ability to repair UV-induced DNA damage like thymine dimers. Following exposure to sunlight, the amount of DNA damage increases in the cells of individuals with XP, causing an increase in mutagenesis and cell death. Cells possessing a mutant Pol η are hindered in their ability to bypass thymine dimers during replication and must resort to using another translesion polymerase for bypass to avoid a block in replication. Because Pol η (but not other translesion polymerases) correctly inserts As across from a thymine dimer, the use of other translesion polymerases may increase the frequency of mutagenesis.

### Recombination Repairs DNA Breaks by Retrieving Sequence Information from Undamaged DNA

Excision repair uses the undamaged DNA strand as a template to replace a damaged segment of DNA on the other strand. How do cells repair double-strand breaks in DNA in which both strands of the duplex are broken? Double-strand break (DSB) repair pathways accomplish this. One recom-



**FIGURE 10-17** Transcription-coupled DNA repair. (Top) RNA polymerase transcribes DNA normally upstream of the lesion. (Middle) Upon encountering the lesion in DNA, RNA polymerase stalls and transcription stops. (Bottom) RNA polymerase recruits the nucleotide excision repair proteins to the site of the lesion, and then it either backs up or dissociates from the DNA to allow the repair proteins access to the lesion. (Adapted, with permission, from Zou Y. and Van Houten B. 1999. *EMBO J.* 18: 4898, Fig. 7. © Macmillan.)

bination-based pathway retrieves sequence information from the sister chromosome. Because of its central role in general homologous recombination as well as in repair, the recombination-based DSB repair pathway is an important topic in its own right, which we shall consider in detail in Chapter 11.

DNA recombination also helps to repair errors in DNA replication. Consider a replication fork that encounters a lesion in DNA (such as a thymine dimer) that has not been corrected by nucleotide excision repair. The DNA polymerase will sometimes stall attempting to replicate over the lesion. Although the template strand cannot be used, the sequence information can be retrieved from the other daughter molecule of the replication fork by recombination (see Chapter 11). Once this recombinational repair is complete, the nucleotide excision system has another opportunity to repair the thymine dimer. Indeed, mutants defective in recombination are known to be sensitive to ultraviolet light. Consider also the situation in which the replication fork encounters a nick in the DNA template. Passage of the fork over the nick will create a DNA break, repair of which can only be accomplished by DSB repair pathways. Although we generally consider recombination as an evolutionary device to explore new combinations of sequences, it may be that its original function was to repair damage in DNA.

### DSBs in DNA Are Also Repaired by Direct Joining of Broken Ends

A DSB is the most cytotoxic of all kinds of DNA damage. If left unmended, a DNA break can have multiple deleterious consequences, such as blocking replication and causing chromosome loss, which result in cell death or neoplastic transformation. Cells typically have multiple overlapping pathways for coping with DNA damage. It should therefore come as no surprise that cells do not rely on recombination alone for mending DSBs. As we have seen and will consider in further detail in Chapter 11, the recombination-based DSB repair pathway relies on DNA sequence information in a sister chromosome to repair broken DNA molecules. This is an effective strategy because the sister chromosome provides a template for the precise restoration of the original sequence across the site of the break. In yeast cells, recombination-based DSB repair is the principal pathway by which breaks are mended. But what happens early in the cell cycle before two sister chromosomes have been generated by DNA replication? If a still-unreplicated chromosome suffers a break, then no sister chromosome is present to serve as a template in the recombination-based DSB repair pathway. Under such conditions, an alternative DSB repair system comes into play known as **non-homologous end joining**, or **NHEJ**. NHEJ is a backup system in yeast, but in higher cells it is the principal pathway by which breaks are repaired (see Box 10-5, Nonhomologous End Joining).

The machinery for performing NHEJ protects and processes the broken ends and then joins them together, as we shall explain. Because sequence information is lost from the broken ends, the original sequence across the break is not faithfully restored during NHEJ. Thus, NHEJ is mutagenic. Of course, the mutagenic consequences of NHEJ-mediated DNA end joining are far less hazardous to the cell than are the consequences leaving broken DNA unrepaired!

What is the mechanism that joins DNA ends together in NHEJ? As its name implies, NHEJ does not involve extensive stretches of homologous sequences. Instead, the two ends of the broken DNA are joined to each other by misalignment between single strands protruding from the broken ends. This misalignment is believed to occur by pairing between tiny stretches

► MEDICAL CONNECTIONS

**Box 10-5 Nonhomologous End Joining**

NHEJ can repair DSBs arising from exposure to exogenous agents, such as ionizing radiation, and from cell-intrinsic insults, such as failures in DNA replication. Remarkably, NHEJ is also used in the entirely normal cell-intrinsic process of adaptive immunity. The immune system produces an enormously diverse group of antibody molecules, which are composed of so-called light and heavy polypeptide chains. The light and heavy chains are generated by a recombinational process that involves the joining, in a bewildering number of combinations, of a large repertoire of protein-coding DNA elements known as V and J segments (and, in the case of the antibody heavy chain, a D segment) for different parts of the polypeptides. As we discuss in Chapter 12, this process is known as **V(D)J recombination**. V(D)J recombination is initiated by the introduction of breaks in the DNA by a process that is specific to lymphocytes and involves an enzyme composed of the proteins RAG1 and RAG2.

Once the breaks are created, the NHEJ pathway, which is not lymphocyte-specific, joins the ends together. In this case, however, the ends of the protein-coding segments are not rejoined to their original partners. Rather, the ends are joined to new partners to create the composite coding sequences for the heavy and light chains. NHEJ also participates in a second example of V(D)J recombination that governs the production of an additional category of immunological polypeptides called T-cell receptors, as discussed in Chapter 12.

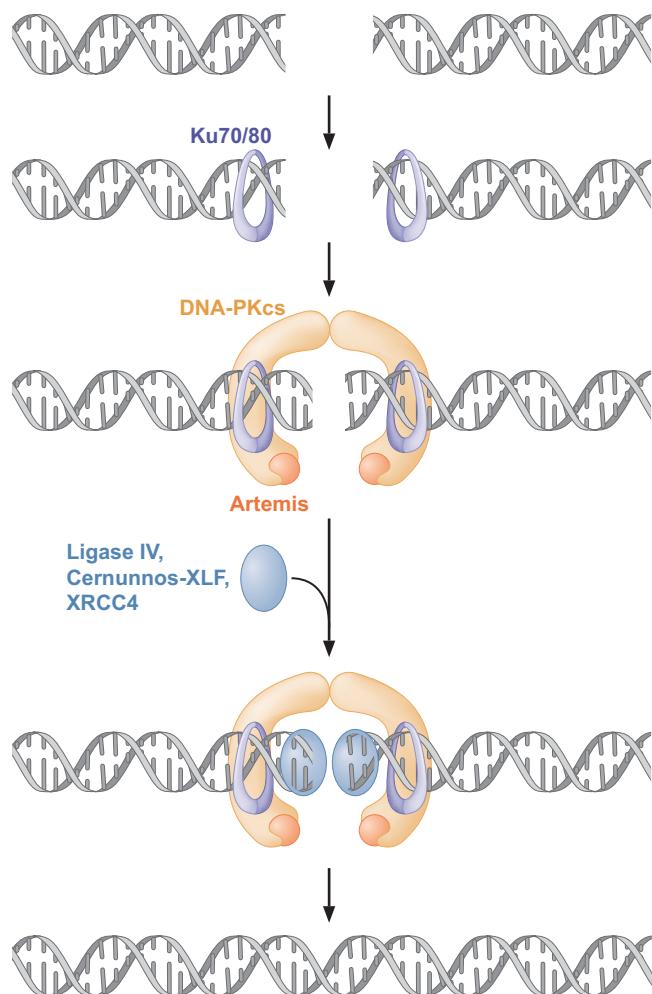
Underscoring the importance of NHEJ in human biology are rare inherited syndromes that are characterized by hypersensitivity to ionizing radiation and DNA-damaging agents and by immunodeficiency, which is attributed to defective V(D)J recombination. Revealingly, patients showing this syndrome harbor mutations in the genes for the Artemis, Ligase IV, or Cernunnos-XLF members of the NHEJ pathway.

(as short as 1 bp) of complementary bases (serendipitous microhomologies). Nucleases remove single-strand tails, and DNA polymerase fills in the gaps.

A growing number of proteins that mediate NHEJ have been identified. To date, seven components of the NHEJ pathway have been discovered in mammalian cells. These proteins, which have formidable-sounding names, are Ku70, Ku80, DNA-PKcs, Artemis, XRCC4, Cernunnos-XLF, and DNA ligase IV (Fig. 10-18). Ku70 and Ku80 are the most fundamental components of NHEJ. They constitute a heterodimer that binds to the DNA ends and recruits DNA-PKcs, which is a protein kinase. DNA-PKcs, in turn, forms a complex with Artemis. Artemis is both a 5'-to-3' exonuclease and a latent endonuclease that is activated by phosphorylation by DNA-PKcs. These nucleolytic activities process the broken ends and prepare them for ligation. Ligase IV performs ligation in a complex with XRCC4 and Cernunnos-XLF.

NHEJ is ubiquitous in eukaryotic organisms, but it occurs, albeit less frequently, in bacteria. Nevertheless, a fascinating specialized example has been discovered in spores of the bacterium *Bacillus subtilis*. *B. subtilis* produces a Ku-like protein and a DNA ligase when it sporulates and packages the proteins into the mature spore. Ku and the DNA ligase, representing a simple, two-protein NHEJ system, repair DNA breaks when the spore germinates. Mutant spores lacking these proteins are highly susceptible to dry heat, a condition that is known to cause breaks in DNA. Upon germination, heated mutant spores are unable to resume growth because they are unable to rejoin the heat-induced breaks.

That germinating spores rely on NHEJ, rather than on the recombination-based DSB repair pathway, to mend breaks makes good sense. Spores have only one chromosome. Therefore, they cannot rely on a sister chromosome to use as a template for repair of the break. Interestingly, the spore chromosome is tightly coiled into a remarkable doughnut-like structure that could hold the ends of breaks in DNA in close proximity to each other. This close juxtaposition could facilitate correct rejoining of ends even if the chromosome has sustained multiple breaks. Spores of *B. subtilis* and related bacteria are able to



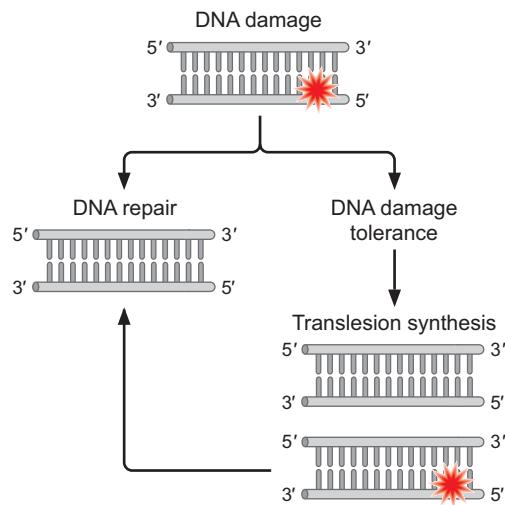
**FIGURE 10-18 Mammalian pathway for NHEJ.** A heterodimer of Ku70 and Ku80 binds to broken DNA ends and recruits the protein kinase DNA-PKcs. DNA-PKcs, in turn, recruits Artemis, an enzyme having exonuclease and endonuclease activities, which processes the broken ends. Finally, a complex of Ligase IV with XRCC4 and Cernunnos-XLF joins the broken ends to each other. (Adapted, with permission, from Sekiguchi J.M. and Ferguson D.O. 2006. *Cell* 124: 260–262. © Elsevier.)

survive extremes of the environment far more effectively than any other kind of dormant cell. NHEJ is part of the basis for this extraordinary robustness.

### Translesion DNA Synthesis Enables Replication to Proceed across DNA Damage

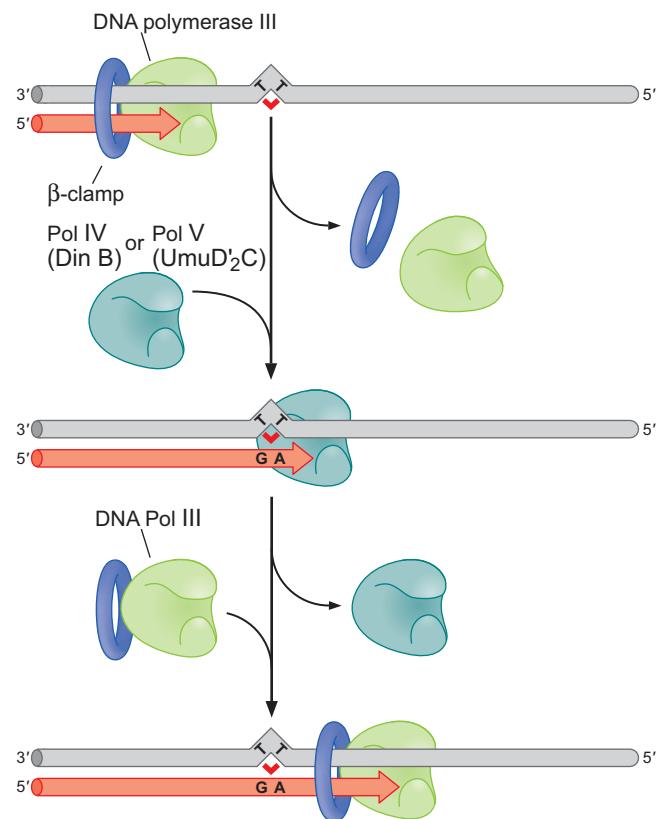
In many of the examples we have considered so far, damage to the DNA is mended by excision, followed by resynthesis using an undamaged template. But such repair systems do not operate with complete efficiency, and sometimes a replicating DNA polymerase encounters a lesion, such as a pyrimidine dimer or an apurinic site, that has not been repaired. Because such lesions are obstacles to progression of the DNA polymerase, the replication machinery must attempt to copy across the lesion or be forced to cease replicating. Even if cells cannot repair these lesions, there is a fail-safe mechanism that allows the replication machinery to bypass these sites of damage or tolerate the DNA damage. One mechanism of DNA damage tolerance is **translesion synthesis**. Although this mechanism is, as we shall see, highly error-prone and thus likely to introduce mutations, translesion synthesis spares the cell the worse fate of an incompletely replicated chromosome. A key feature of DNA damage tolerance is that the DNA lesion remains in the genome. DNA repair pathways can subsequently correct the lesion (Fig. 10-19).

**FIGURE 10-19** Cellular defenses against DNA damage. Cells use DNA repair pathways to restore DNA to its undamaged state. If DNA damage is present when the genome is being replicated, the cell must use DNA damage tolerance to avoid a block in replication and a potentially lethal double-strand break. Translesion synthesis replicates across the DNA lesion, but the lesion remains in the genome until a DNA repair pathway can subsequently correct the damage.

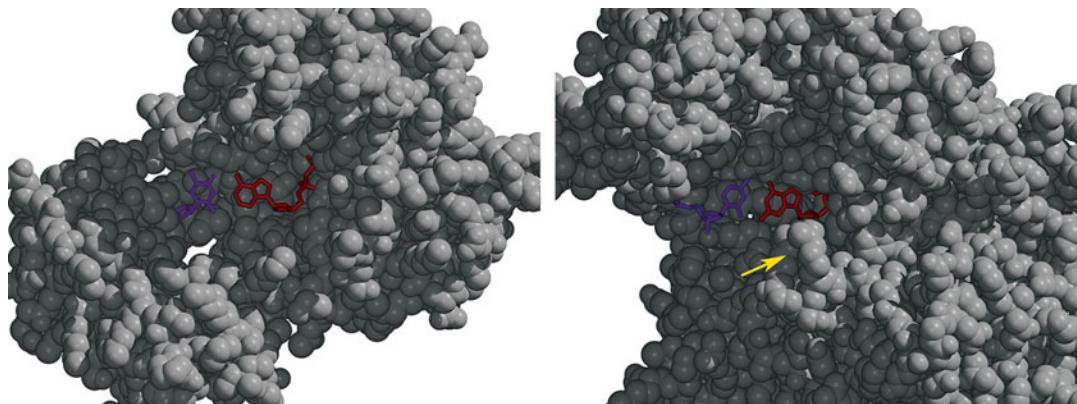


Translesion synthesis is catalyzed by a specialized class of DNA polymerases that synthesize DNA directly across the site of the damage (Fig. 10-20). In *E. coli*, DNA Pol IV (DinB) or DNA Pol V (a complex of the proteins UmuC and UmuD') performs translesion synthesis. DinB and UmuC are members of a distinct family of DNA polymerases found in many organisms known as the Y family of DNA polymerases (Fig. 10-21; see Box 10-6, The Y Family of DNA Polymerases). There are five translesion polymerases known in humans, four of which belong to the Y family.

An important feature of these polymerases is that although they are template-dependent, they incorporate nucleotides in a manner that is



**FIGURE 10-20** Translesion DNA synthesis. Upon encountering a lesion in the template during replication, DNA Pol III with its sliding clamp dissociates from the DNA and is replaced by the translesion DNA polymerase, which extends DNA synthesis across the thymine dimer on the template (upper) strand. The translesion polymerase is then replaced by the DNA polymerase III. (Adapted, with permission, from Woodgate R. 1999. *Genes Dev.* 13: 2191–2195, Fig. 1. © Cold Spring Harbor Laboratory Press.)



**FIGURE 10-21** Crystal structure of a translesion polymerase. The structures shown here represent two different types of DNA polymerases. The structure on the left is a Y-family (lesion bypass) polymerase; that on the right is a typical DNA polymerase from bacteriophage T7. Notice the more open structure around the active site in the Y-polymerase structure, and the absence of the protein region that closes the channel (indicated by the yellow arrow). The incoming nucleotides are in red, and template nucleotides are in blue. (Y polymerase, Ling H. et al. 2001. *Cell* **107**: 91. PDB Code: 1JX4. T7 polymerase, Doublie S. et al. 1998. *Nature* **391**: 251. PDB Code: 1T7P.) Images prepared with MolScript, BobScript, and Raster3D.

independent of base pairing. This explains how the enzymes can synthesize DNA over a lesion on the template strand. But, because the enzyme is not “reading” sequence information from the template, translesion synthesis is often highly error-prone. Consider the case of an apurinic or apyrimidinic site in which the lesion contains no base-specific information. The translesion polymerase synthesizes across the lesion by inserting nucleotides in a manner that is not guided by base pairing. Nonetheless, the nucleotide incorporated may not be random—some translesion polymerases incorporate specific nucleotides. For example, a human member of the Y family of translesion polymerases (DNA Pol  $\eta$ ) correctly inserts two A residues opposite a thymine dimer. Structural studies show that the active site of DNA Pol  $\eta$  is better at accommodating a thymine dimer than is the active site of another translesion DNA polymerase (DNA Pol  $\kappa$ ) (Fig. 10-22).

Because of its high error rate, translesion synthesis (like NHEJ) can be considered a system of last resort. It enables the cell to survive what might otherwise be a catastrophic block to replication, but the price that is paid is a higher level of **mutagenesis**. Mutagenesis is the process by which mutations are introduced and remain in the genome. For this reason, translesion DNA polymerases must be tightly regulated. In *E. coli*, the translesion polymerases are not present under normal circumstances. Rather, their synthesis is induced only in response to DNA damage. Thus, the genes encoding the translesion polymerases are expressed as part of a pathway known as the **SOS response**. Damage leads to the proteolytic destruction of a transcriptional repressor (the LexA repressor) that controls expression of genes involved in the SOS response, including those for DinB, UmuC, and UmuD, the inactive precursor for UmuD'. Interestingly, the same pathway is also responsible for the proteolytic conversion of UmuD to UmuD'. Cleavage of both LexA and UmuD is stimulated by a protein called RecA, which is activated by single-stranded DNA resulting from DNA damage. RecA is a dual-function protein that is also involved in DNA recombination, as we shall see in Chapter 11.

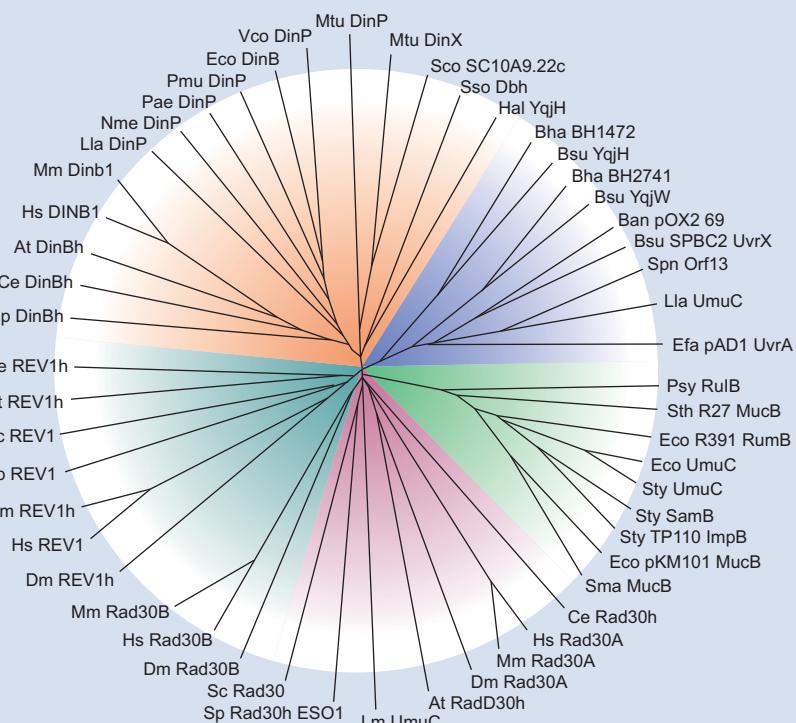
We next address the question of how a translesion polymerase gains access to the stalled replication machinery at the site of DNA damage. In

## ► ADVANCED CONCEPTS

**Box 10-6** The Y Family of DNA Polymerases

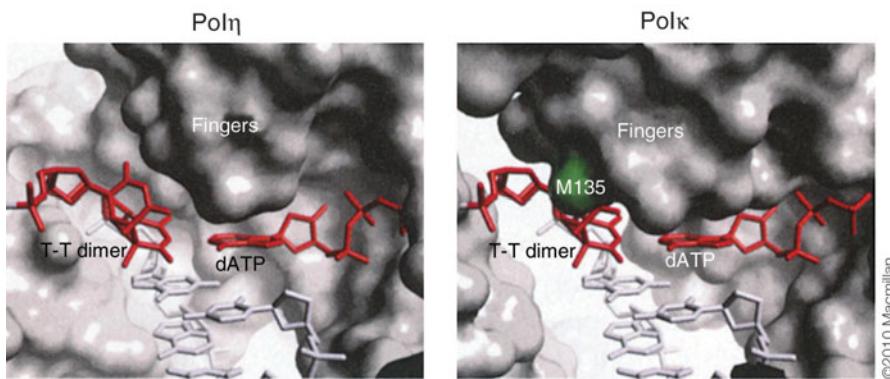
DNA polymerases can be grouped into families, shown in various colors in Box 10-6 Figure 1, based on their amino acid sequence similarities to each other. Recently, UmuC and certain other translesion DNA polymerases are founding members of a large and distinct family of DNA polymerases known as the Y family, which are found in all three domains of life: Bacteria, Archaea, and Eukaryota. Members of the Y family

of DNA polymerases characteristically perform DNA synthesis with low fidelity on undamaged DNA templates but have the capacity to bypass lesions in DNA that block replication by members of the other families of DNA polymerases. Box 10-6 Figure 1 shows a phylogenetic tree for the Y family of translesion DNA polymerases.

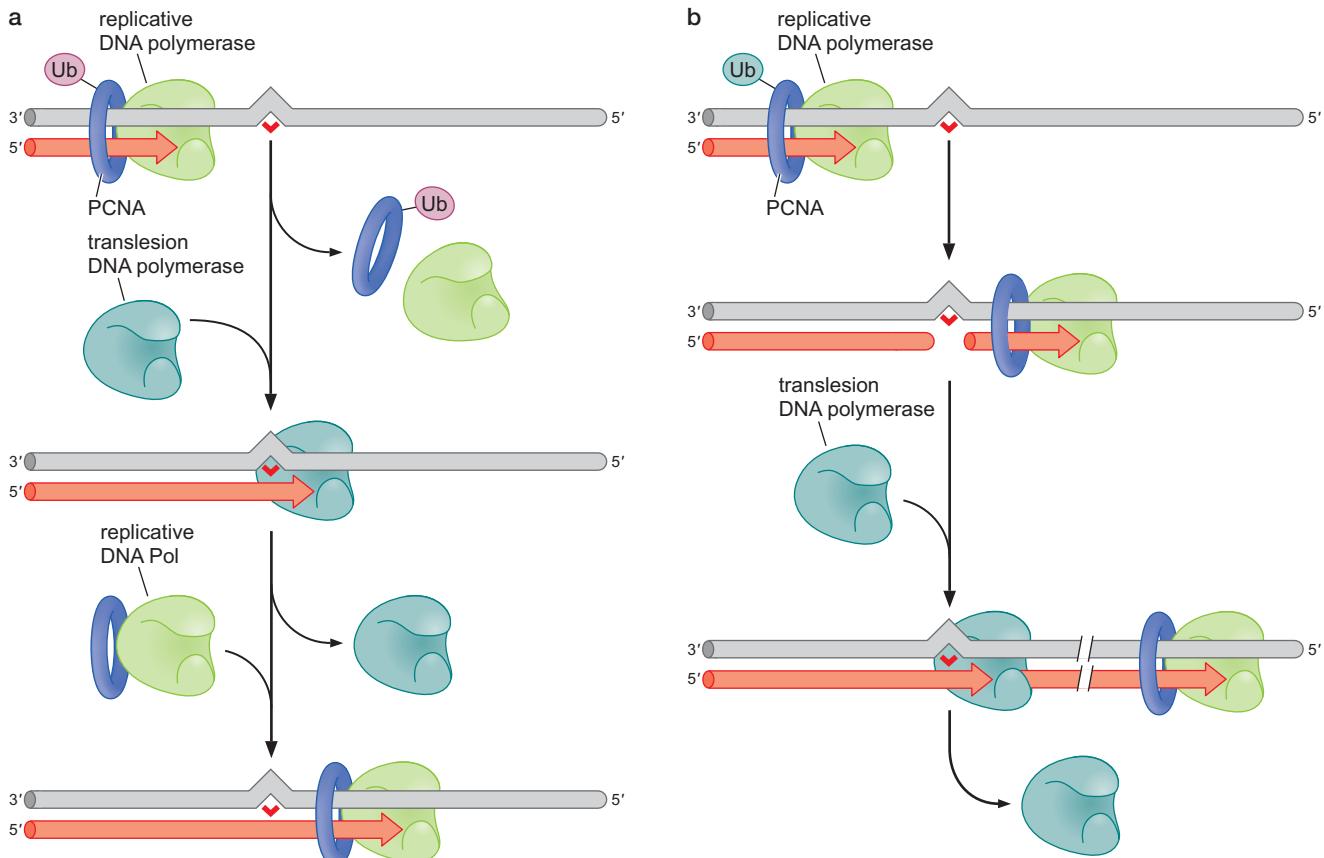


**BOX 10-6 FIGURE 1** The phylogenetic tree of the Y family of DNA polymerases. (Adapted, with permission, from Ohmori H. et al. 2001. *Mol. Cell* 8: 7, Fig 1. © Elsevier.)

mammalian cells, entry into the translesion synthesis pathway is triggered by chemical modification of the sliding clamp. As we saw in Chapter 9, the sliding clamp, which is known as PCNA in eukaryotes, anchors the replicative polymerase to the DNA template. The chemical modification is the covalent attachment to the sliding clamp of a peptide known as ubiquitin in a process known as **ubiquitination**. Ubiquitination is widely used in eukaryotic cells to mark proteins for various processes, such as degradation. Its use in triggering translesion synthesis adds to the growing list of cellular processes that are governed by tagging proteins with the ubiquitin peptide. Once ubiquitinated, the sliding clamp recruits a translesion polymerase, which contains domains that recognize and bind to ubiquitin. The translesion polymerase, in turn, somehow displaces the replicative polymerase from the 3' end of the growing strand and extends it across the site of the damage (Fig. 10-23a). Ubiquitination of the sliding clamp is therefore a distress signal that recruits a translesion polymerase to rescue a replication machine that is stalled at a site of DNA damage. In addition to a polymerase switching mechanism, data support that translesion synthesis also uses a mechanism of gap filling. Following replication, a gap results from the



**FIGURE 10-22 Translesion polymerases favor particular kinds of damage.** (Left) A thymine dimer fits well in the active site of DNA Pol  $\eta$ , allowing the polymerase to correctly insert two As across from the dimer. (Right) A superimposed image of how a thymine dimer does not fit in the active site of another translesion DNA polymerase (DNA Pol  $\kappa$ ) that is better suited for the repair of other DNA lesions. (Reprinted from Silverstein T.D. et al. 2010. *Nature* 465: 1039–1043, Fig. 4b, p. 1042. © Macmillan.)



**FIGURE 10-23 Alternative models for translesion synthesis.** Two models explain the mechanism of translesion synthesis, each likely to be true under particular circumstances. (a) In the polymerase-switching model, the processive, replicative DNA polymerase synthesizes DNA until the polymerase encounters a DNA lesion. The DNA polymerase stalls and is displaced by one or more translesion synthesis polymerases, which step in to replicate across and shortly beyond the lesion. Following this replication bypass, the replicative polymerase returns to displace the translesion polymerase and resume processive replication. (b) In the gap-filling model, the processive, replicative DNA polymerase synthesizes DNA until the polymerase encounters a DNA lesion. Instead of stalling at the DNA lesion, the polymerase skips ahead, continuing DNA synthesis downstream from the DNA lesion and leaving behind a gap. Subsequently, one or more translesion synthesis polymerases synthesize across the lesion to fill in the gap. (Adapted from Waters L.S. et al. 2009. *Microbiol. Mol. Biol. Rev.* 73: 134–154, Fig. 4, p. 146.)

replicative DNA polymerase skipping over the DNA lesion and continuing replication through repriming events or by starting a new Okazaki fragment (Fig. 10-23b).

Finally, several fascinating, but as yet unanswered, questions remain. How exactly does the translesion enzyme replace the normal replicative polymerase in the DNA replication complex? Once DNA synthesis is extended across the lesion, how does the normal replicative polymerase switch back to and replace the translesion enzyme at the replication fork? Translesion polymerases have low processivity, thus perhaps they simply dissociate from the template shortly after copying across a lesion. Nonetheless, this explanation still leaves us with the challenge of understanding how the normal processive enzyme is able to reenter the replication machinery.

## SUMMARY

---

Organisms can survive only if their DNA is replicated faithfully and is protected from chemical and physical damage that would change its coding properties. The limits of accurate replication and repair of damage are revealed by the natural mutation rate. Thus, an average nucleotide is likely to be changed by mistake only about once every  $10^9$  times it is replicated, although error rates for individual bases can vary over a 10,000-fold range. Much of the accuracy of replication is inherent in the way DNA polymerase copies a template. The initial selection of the correct base is guided by complementary pairing. Accuracy is increased by the proofreading activity of DNA polymerase. Finally, in mismatch repair, the newly synthesized DNA strand is scanned by an enzyme that initiates replacement of DNA containing incorrectly paired bases. Despite these safeguards, mistakes of all types occur: base substitutions, small and large additions and deletions, and gross rearrangements of DNA sequences.

Cells have a large repertoire of enzymes devoted to repairing DNA damage that would otherwise be lethal or would alter DNA so as to engender damaging mutations. Some enzymes directly reverse DNA damage, such as photolyases, which reverse pyrimidine dimer formation. A more versatile strategy is excision repair, in which a damaged segment is removed and replaced through new DNA synthesis for which the undamaged strand serves as a template. In base excision repair, DNA glycosylases and endonucleases remove only

the damaged nucleotide, whereas in nucleotide excision repair, a short patch of single-stranded DNA containing the lesion is removed. In *E. coli*, excision repair is initiated by the UvrABC endonuclease, which creates a bubble over the site of the damage and cuts out a 12-nucleotide segment of the DNA strand that includes the lesion. Higher cells perform nucleotide excision repair in a similar manner, but a much larger number of proteins are involved, and the excised, single-stranded DNA is 24–32 nucleotides long.

The most hazardous kind of damage is a DNA break. Recombinational DSB repair is a pathway that mends breaks in which the sequence across the break is copied from a different but homologous duplex. If no template for repair synthesis is available, breaks in DNA are mended by NHEJ, which rejoins the ends but in an error-prone manner. If the cell needs to replicate damaged DNA, translesion synthesis allows the cell to tolerate the lesion. Translesion synthesis enables replication to continue across damage that blocks the progression of a replicating DNA polymerase. Translation synthesis is primarily mediated by a distinct and widespread family of DNA polymerases that are able to perform DNA synthesis in a manner that, although not always accurate, does not depend on base pairing.

Mutagenesis and its repair are of concern to us because they permanently affect the genes that organisms inherit and because cancer is often caused by mutations in somatic cells.

## BIBLIOGRAPHY

---

### Books

- Friedberg E.C., Walker G.C., Siede W., Wood R.D., Schultz R.A., and Ellenberger T. 2005. *DNA repair and mutagenesis*. ASM Press, Washington, DC.
- Kornberg A. and Baker T.A. 1992. *DNA replication*, 2nd ed. W.H. Freeman, New York.

### Replication Errors and Their Repair

- Kunkel T.A. and Erie D.A. 2005. DNA mismatch repair. *Annu. Rev. Biochem.* **76**: 681–710.

### Repair of DNA Damage

- Bridges B.A. 1999. DNA repair: Polymerases for passing lesions. *Curr. Biol.* **9**: R475–R477.
- Citterio E., Vermeulen W., and Hoeijmakers J.H. 2000. Transcriptional healing. *Cell* **101**: 447–450.
- Daley J.M., Palmbos P.L., Wu D., and Wilson T.E. 2005. Nonhomologous end joining in yeast. *Annu. Rev. Genet.* **39**: 431–451.
- de Laat W.L., Jaspers N.G., and Hoeijmakers J.H. 1999. Molecular mechanism of excision nucleotide repair. *Genes Dev.* **13**: 768–785.
- Drapkin R., Reardon J.T., Ansari A., Huang J.C., Zawel L., Ahn K., Sancar A., and Reinberg D. 1994. Dual role of TFIIH in DNA

- excision repair and in transcription by RNA polymerase II. *Nature* **368**: 769–772.
- Kleczkowska H.E., Marra G., Lettieri T., and Jiricny J. 2001. hMSH3 and hMSH6 interact with PCNA and colocalize with it to replication foci. *Genes Dev.* **15**: 724–736.
- Lehmann A.R., McGibbon D., and Orphanet M.S. 2011. Xeroderma pigmentosum. *J Rare Dis* **6**: 70.
- Sekiguchi J.M. and Ferguson D.O. 2006. DNA double-strand break repair: A relentless hunt uncovers new prey. *Cell* **124**: 260–262.
- Silverstein T.D., Johnson R.E., Jain R., Prakash L., Prakash S., and Aggarwal A.K. 2010. Structural basis for the suppression of skin cancers by DNA polymerase η. *Nature* **465**: 1039–1043.
- Verhoeven E.E., Wyman C., Moolenaar G.F., and Goosen N. 2002. The presence of two UvrB subunits in the UvrAB complex ensures damage detection in both DNA strands. *EMBO J.* **21**: 4196–4205.
- Waters L.S., Minesinger B.K., Wiltrot M.E., D’Souza S., Woodruff R.V., and Walker G.C. 2009. Eukaryotic translesion polymerases and their roles and regulation in DNA damage tolerance. *Microbiol. Mol. Biol. Rev.* **73**: 134–154.
- Webster M.P., Jukes R., Zamfir V.S., Kay C.W., Bagnérés C., and Barrett T. 2012. Crystal structure of the UvrB dimer: Insights into the nature and functioning of the UvrAB damage engagement and UvrB–DNA complexes. *Nucleic Acids Res.* doi: 10.1093/nar/gks633.

## QUESTIONS

## MasteringBiology®

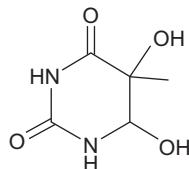
For instructor-assigned tutorials and problems, go to *MasteringBiology*.

For answers to even-numbered questions, see Appendix 2: Answers.

**Question 1.** DNA polymerase mistakenly inserts a C across from a T during replication. Assuming that proofreading and mismatch repair do not correct the mismatch, is the resulting mutation a transition or transversion after the next round of replication? Explain your choice.

**Question 2.** Explain why the deamination of 5-methylcytosine leads to hot spots for spontaneous mutations more than the deamination of cytosine in DNA does.

**Question 3.** Given the structure of the damaged base below, circle the modification(s) present relative to the base normally found in DNA. Name the process that produces this type of modification. Name the DNA repair pathway that you expect would recognize and correct this type of DNA damage.



**Question 4.** The following terms describe the general steps of a DNA repair pathway. Place the steps in the correct order. Name the protein(s) that complete each of the steps in *E. coli* for the mismatch repair, base excision repair, and nucleotide excision repair pathways.

### Ligation, DNA synthesis, Recognition, Excision

#### Question 5.

- A. Calculate the number of mismatches that could occur in one human cell during one round of replication. Assume the size of the human genome is 3.2 billion base pairs.
- B. Calculate the number of mismatches that could occur in one human cell during one round of replication in the *absence of mismatch repair*.

**Question 6.** Given a loss-of-function mutant for *dam* (the gene encoding the Dam methylase) in *E. coli*, predict the phenotype one would observe with respect to spontaneous mutagenesis. Briefly explain your answer.

**Question 7.** Describe a possible advantage and disadvantage of repairing 3-methyladenine through base excision repair relative to repairing *O*<sup>6</sup>-methylguanine through direct reversal by a methyltransferase.

**Question 8.** Exposure of DNA to the chemotherapy drug, cisplatin, causes the formation of intrastrand cross-links between two adjacent guanines in DNA. Explain why the intrastrand cross-link between two adjacent guanines is a better candidate for nucleotide excision repair rather than for base excision repair.

**Question 9.** Predict the immediate consequences to a cell in which the system of transcription-coupled nucleotide excision repair stopped functioning properly.

**Question 10.** Aside from DNA damage tolerance, name the repair pathway that potentially introduces mutations. Describe how this pathway introduces mutations.

**Question 11.** Explain the difference between DNA repair and DNA damage tolerance.

**Question 12.** Consider a loss-of-function mutant in the nucleotide excision repair and translesion synthesis pathway. Predict the level of DNA damage, percent survival, and level of mutagenesis relative to wild type for each mutant after exposure to UV light. In the table below, fill each blank with increase, decrease, or stay the same.

Mutant Pathway	DNA Damage	Percent Survival	Mutagenesis
NER	—	—	—
Translesion synthesis	—	—	—

**Question 13.** You want to test two strains of *E. coli* for sensitivity to the DNA damaging agent, MMS (methyl methanesulfonate), an alkylating agent that methylates specific bases in DNA. After treating cells with a range of doses of MMS (by incubating the cells with MMS in liquid culture for various times), you plate the cells on a solid media to count survivors (given below). Plot the data for the two strains (percent survival vs. time of MMS exposure). Which strain is more sensitive to MMS? (Note that to count single colonies, you would have to serial-dilute the culture. For this example, we will ignore the serial dilution step.)

Strain	0 Min in MMS	5 Min in MMS	10 Min in MMS
Strain A	254	251	249
Strain B	325	253	189

**Question 14.** There are many claims that certain chemicals that you encounter in daily life are mutagenic. You are interested in learning if chemicals that you commonly use are mutagenic.

To do so, you choose to use the Ames test to test for reversion of a point mutation in the *HisG* gene in *Salmonella typhimurium*. You added a chemical into the growth medium for the bacteria. Assume you plated an equal number of cells for each mutagenesis plate. You also calculated percent survival for chemical-treated cells relative to untreated cells. Remember that the plate media for survival is not selective. A summary of the results is shown below.

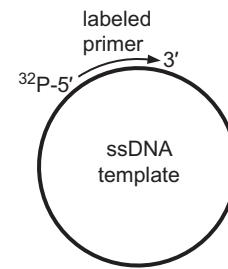
Chemical	Percent Survival	Number of Revertants (# of His <sup>+</sup> Colonies/Selective Plate)
No chemical added	100	28
Chemical A	50	1400
Chemical B	70	20
Chemical C	100	7

- A. What medium must be used in the selective plate as part of the Ames test? Explain how a mutation gives rise to a revertant in this experiment. Be specific.
- B. You are initially surprised to see revertants in the absence of any chemical that you are testing, but you realize that this is normal. Give a specific example of how a revertant can arise in the absence of an added mutagen.
- C. Which chemical(s) would you identify as containing a mutagen? Explain your reasoning.
- D. Which chemical(s) would you identify as possibly antimutagenic? Explain your reasoning.

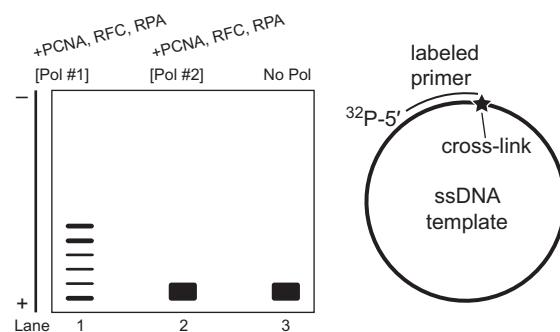
**Question 15.** You perform a follow-up experiment to that discussed in the Questions for Chapter 9. Here is a review of the setup for the experiment. You have just discovered two new eukaryotic DNA polymerases and want to learn more about their properties. To begin, you obtain purified protein of each DNA polymerase and perform polymerase processivity assays. You use a short primer that is <sup>32</sup>P-labeled at the 5' end and binds to a circular single-stranded (ssDNA) DNA template (shown below). You complete the following steps to obtain the processivity

assay results for DNA polymerase #1 (Pol #1) and DNA polymerase #2 (Pol #2).

- In the appropriate buffer conditions, you preincubate the primed, circular ssDNA with Pol #1 or Pol #2 for 5 min at 37°C.
- You add dNTPs and a large excess of ssDNA to initiate the reaction.
- You allow the reaction to proceed for 10 min, quench the reactions with the addition of SDS, and run the samples on a polyacrylamide gel (which resolves single nucleotides) under denaturing conditions.

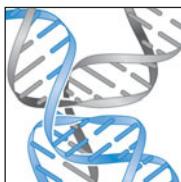


You perform a follow-up experiment, and the results are shown below. In this experiment, you separately incubate each DNA polymerase with the accessory proteins (PCNA, RFC, and RPA) and a DNA template that now includes an intrastrand cross-link of two guanines (induced by cisplatin). The cross-link is located immediately adjacent to the 3' end of the primer as indicated by a star (pictured below to the right).



- A. Describe the processivity for each DNA polymerase in the presence of a template with a cisplatin-induced intrastrand cross-link.
- B. Propose a cellular role for DNA polymerase #1 given this data. Be specific.

CHAPTER 11



# Homologous Recombination at the Molecular Level

ALL DNA IS RECOMBINANT DNA. Genetic exchange works constantly to blend and rearrange chromosomes, most obviously during meiosis, when homologous chromosomes pair before the first nuclear division. During this pairing, genetic exchange between the chromosomes occurs. This exchange, classically termed **crossing over**, is one of the results of **homologous recombination**. Recombination involves the physical exchange of DNA sequences between the chromosomes. The frequency of crossing over between two genes on the same chromosome depends on the physical distance between these genes, with long distances giving the highest frequencies of exchange. In fact, genetic maps derived from early measurements of crossing-over frequencies gave the first real information about chromosome structure by revealing that genes are arranged in a fixed, linear order.

Sometimes, however, gene order *does* change: For example, movable DNA segments called **transposons** occasionally “jump” around chromosomes and promote DNA rearrangements, thus altering chromosomal organization. The recombination mechanisms responsible for transposition and other genome rearrangements are distinct from those of homologous recombination. These mechanisms are discussed in detail in Chapter 12.

Homologous recombination is an essential cellular process catalyzed by enzymes specifically made and regulated for this purpose. Besides providing genetic variation, recombination allows cells to retrieve sequences lost through DNA damage by replacing the damaged section with an undamaged DNA strand from a homologous chromosome. Recombination also provides a mechanism to restart stalled or damaged replication forks (“replication restart”). Furthermore, special types of recombinations regulate the expression of some genes. For example, by switching specific segments within chromosomes, cells can put otherwise dormant genes into sites where they are expressed.

In addition to providing an explanation for genetic processes, elucidating the molecular mechanisms of recombination has led to the development of methods to manipulate genes. It is, for example, now routine to generate gene “knock-out” and “transgenic” variants in many different experimental organisms (see Appendix 1). These methods for deleting and introducing

## OUTLINE

- DNA Breaks Are Common and Initiate Recombination, 342
  - Models for Homologous Recombination, 342
  - Homologous Recombination Protein Machines, 349
  - Homologous Recombination in Eukaryotes, 362
  - Mating-Type Switching, 369
  - Genetic Consequences of the Mechanism of Homologous Recombination, 371
- Visit Web Content for Structural Tutorials and Interactive Animations

genes within the context of a whole organism rely on recombination and are exceedingly powerful for determining gene function.

## DNA BREAKS ARE COMMON AND INITIATE RECOMBINATION

---

Double-stranded breaks (DSBs) in DNA arise frequently. If these breaks are not repaired, the consequence to the cell is disastrous. For example, a single DSB in the *Escherichia coli* chromosome is lethal to a cell that lacks the ability to repair it. The major mechanism used to repair DSBs in most cells is **homologous recombination**. Some cell types also use a simpler mechanism, such as **nonhomologous end joining** (NHEJ) to heal their chromosomes. This process is described in Chapter 10 (see text and Box 10-5).

In bacteria, the major biological role of homologous recombination is to repair DSBs. These broken DNA ends arise by several means (see also Chapter 10). Ionizing radiation and other damaging agents can directly break both strands of the DNA backbone. Many types of DNA damage also indirectly give rise to DSBs by interfering with the progress of a replication fork. For example, an unrepaired nick in one DNA strand will lead to collapse of a passing replication fork (Fig. 11-1). Similarly, a lesion in DNA that makes a strand unable to serve as a template will stop a replication fork. This type of stalled fork can be processed by several different pathways (e.g., fork regression or nuclease digestion) (see Fig. 11-1) that give rise to a DNA end with a DSB. These broken DNA ends then initiate recombination with a homologous DNA molecule, a process that will, in turn, heal the break.

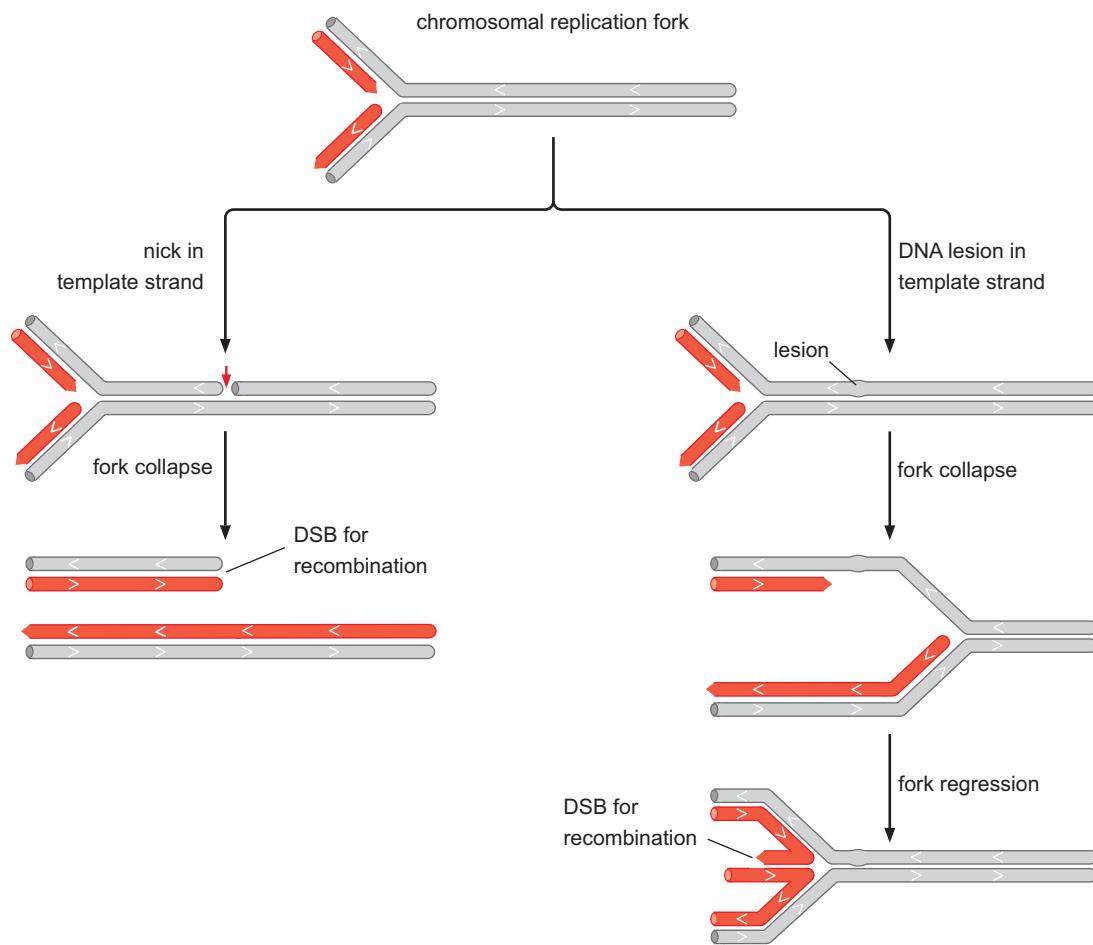
In addition to repairing DSBs in chromosomal DNA, homologous recombination promotes genetic exchange in bacteria. This exchange occurs between the chromosome of one cell and the DNA that enters that cell via phage-mediated transduction or cell-to-cell conjugation (as we see in Chapter 7). In these cases, the new DNA enters the cell as a linear molecule and thus provides the critical “broken” DNA end needed to initiate recombination.

In eukaryotic cells, homologous recombination is critical for repairing DNA breaks and collapsed replication forks. This role of chromosome repair and replication restart is the principal function of homologous recombination in most somatic cells in complex organisms as well as in vegetatively growing single-cellular eukaryotes. However, there are other times when recombination for genetic exchange and chromosome maintenance is specifically needed. As described below, recombination is *essential* to the process of chromosome pairing during meiosis. In this case, as cells enter meiosis, they produce a specific protein to introduce DSBs into the DNA and thereby initiate the recombination pathway. Thus, although they arise from many different sources, the appearance of DSBs in DNA is a key early event in homologous recombination.

## MODELS FOR HOMOLOGOUS RECOMBINATION

---

Elegant early experiments using heavy isotopes of atoms incorporated into DNA provided the first molecular view of the process of homologous recombination. This is the same approach used by Matthew Meselson and Franklin W. Stahl to show that DNA replicates in a semiconservative manner (see Chapter 2). In their experiments, Meselson and Stahl showed that the products of replication contain one old and one newly synthesized DNA



**FIGURE 11-1** Damage in the DNA template can lead to DSB formation during DNA replication. This is easiest to see when the template contains a nick (left panel), but it also can occur when the template carries a fork-stopping lesion (right panel). In this case, the two newly synthesized strands (red) can base-pair and the fork can regress. This structure can be further processed by a number of means. The broken end can serve to initiate recombination.

strand. In contrast, this same experimental approach revealed that the recombination process under investigation involved the direct breakage and rejoining of DNA molecules. As we shall see in the following sections, we now understand that breakage and joining of DNA is a central aspect of homologous recombination. But recombination usually involves also at least the limited destruction and resynthesis of DNA strands. In the years since these initial experiments, numerous models have been proposed to explain the molecular mechanism of genetic exchange. Key steps of homologous recombination present in these models include the following:

1. *Alignment of two homologous DNA molecules.* By “homologous” we mean that the DNA sequences are identical or nearly identical for a region of at least 100 bp or so. Despite this high degree of similarity, DNA molecules can have small regions of sequence difference and may, for example, carry different sequence variants, known as **alleles**, of the same gene.
2. *Introduction of breaks in the DNA.* Once the breaks are formed, the ends at the breaks are further processed to generate regions of single-stranded DNA.

3. *Strand invasion.* Initial short regions of base pairing are formed between the two recombining DNA molecules. This pairing occurs when a single-stranded region of DNA originating from one parental molecule pairs with its complementary strand in the homologous duplex DNA molecule. This event is called **strand invasion**. As a result of the strand invasion process, regions of new duplex DNA are generated; this DNA, which often contains some mismatched base pairs, is called **heteroduplex DNA**.
4. *Formation of the Holliday junction.* After strand invasion, the two DNA molecules become connected by crossing DNA strands to form a structure that is called a **Holliday junction**. This junction can move along the DNA by the repeated melting and formation of base pairs. Each time the junction moves, base pairs are broken in the parental DNA molecules while identical base pairs are formed in the recombination intermediate. This process is called **branch migration**.
5. *Resolution of the Holliday junction.* The process to regenerate DNA molecules and therefore finish genetic exchange is called **resolution**. Resolution can be achieved in one of two ways, either by cleavage of the Holliday junction or (in eukaryotes) by a process of “dissolution.” In the first, cutting the DNA strands within the Holliday junction regenerates two separate duplexes. As we shall see, which of the two pairs of DNA strands in the Holliday junction is cut during resolution has a large impact on the extent of DNA exchange that occurs between the two recombining molecules (see Interactive Animation 11-1). In the second (alternative) process, resolution is achieved by dissolution, a sort of convergence/collapse mechanism, which we describe in more detail below.

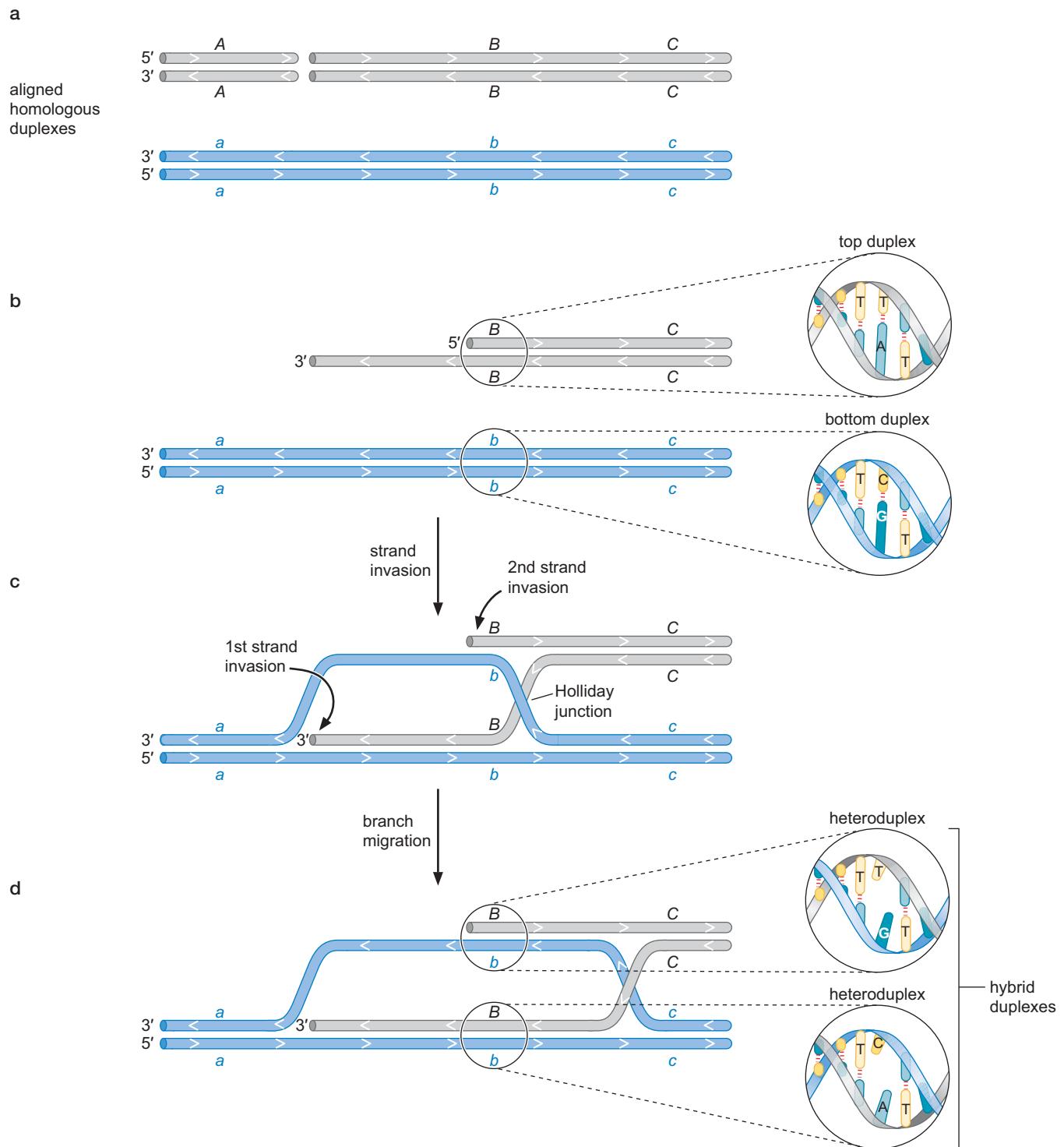


### Strand Invasion Is a Key Early Step in Homologous Recombination

When illustrating the steps of homologous recombination, it is useful to picture the two homologous, double-stranded DNA molecules aligned, as shown in Figure 11-2a. These molecules, although nearly identical, carry different alleles of the same gene (as is denoted by the *A/a*, *B/b*, and *C/c* symbols in Fig. 11-2), which are helpful for following the outcome of recombination.

Recombination is initiated by the presence of a DSB in one of the DNA molecules (Fig. 11-2b). DNA strands near the break site can then be “peeled” away from their complementary strands, freeing these strands to invade and ultimately base-pair with the homologous duplex (Fig. 11-2c). Processing of the strands near the break site is described in more detail below. Strand invasion is the central step in homologous recombination, because it is this invasion and then pairing of complementary strands between the two homologous duplexes that establishes the stable pairing between these molecules. This process also initiates the exchange of DNA strands between the two “parental” DNAs. As we shall see below, the enzymes that catalyze strand invasion are called **strand-exchange proteins** because they promote this critical reaction.

Strand invasion generates a Holliday junction that can then move along the DNA by branch migration. This migration increases the length of the DNA exchanged. If the two DNA molecules are not identical—but, for example, carry a few small sequence differences, as is often true between two alleles of the same gene—branch migration through these regions of sequence difference generates DNA duplexes carrying one or a few sequence mismatches (see *B* and *b* alleles in Fig. 11-2d and the inset). Repair of these mismatches in the heteroduplex DNA can have important genetic consequences, a point we return to at the end of the chapter.



**FIGURE 11-2** Holliday model through the steps of branch migration. The small arrowheads on the DNA single strands point in the 5'-to-3' direction. Note that A and *a*, B and *b*, and C and *c* specify different alleles and have slightly different DNA sequences. Therefore, heteroduplex DNA containing those genes (shown in the expanded section in panel d) will have some mismatches.

## Resolving Holliday Junctions Is a Key Step to Finishing Genetic Exchange

Finishing recombination usually requires resolution of the Holliday junction by cutting the DNA strands near the site of the cross; this reaction separates the two recombining DNA molecules and thus completes the genetic exchange. Figure 11-3 shows two homologous DNA duplexes connected by a single Holliday junction. Resolution occurs in one of two ways, and, therefore, gives rise to two distinct classes of DNA products, as we now describe.

Figure 11-3 illustrates where the alternative pairs of DNA cut sites occur during resolution on this simple branched DNA generated by exchange between two similar duplex DNA molecules. To make these cut sites easier to visualize, the Holliday junction is “rotated” to give a square, planar structure with no crossing strands. The two strands with the same sequence and polarity must be cleaved; the two alternative choices for cleavage sites are labeled site 1 and site 2 in Figure 11-3.

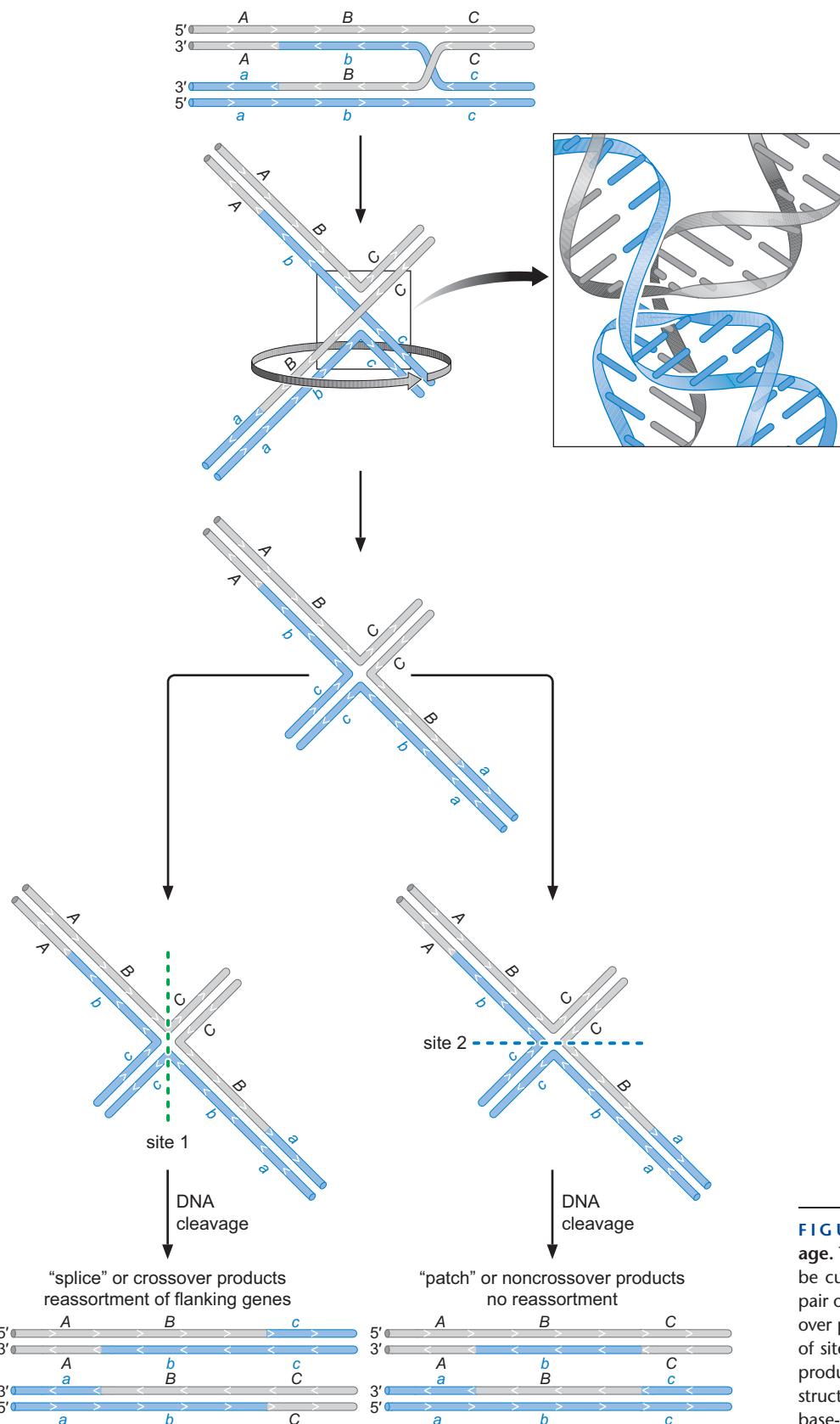
In this example, the cut sites marked 1 occur in the two DNA strands that are composed entirely of DNA from one of the two parental DNA molecules (e.g., the solid blue and solid gray strands). If these strands are now cut and then covalently joined (the second reaction is catalyzed by DNA ligase as discussed below), the resulting DNA molecules will have the structure and sequence shown on the left in the bottom of the figure. These products are referred to as “splice” recombination products, because the two original duplexes are now “spliced together” such that regions from the parental DNA molecules are covalently joined together by a region of hybrid duplex. As seen by following the allele markers, generation of splice products results in reassortment of genes that flank the site of recombination. Therefore, this type of recombinant is also called the **crossover product**, because, within this DNA molecule, crossing over has occurred between the *A* and *C* genes.

In contrast, the alternative pair of cut sites in the Holliday junction (labeled 2 in Fig. 11-3) is in the two DNA strands that *contain regions of sequence from both parental molecules* (e.g., both blue and gray segments). After resolution and covalent joining of the strands at these sites, the resulting DNA molecules contain a region or “patch” of hybrid DNA. These molecules are thus known as the **patch products**. In these products, recombination does not result in reassortment of the genes flanking the site of initial cleavage (see the fate of the *A/a* and *C/c* allele markers in Fig. 11-3). These molecules are therefore also known as the **noncrossover products**. Factors that influence the site and polarity of resolution are discussed below.

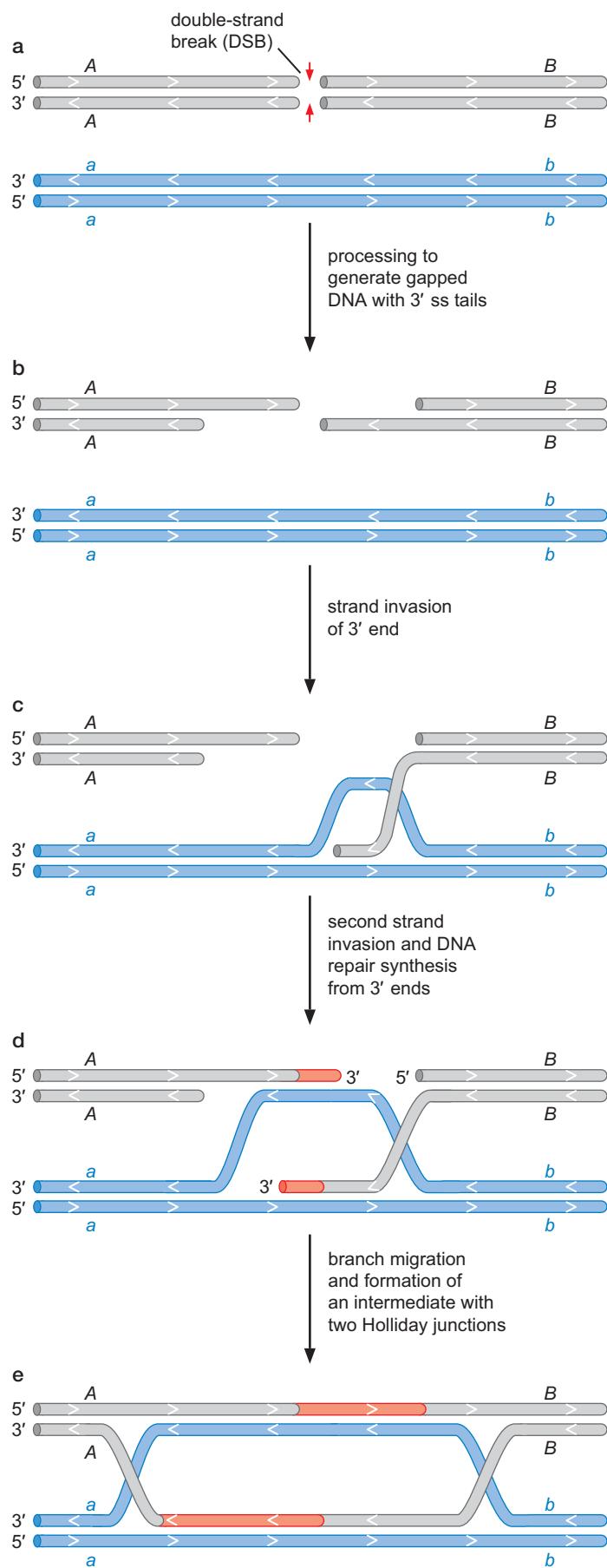
## The Double-Strand Break–Repair Model Describes Many Recombination Events

Homologous recombination is often initiated by DSBs in DNA. A common model describing this type of genetic exchange is the **double-strand break–repair pathway** (Fig. 11-4). This pathway starts with the introduction of a DSB in one of two homologous duplex DNA molecules (Fig. 11-4a). The other DNA duplex remains intact. The asymmetric initial breakage of the two DNA molecules in the DSB-repair model necessitates that later stages in the recombination process are also asymmetric (i.e., the two duplexes are treated differently, as we shall see).

After introduction of the DSB, a DNA-cleaving enzyme sequentially degrades the broken DNA molecule to generate regions of single-stranded DNA (ssDNA) (Fig. 11-4b). This processing creates single-strand extensions, known as ssDNA tails, on the broken DNA molecules; these ssDNA tails terminate with 3' ends. In some cases, both strands at a DSB are processed, whereas in other cases, only the 5'-terminating strand is degraded.



**FIGURE 11-3** Holliday junction cleavage. Two alternative pairs of DNA sites can be cut during resolution. Cleavage at one pair of sites generates the "splice" or crossover products. Cleavage at the second pair of sites yields the "patch" or noncrossover products. (Inset) A Holliday junction DNA structure. Notice that the DNA is completely base-paired in this structure.



**FIGURE 11-4** DSB-repair model for homologous recombination. Shown are the steps leading to generation of a recombination intermediate with two Holliday junctions.

The ssDNA tails generated by this process then invade the unbroken homologous DNA duplex (Fig. 11-4c). This panel of the figure shows one strand invasion, as likely occurs initially, whereas the next panel shows the two invading strands. In each case, the invading strand base-pairs with its complementary strand in the other DNA molecule. Because the invading strands end with 3' termini, they can serve as primers for new DNA synthesis. Elongation from these DNA ends—using the complementary strand in the homologous duplex as a template—serves to regenerate the regions of DNA that were destroyed during the processing of the strands at the break site (Fig. 11-4d,e).

If the two original DNA duplexes were not identical in sequence near the site of the break (e.g., having single-base-pair changes as described above), sequence information could be lost during recombination by the DSB-repair pathway. In the recombination event shown in Figure 11-4, sequence information lost from the gray DNA molecule as a result of DNA processing is replaced by the sequence present on the blue duplex as a result of DNA synthesis. This nonreciprocal step in DSB repair sometimes leaves a genetic trace—giving rise to a **gene conversion** event—a point that we shall return to at the end of the chapter.

The two Holliday junctions found in the recombination intermediates generated by this model move by branch migration and ultimately are resolved to finish recombination. Once again, the strands that are cleaved during resolution of these Holliday junctions determine whether or not the product DNA molecules will contain reassorted genes in the regions flanking the site of recombination (i.e., result in crossing over). The different ways to resolve a recombination intermediate containing two Holliday junctions are explained in Box 11-1, How to Resolve a Recombination Intermediate with Two Holliday Junctions.

## HOMOLOGOUS RECOMBINATION PROTEIN MACHINES

Organisms from all branches of life encode enzymes that catalyze the biochemical steps of recombination. In some cases, members of homologous protein families provide the same function in all organisms. In contrast, other recombination steps are catalyzed by different classes of proteins in different organisms but with the same general outcome (see Interactive Animation 11-2). Our most detailed understanding of the mechanism of recombination comes from studies of *E. coli* and its phage. Thus, in the following sections, we first focus on the proteins that promote recombination in *E. coli* via a major DSB-repair pathway, known as the **RecBCD pathway**. Homologous recombination in eukaryotic cells and the proteins involved in these events are considered in sections below.



Table 11-1 lists the proteins that catalyze critical recombination steps in bacteria as well as those that serve these same functions in eukaryotes (the budding yeast *Saccharomyces cerevisiae* is the best-understood example). These proteins provide activities needed to complete important steps in the DSB-repair pathway. In addition to these dedicated recombination proteins, DNA polymerases, ssDNA-binding proteins, topoisomerases, and ligases also have critical roles in the process of genetic exchange.

Note that absent from the list in Table 11-1 is an *E. coli* protein that introduces DSBs in DNA, despite the fact that recombination via the RecBCD pathway requires a DSB on one of the two recombining DNA molecules. As discussed above, in bacteria, no specific protein has been found that performs

## ► ADVANCED CONCEPTS

**Box 11-1** How to Resolve a Recombination Intermediate with Two Holliday Junctions

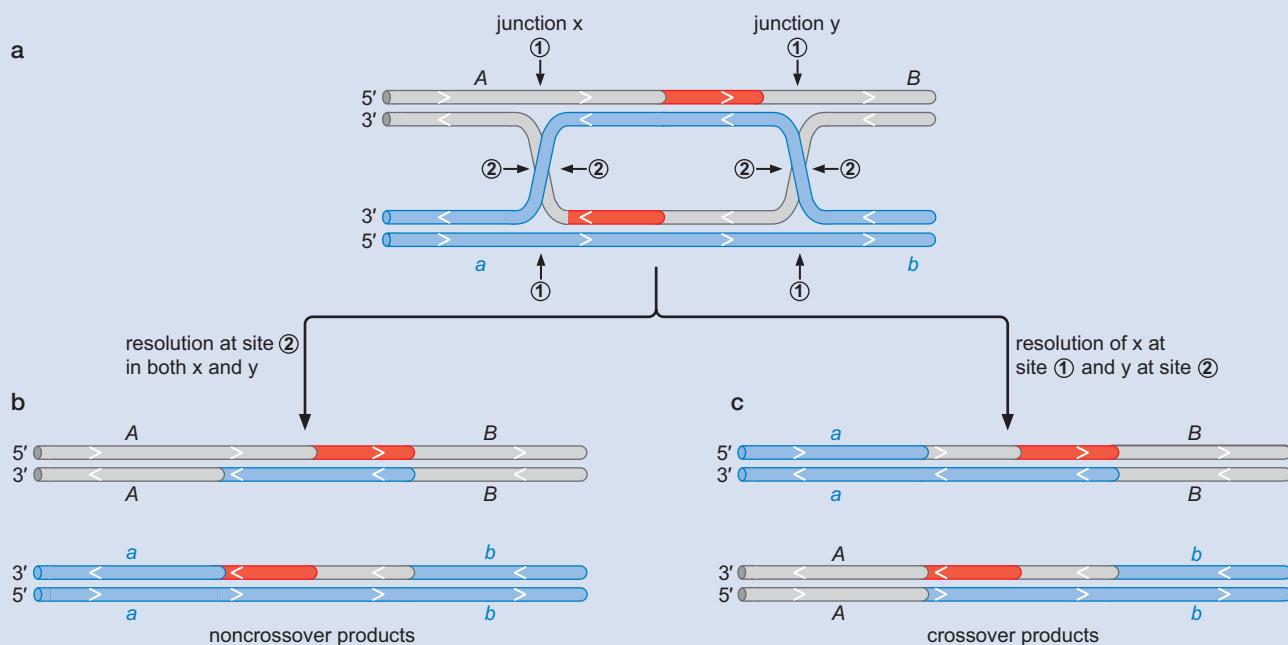
How the Holliday junctions present in a recombination intermediate are cleaved has a huge impact on the structure of the product DNA molecules. Products will either have the DNA flanking the site of recombination reassorted (in the splice/crossover products) or not (in the patch/noncrossover products) depending on how resolution is achieved. Because the intermediates generated by the DSB-repair pathway contain two Holliday junctions, it can be difficult to see which products are generated by the different possible combinations of Holliday junction–cleavage events. In fact, there is a simple pattern that determines whether crossover or noncrossover products are generated.

To explain the different possible ways these intermediates can be resolved, consider the two junctions (labeled x and y) in Box 11-1 Figure 1. For each junction, there are two possible cleavage sites (labeled site 1 and site 2). The simple rule that determines whether or not resolution will result in crossover versus noncrossover products is as follows. If both junctions are cleaved *in the same way*, that is, either both at site 1 or both at site 2, then noncrossover products will be generated. An example of this type of product is shown in panel b of the figure; these are the molecules generated when both Holliday junctions are cleaved at site 2. Note that the allele markers A/B and a/b are still on the same DNA molecules as they were in the parental chromosomes. Cleavage of both junctions at site 1 also generates noncrossover products.

In contrast, when the two Holliday junctions are cleaved *using different sites*, then the crossover products are generated. An example of this type of resolution is shown in Box 11-1 Figure 1c. Here junction x was cleaved at site 1, whereas junction y was cleaved at site 2. Note that now gene A is linked to gene b, whereas gene a is linked to gene B; thus, reassortment of the flanking genes has occurred. Cleavage of junction x at site 2 and junction y at site 1 also generates crossover products.

Why is the simple rule true? To understand this, compare the junctions shown here with the single Holliday junction shown in Figure 11-3. It can be seen that, at a single junction, cleavage at site 1 would give the splice products, whereas cleavage at site 2 would generate patch products. Therefore, when the results of cleavage at the two junctions are combined, this is what happens:

- Cleavage of both junctions at site 2 will give a patch product (patch + patch = patch, noncrossover products).
- Cleavage at both junctions at site 1 also gives a patch product (splice + splice = patch) because the second splice-type resolution essentially “undoes” the rearrangement caused by the first cleavage.
- Cleavage of one junction at site 1, but the other at site 2 therefore generates crossover products (splice + patch = splice) because the rearrangement caused by the site 1 cleavage is retained in the final product.



**BOX 11-1 FIGURE 1** Two possible ways of resolving an intermediate from the DSB-repair pathway. The parental DNA molecules were like those in Figure 11-4. The regions of red DNA are those that were resynthesized during recombination.

**TABLE 11-1** Prokaryotic and Eukaryotic Factors That Catalyze Recombination Steps

Recombination Step	<i>E. coli</i> Protein Catalyst	Eukaryotic Protein Catalyst
Pairing homologous DNAs and strand invasion	RecA protein	Rad51 Dcm1 (in meiosis) Spo11 (in meiosis)
Introduction of DSB	None	HO (for mating-type switching)
Processing DNA breaks to generate single strands for invasion	RecBCD helicase/nuclease	MRX protein (also called Rad50/58/60 nuclease)
Assembly of strand-exchange proteins	RecBCD and RecFOR	Rad52 and Rad59
Holliday junction recognition and branch migration	RuvAB complex	Not well characterized
Resolution of Holliday junctions	RuvC	Rad51c–XRCC3 complex, WRN, and BLM

this task. Rather, breaks generated as a result of DNA damage, “mis-steps” in DNA repair, or collapse of a replication fork are the major source of these initiating events in chromosomal DNA. Alternatively, during genetic exchange reactions, such as phage-mediated transduction (which we shall consider in Appendix 1), the infecting DNA segment carries broken DNA ends.

The following sections describe the *E. coli* recombination proteins and how they perform their functions during recombination by the DSB-repair pathway. These proteins are discussed in the order in which they appear during the reaction pathway. First, we consider how the RecBCD enzyme processes DNA at the site of the DSB to generate single-strand regions. Next, the structure and mechanism of RecA, the strand-exchange protein, are described. RecA, after assembling on the ssDNA, finds regions of sequence homology in the DNA molecules and generates new base-pairing partners between these regions. The RuvA and RuvB proteins that drive DNA branch migration are then described. Finally, the Holliday junction–resolving enzyme, RuvC, is considered.

### The RecBCD Helicase/Nuclease Processes Broken DNA Molecules for Recombination

DNA molecules with ssDNA extensions or tails are the preferred substrate for initiating strand exchange between regions of homologous sequence. The **RecBCD enzyme** processes broken DNA molecules to generate these regions of ssDNA. RecBCD also helps load the RecA strand-exchange protein onto these ssDNA ends. In addition, as we shall see, the multiple enzymatic activities of RecBCD provide a means for cells to “determine” whether to recombine with or destroy DNA molecules that enter a cell.

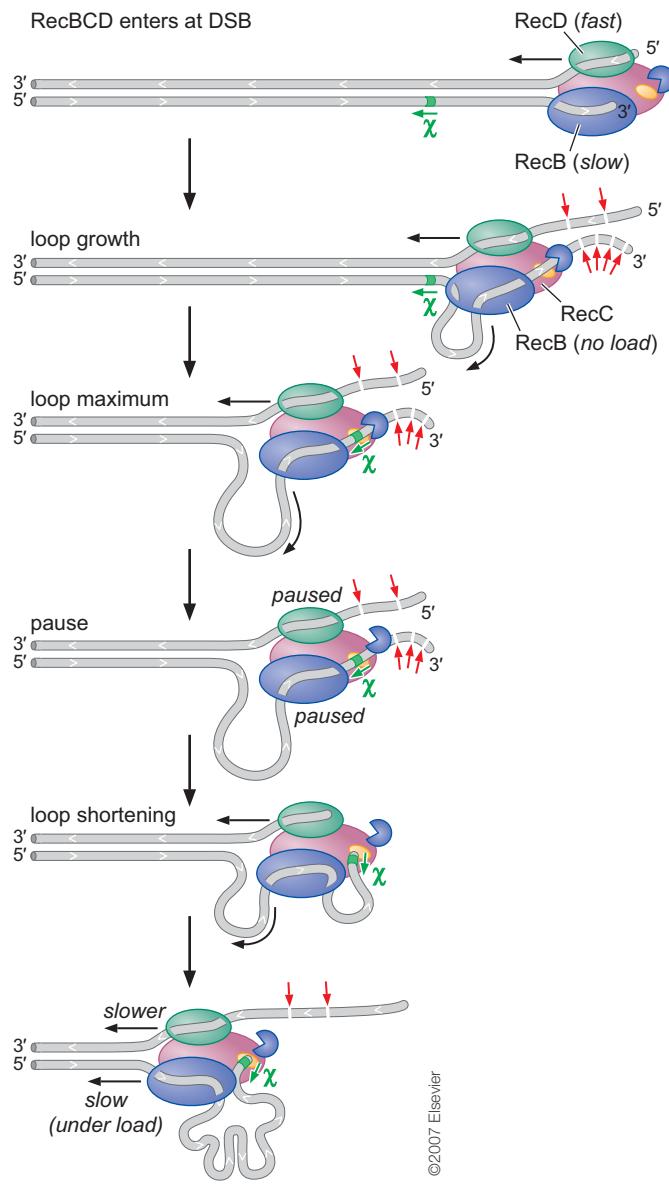
RecBCD is composed of three subunits (the products of the *recB*, *recC*, and *recD* genes) and has both DNA helicase and nuclease activities. The complex binds to DNA molecules at the site of a DSB and tracks along DNA using the energy of ATP hydrolysis. As a result of its action, the DNA is unwound, with or without the accompanying nucleolytic destruction of one or both of the DNA strands. The activities of RecBCD are controlled by specific DNA sequence elements known as **Chi sites** (for “crossover hot spot instigator”). Chi sites were discovered because they stimulate the frequency of homologous recombination.

The RecB and RecD subunits are both DNA helicases, that is, enzymes that use ATP hydrolysis to melt and unwind DNA base pairs (see Chapter 9). The RecB subunit contains a 3'-to-5' helicase and has also a multifunctional nuclease domain that digests the DNA as it moves along. RecD is a 5'-to-3' helicase, and RecC functions to recognize Chi sites. But how do these various sub-

units of this complex multifunctional machine work together to move along DNA, and what actually happens when the complex encounters a Chi site?

Various studies, including those based on single-molecule techniques, have shown that the RecB and RecD helicase “motors” move independently along opposite strands of the DNA duplex and at different speeds. Together, they are capable of “driving” the RecBCD complex along the DNA at rates of >1000 bp per second! Chi sites within DNA act as a sort of “molecular throttle” to regulate the activities of the helicases and therefore the speed of DNA translocation. Figure 11-5 shows an overall schematic of RecBCD processing a DNA molecule containing a single Chi site to activate this DNA for recombination. RecBCD enters the DNA at the site of the DSB and moves along the DNA, unwinding the strands. But during this initial phase, the two motors of the complex are not moving at equal rates—the RecD subunit runs faster than RecB and therefore leads the complex. As RecB tries to keep up, a loop of ssDNA from the 3' end bulges out ahead of the complex.

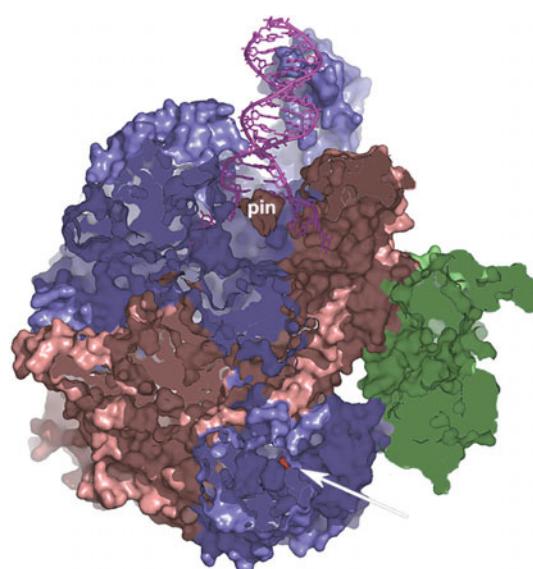
Upon encountering the Chi sequence, the complex pauses for a few seconds, then continues at about one-half the initial rate. During the pause, three



**FIGURE 11-5** Steps of DNA processing by RecBCD. Note that RecBCD protein could have entered this DNA molecule from either or both broken ends. However, Chi sites function only in one orientation. On the DNA molecule shown, the Chi site is oriented such that it will only modify a RecBCD enzyme moving from right to left. The RecBCD enzyme has two DNA helicases: RecD (green), which moves rapidly on the 5'-ending strand (top strand), and RecB (purple), which moves slowly on the 3'-ending strand (top strand). Because these two subunits travel at different speeds, RecB accumulates a single-stranded DNA loop in the lower strand during unwinding. After the enzyme encounters the Chi site, this loop is “reeled in” and its 3' end, now containing Chi, is available for RecA assembly. (Adapted, with permission, from Spies et al. 2007. *Cell* 131: 694–705, Fig. 5, p. 701. © Elsevier.)

events occur to change the activity of the complex. First, the looped-out ssDNA is pulled or reeled back in by the RecB subunit, and RecB becomes the primary motor now leading the complex; second, a possible conformational change occurs that results in uncoupling of the RecD subunit; and, third, the nuclease activity of the RecBCD complex is altered. As RecBCD moves into the sequence beyond the Chi site, the nuclease no longer cleaves the DNA strand with  $3' \rightarrow 5'$  polarity. Furthermore, the opposite DNA strand is cleaved more frequently than it was before the Chi site was encountered. As a result of this change in activity, the DNA duplex now has a  $3'$  single-strand extension terminating with the Chi sequence at the  $3'$  end. This structure is ideal for assembly of RecA and initiation of strand exchange (see below). We now consider the molecular basis of the change in RecBCD's enzyme activity after the encounter with a Chi site and the change in the way the DNA travels through the multi-subunit RecBCD complex.

The structure of the RecBCD complex bound to DNA provides further insight into how this three-subunit machine functions and how its activity changes upon encountering a Chi site (see Structural Tutorial 11-1). As shown in Figure 11-6, the protein complex has an overall triangular shape, with duplex DNA entering the protein from the top point of the triangle. Here, the DNA encounters a “pin” structure protruding from the RecC subunit that splits the duplex and guides the two individual strands of DNA to the two motors within the enzyme. The RecC subunit channels the  $3'$  strand to the RecB motor and the  $5'$  strand to the RecD motor. In this manner, RecC, which is not itself a helicase, contributes to the overall efficiency of the helicase activity of the complex. The organization of the channels within the enzyme causes the  $3'$  DNA tail to be fed along a groove that emerges at the nuclease active site on the RecB subunit. As a result, before the enzyme complex encounters a Chi site, this strand is efficiently and processively degraded. The  $5'$  DNA tail also moves past the nuclease active site upon leaving the RecD motor, but it is digested less frequently than the  $3'$  tail, because it must compete with the more favorably positioned  $3'$  strand. However, upon encountering a Chi site, the situation changes. RecC recognizes and binds tightly to this DNA site, and once this  $3'$  end is bound, it is prevented from entering the nuclease. This binding therefore both prevents further digestion of the  $3'$  tail and promotes digestion of the  $5'$  tail, by removing its competitor.



**FIGURE 11-6** The structure of the RecBCD–DNA complex. Here RecB is shown in blue, RecC in magenta, and RecD in green. The bound DNA (purple) enters through the “top” of the complex, and the white arrow points to a bound calcium ion (red) in the RecB nuclease active site. The structure shows a cutaway view to reveal the DNA. RecC contacts both DNA strands and splits them, directing the  $3'$  strand to RecB on the left and the  $5'$  tail to RecD on the right. (Singleton M.R. et al. 2004. *Nature* 432: 7015; PDB Code: 1W36.) Image prepared with PyMOL (DeLano W.L. 2002. *The PyMOL molecular graphics system*. DeLano Scientific, Palo Alto, California. <http://www.pymol.org>).

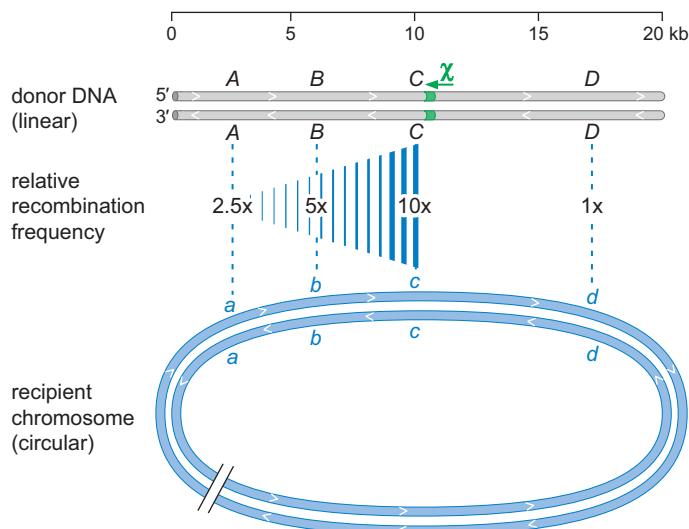
The ssDNA tail generated by RecBCD must be coated by the RecA protein for recombination to occur. However, cells also contain ssDNA-binding protein (SSB) that can bind to this DNA. To ensure that RecA, rather than SSB, binds these ssDNA tails, RecBCD interacts directly with RecA and promotes its assembly. This loading activity involves a direct protein–protein interaction between the nuclease domain of the RecB subunit and the RecA protein and serves to load RecA on the DNA with the 3' tail.

### Chi Sites Control RecBCD

Chi sites increase the frequency of recombination ~10-fold. This stimulation is most pronounced directly adjacent to the Chi site. Although elevated recombination frequencies are observed for ~10 kb distal to the Chi site, they drop off gradually over this distance (Fig. 11-7). The observation that recombination is stimulated specifically only on one “side” of the Chi site was initially puzzling. It is now clear, however, why this pattern is observed: The DNA between the DSB (where RecBCD enters) and the Chi site is cut into small pieces by the enzyme and is therefore not available for recombination. In contrast, DNA sequences met by RecBCD after its encounter with Chi are preserved in a recombinogenic, single-strand form and are specifically loaded with RecA.

The ability of Chi sites to control the nuclease activity of RecBCD also helps bacterial cells protect themselves from foreign DNA that may enter via phage infection or conjugation. The 8-nucleotide Chi site (GCTGGTGG) is highly overrepresented in the *E. coli* genome: Although, mathematically, it is predicted to occur only once every 65 kb, or about 80 times, the chromosomal sequence reveals the presence of 1009 Chi sites! Because of this overrepresentation, *E. coli* DNA that enters an *E. coli* cell is likely to be processed by RecBCD in a manner that generates the 3' ssDNA tails, and thus to be activated for recombination. In contrast, DNA from a bacteriophage or from another species (in which *E. coli* Chi sites are not overrepresented) will lack frequent Chi sites. RecBCD action on this DNA will lead to its extensive degradation, rather than activation for recombination.

In summary, the DNA-degradation activity of RecBCD has multiple consequences: This degradation is needed to process DNA at a break site for the subsequent steps of RecA assembly and strand invasion. In this manner, RecBCD promotes recombination. However, because RecBCD degrades DNA to acti-



**FIGURE 11-7** Polar action of Chi. This schematic shows that a Chi site specifically elevates recombination frequencies directly at the site, as well as in the distal sequences. The recombination event shown represents exchange between a transferred linear DNA segment introduced into a cell by transduction or conjugation and the bacterial chromosome. The actual DNA segments participating may be much longer. For example, phage transduction often delivers an ~80-kb segment of DNA. The *E. coli* chromosome is ~5 Mb.

vate it, the overall process of homologous recombination must also involve DNA synthesis to regenerate the degraded strands. In addition, RecBCD sometimes functions simply to destroy DNA, as it does when foreign DNA lacking frequent Chi sites enters cells. In this way, RecBCD can protect cells from the potentially deleterious consequences of taking up foreign sequences, which, for example, may carry a bacteriophage or other harmful agent.

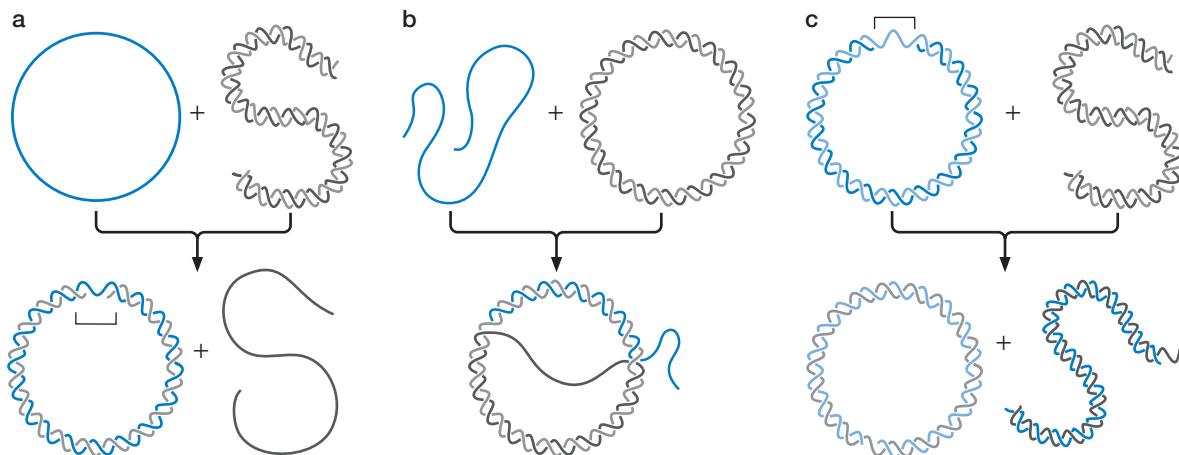
### RecA Protein Assembles on Single-Stranded DNA and Promotes Strand Invasion

RecA is the central protein in homologous recombination. It is the founding member of a family of enzymes called **strand-exchange proteins**. These proteins catalyze the pairing of homologous DNA molecules. Pairing involves both the search for sequence matches between two molecules and the generation of regions of base pairing between these molecules.

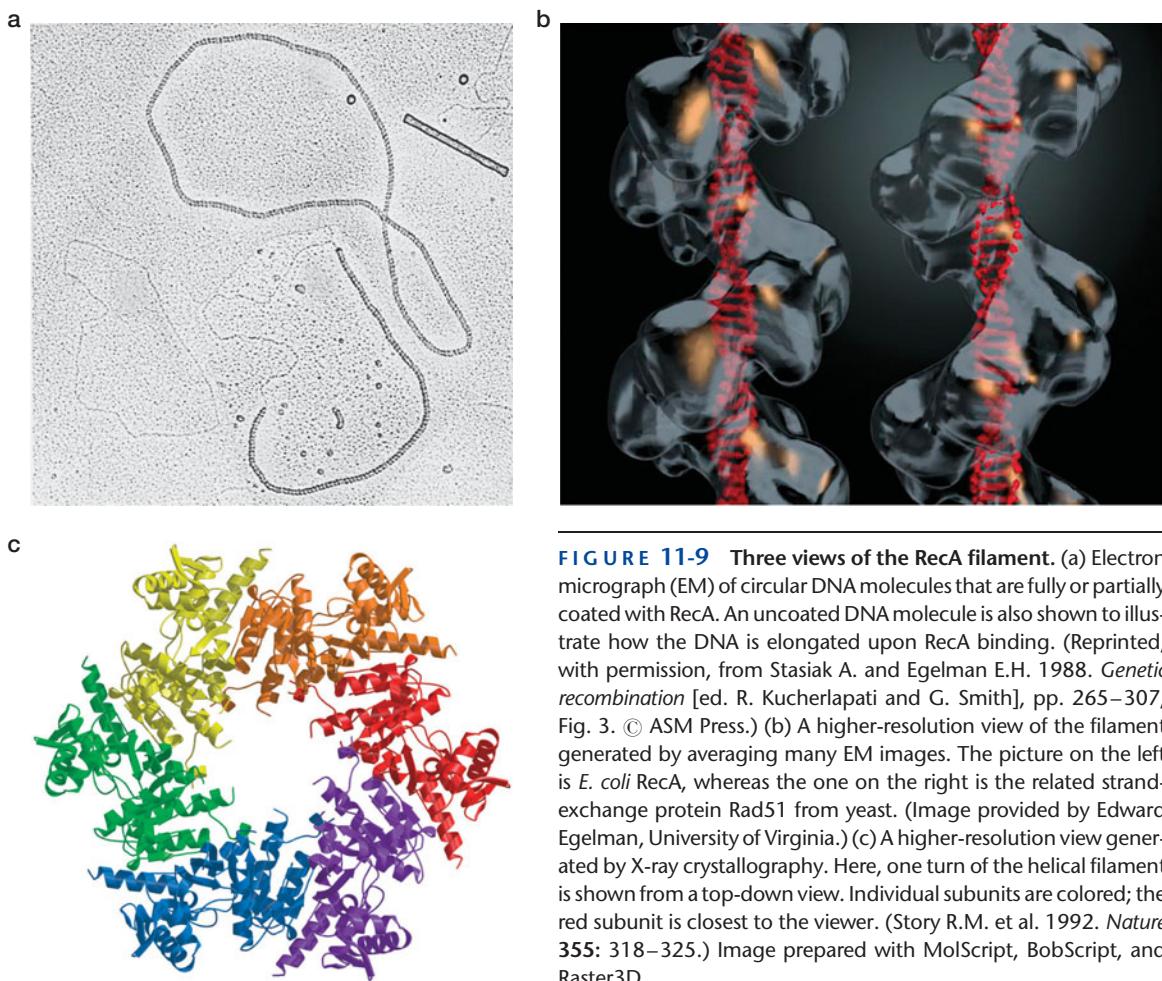
The DNA pairing and strand-exchange activities of RecA can be observed using simple DNA substrates *in vitro*; examples of DNA pairing and strand-exchange reactions useful for showing the biochemical activities of RecA are shown in Figure 11-8. The important features of these DNA molecules are (1) DNA sequence complementarity between the two partner molecules; (2) a region of ssDNA on at least one molecule to allow RecA assembly; and (3) the presence of a DNA end within the region of complementarity, enabling the DNA strands in the newly formed duplex to intertwine.

The active form of RecA is a protein–DNA filament (Fig. 11-9). Unlike most proteins involved in molecular biology that function in smaller discrete protein units, such as monomers, dimers, or hexamers, the RecA filament is huge and variable in size; filaments that contain approximately 100 subunits of RecA and 300 nucleotides of DNA are common. The filament can accommodate one, two, three, or even four strands of DNA. As described below, filaments with either one or three bound strands are most common in recombination intermediates.

The structure of DNA within the filament is highly extended compared with either uncoated ssDNA or a standard B-form helix. On average, the distance between adjacent bases is 5 Å, rather than the 3.4 Å spacing normally observed (Chapter 4). Thus, upon RecA binding, the length of a DNA mole-



**FIGURE 11-8** Substrates for RecA strand exchange. Shown here are three possible structure combinations that participate in DNA pairing and strand exchange. Note that the brackets in parts a and c show the location of a gap in one of the strands.



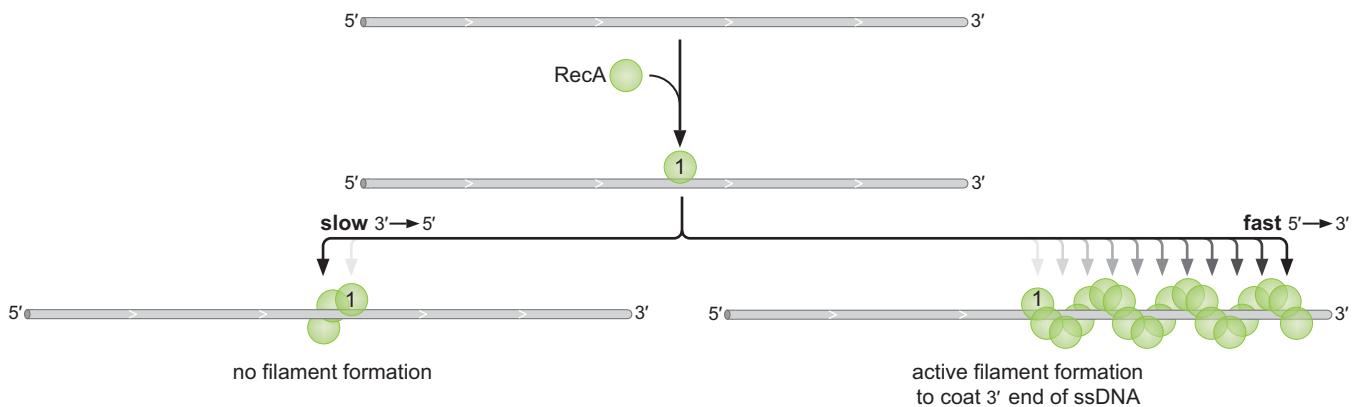
**FIGURE 11-9** Three views of the RecA filament. (a) Electron micrograph (EM) of circular DNA molecules that are fully or partially coated with RecA. An uncoated DNA molecule is also shown to illustrate how the DNA is elongated upon RecA binding. (Reprinted, with permission, from Stasiak A. and Egelman E.H. 1988. *Genetic recombination* [ed. R. Kucherlapati and G. Smith], pp. 265–307, Fig. 3. © ASM Press.) (b) A higher-resolution view of the filament generated by averaging many EM images. The picture on the left is *E. coli* RecA, whereas the one on the right is the related strand-exchange protein Rad51 from yeast. (Image provided by Edward Egelman, University of Virginia.) (c) A higher-resolution view generated by X-ray crystallography. Here, one turn of the helical filament is shown from a top-down view. Individual subunits are colored; the red subunit is closest to the viewer. (Story R.M. et al. 1992. *Nature* 355: 318–325.) Image prepared with MolScript, BobScript, and Raster3D.

cule is extended ~1.5-fold (Fig. 11-9a). It is within this RecA filament that the search for homologous DNA sequences is conducted and the exchange of DNA strands executed.

To form a filament, subunits of RecA bind cooperatively to DNA. RecA binding and assembly are much more rapid on ssDNA than on double-stranded DNA, thus explaining the need for regions of ssDNA in strand-exchange substrates. The filament grows by the addition of RecA subunits in the 5'-to-3' direction, such that a DNA strand that terminates in 3' ends is most likely to be coated by RecA (Fig. 11-10). Note that in the DSB-repair model for recombination, it is DNA molecules with just this configuration that participate in strand invasion.

### Newly Base-Paired Partners Are Established within the RecA Filament

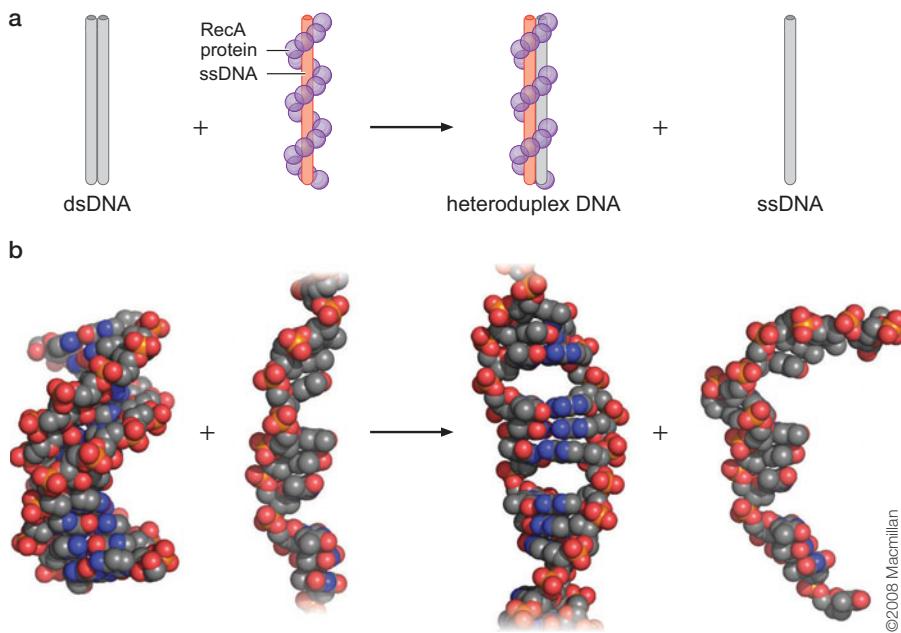
RecA-catalyzed strand exchange can be divided into distinct reaction stages. First, the RecA filament must assemble on one of the participating DNA molecules. Assembly occurs on a molecule containing a region of ssDNA, such as an ssDNA tail. This RecA–ssDNA complex is the active form that participates in the search for a homology. During this search, RecA must “look” for base-pair complementarity between the DNA within the filament and a new DNA molecule. The structures of RecA with ssDNA and with dsDNA reveal



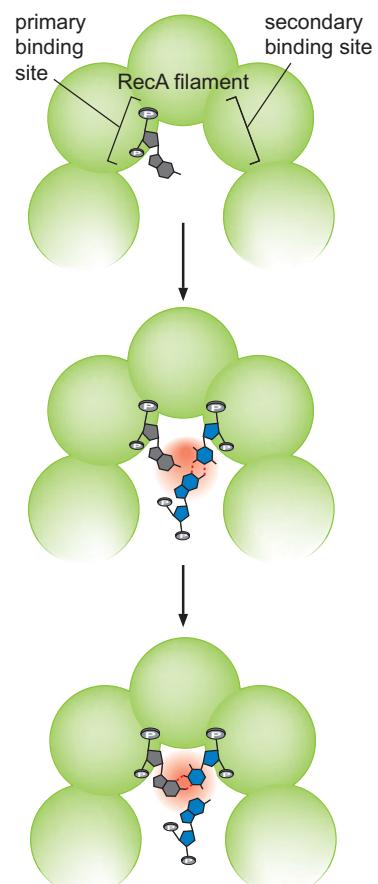
**FIGURE 11-10** Polarity of RecA assembly. Note that new subunits of RecA join the filament on the DNA 3' side to an existing subunit much faster than these subunits join on the 5' side. Because of this polarity of assembly, DNA molecules with 3' ssDNA extensions will be efficiently coated with RecA. In contrast, molecules with 5' ssDNA extensions would not serve as substrates for filament assembly.

varied DNA stretching and imply a mechanism for strand exchange (Fig. 11-11).

This homology search is promoted by RecA because the filament structure has two distinct DNA-binding sites: a primary site (bound by the first DNA molecule) and a secondary site (Fig. 11-12). This secondary DNA-binding site can be occupied by double-stranded DNA. Binding to this site is rapid, weak, transient, and—importantly—-independent of DNA sequence. In this way, the RecA filament can bind and rapidly “sample” huge stretches of DNA for sequence homology.



**FIGURE 11-11** DNA view of the strand-exchange reaction promoted by RecA. (a) dsDNA pairs with the RecA presynaptic filament that consists of RecA and ssDNA, to produce heteroduplex DNA–RecA and ssDNA. (b) The structures of the participating DNA molecules are shown: B-form DNA; ssDNA within the presynaptic filament; and randomly coiled ssDNA. (Adapted, with permission, from Kowalczykowski S.C. 2008. *Nature* **453**: 463, Fig. 1, p. 465. © Macmillan.)

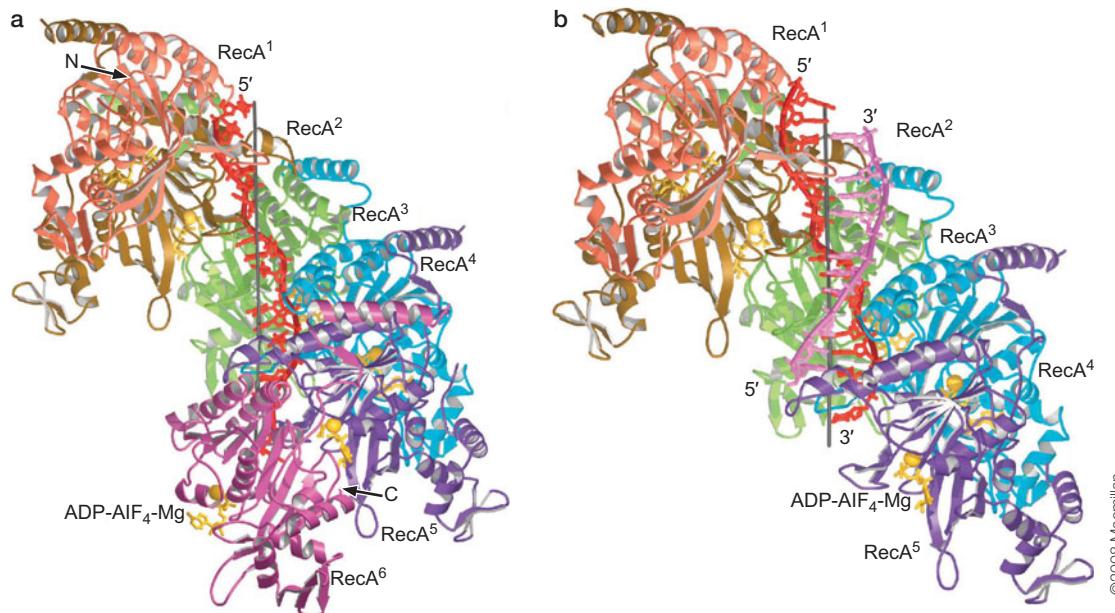


**FIGURE 11-12** Model of two steps in the search for homology and DNA strand exchange within the RecA filament. Here, the RecA filament is represented from a top-down view as in Fig. 11-9c. The incoming DNA duplex is shown in blue. (a) This panel shows a cross section of a single DNA strand bound to RecA protein. (b) DNA in the secondary site is tested for complementarity. (c) Base pairing between the strands is switched. (Adapted from Howard-Flanders et al. 1984. *Nature* **309**: 215–220.)

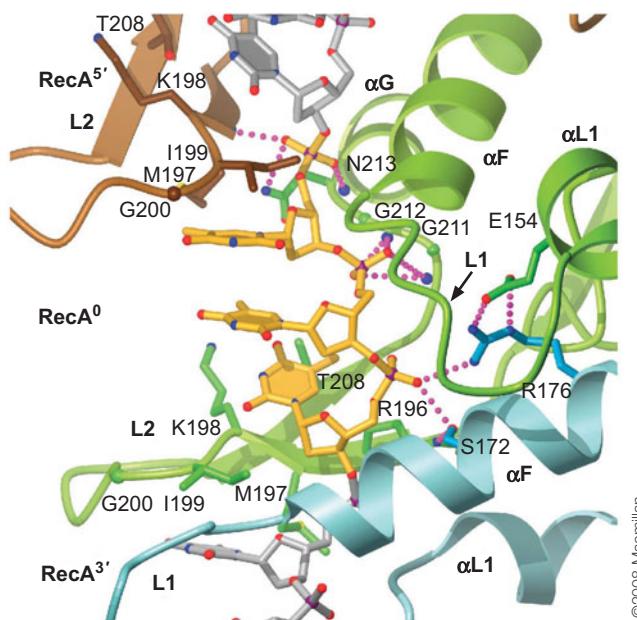
How does the RecA filament sense sequence homology? Details of this mechanism are not yet clear, but what we do know now, largely from structural studies, is how ssDNA and ATP bind to RecA to form a helical “presynaptic” filament. The ssDNA within the filament becomes underwound and stretched, a change that likely allows for more optimal Watson–Crick base pairing with dsDNA, to form a synaptic filament that samples for homology between the ssDNA and dsDNA (compare the structures shown in Fig. 11-13). Within the presynaptic filament, ssDNA binds with a stoichiometry of exactly three nucleotides per RecA, arranged in a B DNA-like conformation—recall from Chapter 4 that the repeating unit of the DNA structure is three nucleotides, a “nucleotide triplet” (see Fig. 11-14). After each three-nucleotide unit of B-like DNA in the presynaptic filament there is a large “step” before the next base; these large steps are principally responsible for the expended helix compared to naked DNA (Figs. 11-13 and 11-14).

The DNA in the secondary binding site is transiently opened and tested for complementarity with the ssDNA in the primary site. This “testing” occurs via base-pairing interactions, although it occurs initially without disrupting the global base pairing between the two strands of the DNA in the secondary site. In vitro experiments indicate that a sequence match of just 15 bp provides a sufficient signal to the RecA filament that a match has been found and thereby triggers strand exchange.

Once a region of base-pair complementarity is located, RecA promotes the formation of a stable complex with complete Watson–Crick hydrogen bonding between these two DNA molecules (Fig. 11-15a). The repeating unit is now a triplet of stacked base pairs (the base-pair triplet) that is quite similar to B-form DNA (Fig. 11-15b). This RecA-bound three-stranded structure is called a **joint molecule** and usually contains several hundred base pairs of hybrid DNA. It is within this joint molecule that the actual exchange of DNA strands occurs. The DNA strand in the primary binding site becomes base-paired with its complement in the DNA duplex bound



**FIGURE 11-13** RecA ssDNA and dsDNA structures. (a) Structure of the presynaptic nucleoprotein filament. (b) Structure of the postsynaptic nucleoprotein filament. (Reproduced, with permission, from Chen et al. 2008. *Nature* **453**: 489. Part a is Fig. 1A, p. 490; part b is Fig. 4A, p. 492. © Macmillan.)



**FIGURE 11-14** Close-up view of ssDNA in RecA. Each nucleotide triplet is bound by three RecA units. (Reproduced, with permission, from Chen et al. 2008. *Nature* 453: 489, Fig. 2, p. 491. © Macmillan.)

in the secondary site. Strand exchange thus requires the breaking of one set of base pairs and the formation of a new set of identical base pairs. Completion of strand exchange also requires that the two newly paired strands be intertwined to form a proper double helix. RecA binds preferentially to the DNA products after strand exchange has occurred, and it is this binding energy that actually drives the exchange reaction toward the new DNA configuration.

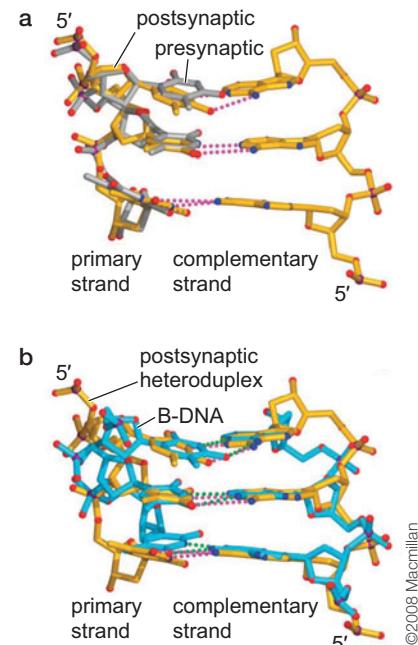
### RecA Homologs Are Present in All Organisms

Strand-exchange proteins of the RecA family are present in all forms of life. The best-characterized members are RecA from Eubacteria, RadA from Archaea, Rad51 and Dmc1 from Eukaryota, and the bacteriophage T4 UvsX protein. These proteins form filaments similar to those made by RecA (Fig. 11-16) and likely function in an analogous manner (although some features of the proteins are specifically tailored for their specific cellular roles and interaction partners). We discuss the roles of Rad51 and Dmc1 recombination in eukaryotic cells in a later section.

### The RuvAB Complex Specifically Recognizes Holliday Junctions and Promotes Branch Migration

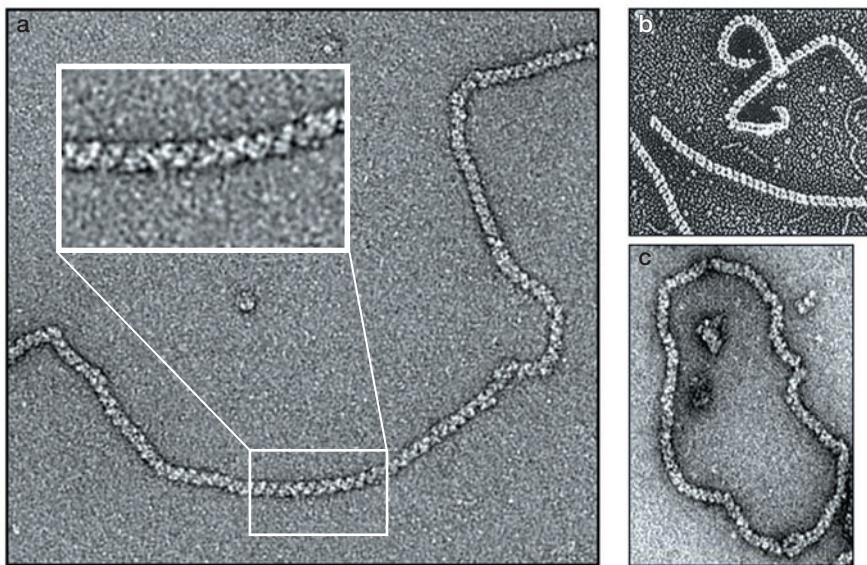
After the strand invasion step of recombination is complete, the two recombining DNA molecules are connected by a DNA branch known as a Holliday junction (see above). Movement of the site of this branch requires exchange of DNA base pairs between the two homologous DNA duplexes. Cells encode proteins that greatly stimulate the rate of branch migration.

RuvA protein is a Holliday junction–specific DNA-binding protein that recognizes the structure of the DNA junction, regardless of its specific DNA sequence (see Structural Tutorial 11-2). RuvA recognizes and binds to Holliday junctions and recruits the RuvB protein to this site. RuvB is a hexameric ATPase, similar to the hexameric helicases involved in DNA replication (see Chapter 9). The RuvB ATPase provides the energy to drive the

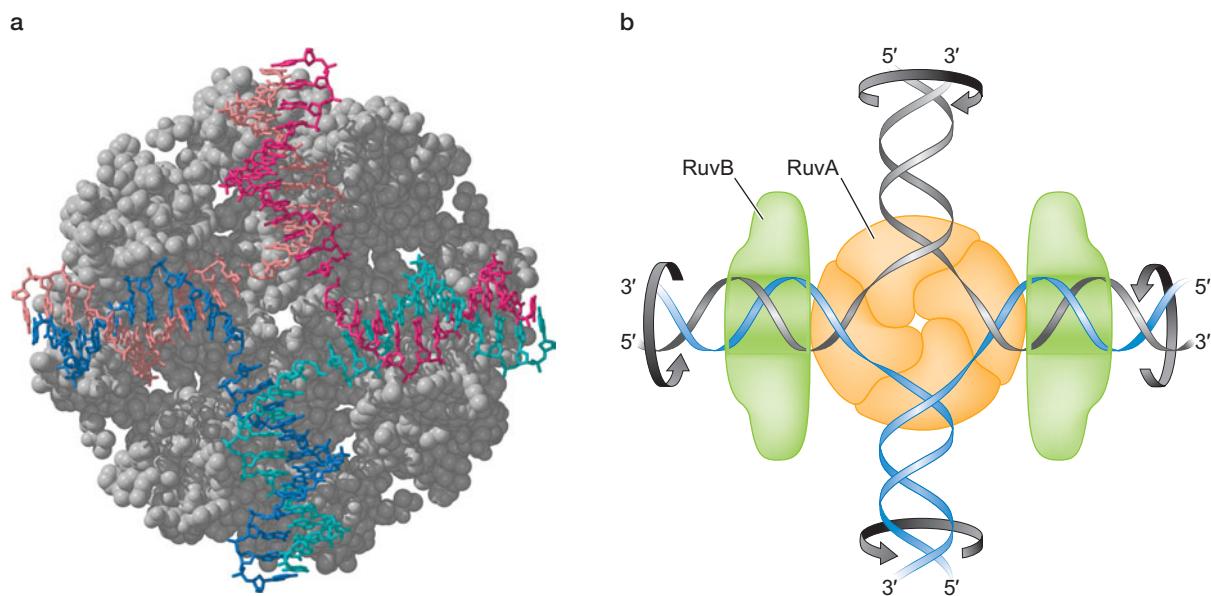


**FIGURE 11-15** DNA structure in the filament does not change very much. (a) The presynaptic nucleotide triplet (in gray) is superimposed with the postsynaptic base-pair triplet (yellow). Watson-Crick hydrogen bonds are indicated by dotted lines. (b) The postsynaptic base-pair triplet (yellow) is superimposed with B-type DNA (cyan). Watson-Crick hydrogen bonds are shown as dotted lines, colored magenta for the heteroduplex and green for B-DNA. (Reproduced, with permission, from Chen et al. 2008. *Nature* 453: 489, Fig. 5a,b, p. 491. © Macmillan.)

**FIGURE 11-16** RecA-like proteins in three branches of life. Nucleoprotein filaments are shown for (a) human Rad51, (b) *E. coli* RecA, and (c) *Archaeoglobus fulgidus* RadA proteins. The Rad51 and RecA proteins are also shown in Figure 11-8. Notice the similar helical structure of the filaments revealed by the stripes in these EM images. (Reprinted, with permission, from West S.C. et al. 2003. *Nat. Rev. Mol. Cell Biol.* **4**: 435–445. © Macmillan. Images provided by A. Stasiak, University of Lausanne, Switzerland.)



exchange of base pairs that moves the DNA branch. This energy is needed to move the branch rapidly and in one direction. Structural models for RuvAB complexes at a Holliday junction show how a tetramer of RuvA together with two hexamers of RuvB work together to power this DNA-exchange process (Fig. 11-17).



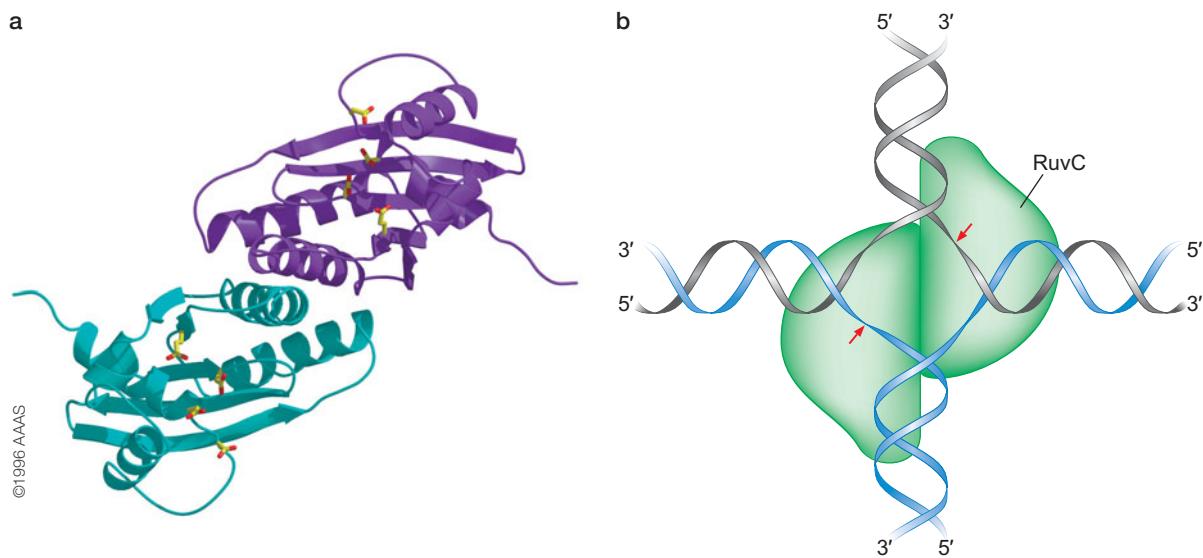
**FIGURE 11-17** High-resolution structure of RuvA and schematic model of the RuvAB complex bound to Holliday junction DNA. (a) The crystal structure of the RuvA tetramer shows the fourfold symmetry of the protein. (Ariyoshi M. et al. 2000. *Proc. Natl. Acad. Sci.* **97**: 8257–8262.) Image prepared with MolScript, BobScript, and Raster3D. (b) A schematic model of the crystal structure is shown with two RuvB hexamers. Notice how a tetramer of RuvA binds with four-fold symmetry to the junction. Two hexamers of RuvB bind on opposite sides of RuvA and function as a motor to pump DNA through the junction. The RuvB hexamers are shown in cross sections, so that the DNA threading through these complexes can be seen. (Redrawn from Yamada K. et al. 2002. *Mol. Cell* **10**: 671–681, Fig. 4.)

### RuvC Cleaves Specific DNA Strands at the Holliday Junction to Finish Recombination

Completion of recombination requires that the Holliday junction (or junctions) between the two recombining DNA molecules be resolved. In bacteria, the major Holliday junction resolving endonuclease is RuvC. RuvC was discovered and purified based on its ability to cut DNA junctions made by RecA in vitro. Evidence indicates that it functions in concert with RuvA and RuvB.

Resolution by RuvC occurs when RuvC recognizes the Holliday junction—likely in a complex with RuvA and RuvB—and specifically nicks two of the homologous DNA strands that have the same polarity. This cleavage results in DNA ends that terminate with 5'-phosphates and 3'-OH groups that can be directly joined by DNA ligase. Depending on which pair of strands is cleaved by RuvC, the resulting ligated recombination products will be of either the “splice” (crossover) or “patch” (noncrossover) type. The structure of RuvC and a model schematic proposing how it may interact with junction DNA are shown in Figure 11-18.

Despite recognizing a structure rather than a specific sequence, RuvC cleaves DNA with modest sequence specificity. Cleavage takes place only at sites conforming to the consensus 5'-A/T-T-T-G/C. Cleavage occurs after the second T in this sequence. Sequences with this consensus are found frequently in DNA, averaging once every 64 nucleotides. This modest sequence selectivity ensures that at least some branch migration occurs before resolution. Without this sequence selectivity, RuvC might simply cleave Holliday junctions as soon as they are formed, thereby restricting the region of DNA that participates in strand exchange.



**FIGURE 11-18** High-resolution structure of the RuvC resolvase and schematic model of the RuvC dimer bound to Holliday junction DNA. (a) The crystal structure of the RuvC protein. (Ariyoshi M. et al. 1994. *Cell* 78: 1063–1072.) Image prepared with MolScript, BobScript, and Raster3D. (b) Model for binding of a RuvC dimer to a Holliday junction. Notice how, in this model, a dimer of RuvC can bind the Holliday junction and introduce symmetrical cleavages into the two identical DNA strands. (Adapted, with permission, from Rafferty J.B. et al. 1996. *Science* 274: 415–421, Fig. 1b. © AAAS.)

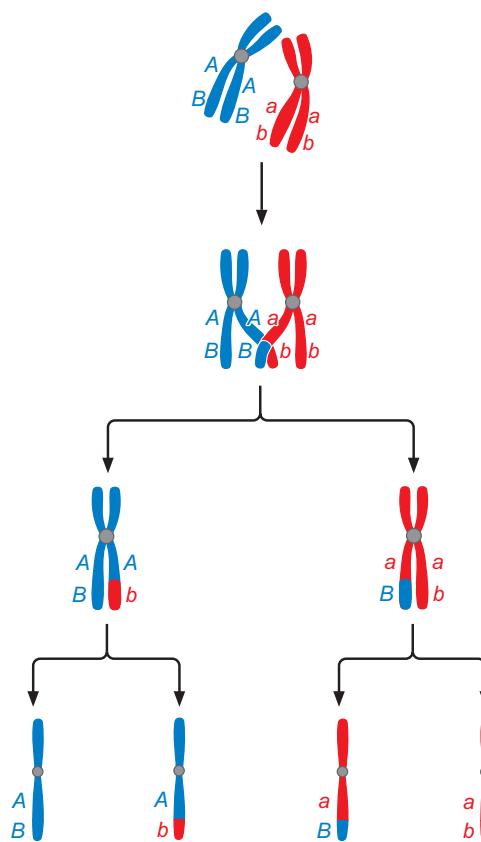
## HOMOLOGOUS RECOMBINATION IN EUKARYOTES

### Homologous Recombination Has Additional Functions in Eukaryotes

As we have just described, homologous recombination in bacteria is required to repair DSBs in DNA, to restart collapsed replication forks, and to allow a cell's chromosomal DNA to recombine with DNA that enters via phage infection or conjugation. Homologous recombination is also required for DNA repair and the restarting of collapsed replication forks in eukaryotic cells. This requirement is illustrated by the fact that cells with defects in the proteins that promote recombination are hypersensitive to DNA-damaging agents, especially those that break DNA strands. Furthermore, animals carrying mutations that interfere with homologous recombination are predisposed to certain types of cancer as well as conditions of premature aging. However, as we discuss below, homologous recombination plays important additional roles in eukaryotic organisms. Most importantly, homologous recombination is critical for meiosis. During meiosis, homologous recombination is *required* for proper chromosome pairing and thus for maintaining the integrity of the genome. This recombination also reshuffles genes between the parental chromosomes, ensuring variation in the sets of genes passed to the next generation.

### Homologous Recombination Is Required for Chromosome Segregation during Meiosis

As we saw in Chapter 9, **meiosis** involves two rounds of nuclear division, resulting in a reduction of the DNA content from the normal content of diploid cells ( $2N$ ) to the content present in gametes ( $1N$ ). Figure 11-19 schematically



**FIGURE 11-19** DNA dynamics during meiosis. Here, only one type of chromosome is shown for clarity. The two homologs are shown, in red and blue, after they have been duplicated by a round of DNA replication. Homologous recombination is required to pair these homologous chromosomes in preparation for the first nuclear division. This recombination can also lead to crossing over, as is shown here between the *A* and *B* genes.

shows how the chromosomes are configured during these two division cycles. Before division, the cell has two copies of each chromosome (the **homologs**), one each that was inherited from its two parents. During S phase, these chromosomes are replicated to give a total DNA content of 4N (Fig. 11-20). The products of replication—that is, the **sister chromatids**—stay together. Then, in preparation for the first nuclear division, these *duplicated homologous chromosomes must pair* and align at the center of the cell. It is this pairing of homologs that requires homologous recombination (Fig. 11-19). These events are carefully timed. Recombination must be complete before the first nuclear division to allow the homologs to properly align and then separate. During this process, sister chromatids remain paired (Fig. 11-20; see also Chapter 8, Fig. 8-16). Then, in the second nuclear division, it is the sister chromatids that separate. The products of this division are the four gametes, each with one copy of each chromosome (i.e., the 1N DNA content).

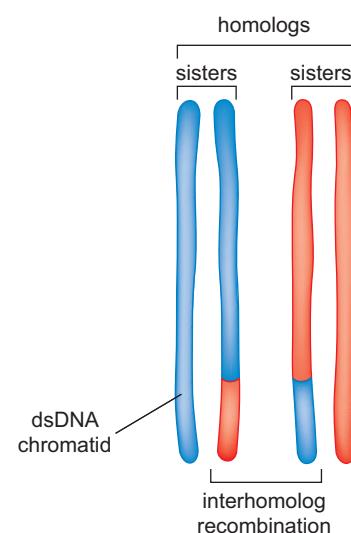
In the absence of recombination, chromosomes often fail to align properly for the first meiotic division, and, as a result, there is a high incidence of chromosome loss. This improper segregation of chromosomes, called **nondisjunction**, leads to a large number of gametes without the correct chromosome complement. Gametes with either too few or too many chromosomes cannot develop properly once fertilized; thus, a failure in homologous recombination is often reflected in poor fertility. The homologous recombination events that occur during meiosis are called **meiotic recombination**.

Meiotic recombination also frequently gives rise to crossing over between genes on the two homologous parental chromosomes. This genetic exchange, shown schematically in Figure 11-20, can be observed cytologically (Fig. 11-21a). An important consequence is that the alleles present on the parental DNA molecules are reassorted for the next generation.

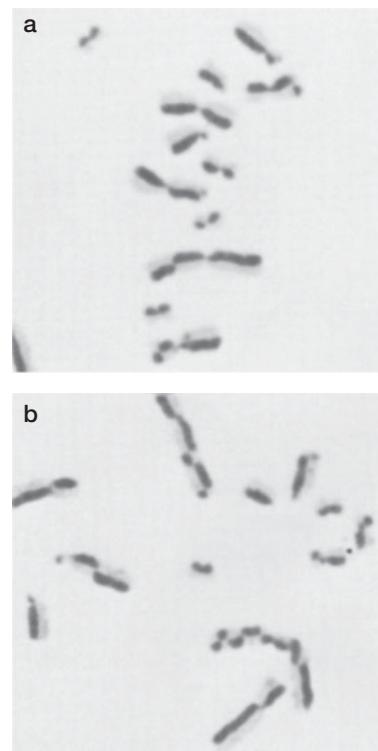
### Programmed Generation of Double-Stranded DNA Breaks Occurs during Meiosis

The developmental program needed for cells to complete meiosis successfully involves turning on the expression of many genes that are not needed during normal growth. One of these is *SPO11*. This gene encodes a protein that introduces DSBs in chromosomal DNA to initiate meiotic recombination.

The Spo11 protein cuts the DNA at many chromosomal locations, with little sequence selectivity, but at a very specific time during meiosis. Spo11-mediated DNA cleavage occurs right around the time when the replicated homologous chromosomes start to pair. Spo11 cut sites, although frequent, are not randomly distributed along the DNA. Rather, the cut sites are located most commonly in chromosomal regions that are not tightly packed with nucleosomes, such as promoters controlling gene transcription (see Chapters 8 and 19). Regions of DNA that experience a high frequency of DSBs

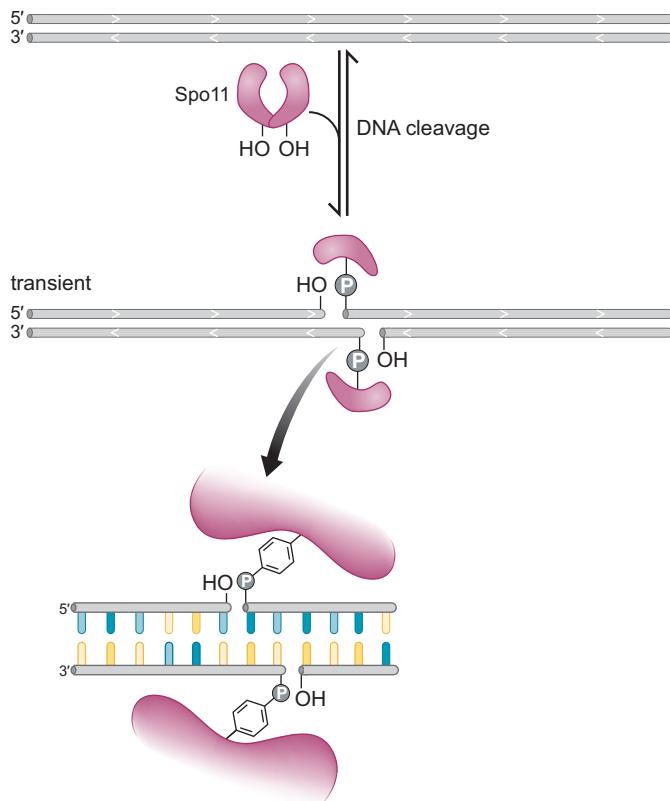


**FIGURE 11-20** Meiotic recombination between homologous chromatids. Each structure shown is a replicated, double-stranded DNA molecule called a chromatid. The pairs are called sister chromatids, and recombination that occurs between non-sister pairs is mediated by Dmc1 (see Fig. 11-19).



**FIGURE 11-21** Cytological view of crossing over. Reciprocal crossing over directly visualized in hamster cells in tissue culture. Chromosomes whose DNA contains bromodeoxyuridine in place of thymidine in both strands appear light after treatment with Giemsa stain, whereas those containing DNA substituted in only one strand appear dark. After two generations of growth in bromodeoxyuridine, one newly replicated chromatid has only one of its strands substituted, whereas its sister has both substituted. Thus, sister chromatids can be distinguished by staining. Then, crossovers are easily detected as alternating lengths of light and dark (a). Similar recombinant chromosomes are also seen when mitotically growing cells are treated with a DNA-damaging agent (b). (Courtesy of Sheldon Wolff and Jody Bodycote.)

**FIGURE 11-22** Mechanism of cleavage by Spo11. The OH group of a tyrosine in the Spo11 protein attacks the DNA to form a covalent protein–DNA linkage. Two subunits of Spo11 are required to generate a double-stranded DNA break, one to attack each of the two DNA strands. Note that because of this cleavage mechanism, the DSB can be resealed by the simple reversal of the cleavage reaction.



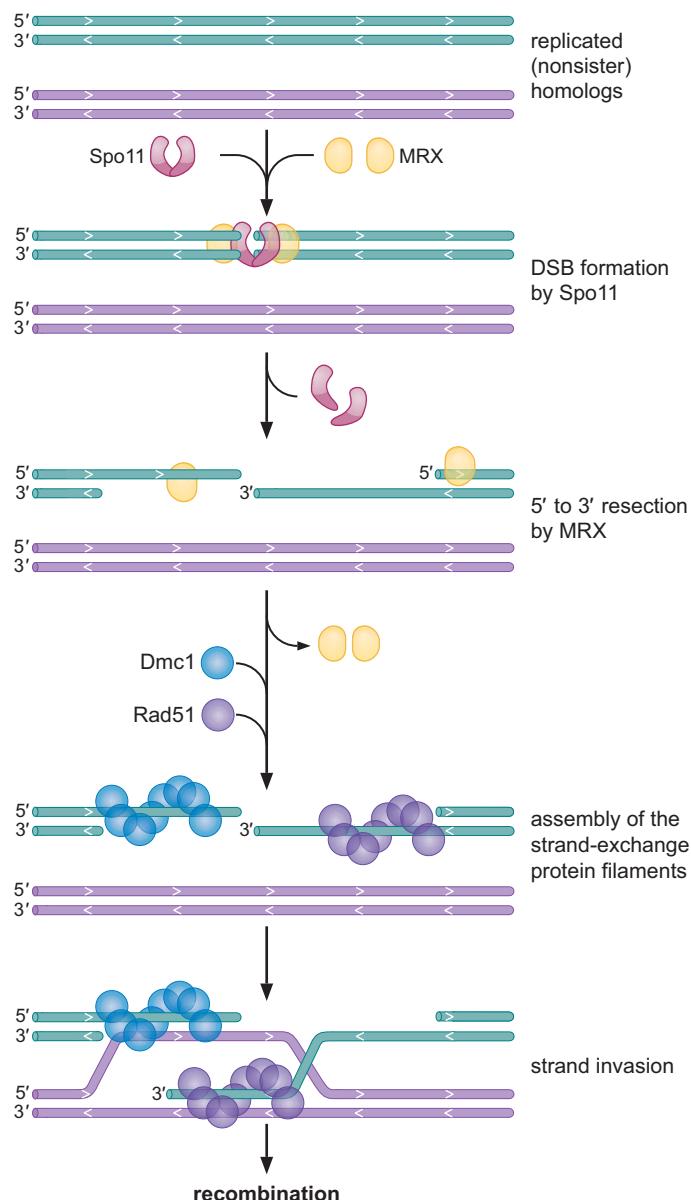
also show a high frequency of recombination. Thus, the most commonly used Spo11 DNA cleavage sites, like Chi sites, are hot spots for recombination.

The mechanism of Spo11 DNA cleavage is as follows: A specific tyrosine side chain in the Spo11 protein attacks the phosphodiester backbone to cut the DNA and generate a covalent complex between the protein and the severed DNA strand (Fig. 11-22). Two subunits of Spo11 cleave the DNA two nucleotides apart on the two DNA strands to make a staggered DSB. Spo11 shares this DNA cleavage mechanism with the DNA topoisomerases and the site-specific recombinases (see Chapters 4 and 12). Protein sequence comparisons reveal that Spo11 appears to be a distant cousin of these enzymes.

The fact that Spo11 cleavage involves a covalent protein–DNA complex has two consequences: First, the 5' ends of the DNA at the site of Spo11 cleavage are covalently bound to the enzyme. It is these Spo11-linked 5' DNA ends that are the initial sites of DNA processing to create the ssDNA tails required for assembly of RecA-like proteins and initiation of DNA strand invasion (see below). Second, the energy of the cleaved DNA phosphodiester bond is stored in the bound protein–DNA linkage, and thus the DNA strands can be resealed by a simple reversal of the cleavage reaction (see Fig. 12-5 for chemical mechanism). This resealing can occur when cells receive a signal to stop proceeding with meiosis.

### MRX Protein Processes the Cleaved DNA Ends for Assembly of the RecA-Like Strand-Exchange Proteins

The DNA at the site of the Spo11-catalyzed DSB is processed to generate single-strand regions needed for assembly of the RecA-like strand-exchange



**FIGURE 11-23** Overview of meiotic recombination pathway. Formation of the DSBs during meiosis requires the presence of both Spo11 and the MRX complex. This observation suggests that DSB formation and subsequent strand processing are normally coupled by the coordinated action of several proteins. MRX protein is responsible for resection of the 5'-ending strands at the break site. The strand-exchange proteins Dmc1 and Rad51 then assemble on the ssDNA tails. Both proteins participate in recombination, but how they work together is not known. They are shown forming separate filaments for clarity. (Redrawn, with permission, from Lichten M. 2001. *Curr. Biol.* 11: R253–R256, Fig. 2. © Elsevier.)

proteins. As was observed in the RecBCD pathway from bacteria, this processing generates long segments of ssDNA that terminate in 3' ends (Fig. 11-23). During meiotic recombination, the MRX–enzyme complex is responsible for this DNA-processing event. This complex, although not homologous to RecBCD, is also a multi-subunit DNA nuclease. MRX is composed of protein subunits called Mre11, Rad50, and Xrs2; the first letters of these subunits give the complex its name.

Processing of the DNA at the break site occurs exclusively on the DNA strand that terminates with a 5' end—that is, the strands covalently attached to the Spo11 protein (as described above). The strands terminating with 3' ends are not degraded. This DNA-processing reaction is therefore called 5'-to-3' resection. The MRX-dependent 5'-to-3' resection generates the long ssDNA tails with 3' ends that are often 1 kb or longer. The MRX complex is also thought to remove the DNA-linked Spo11.

### Dmc1 Is a RecA-Like Protein That Specifically Functions in Meiotic Recombination

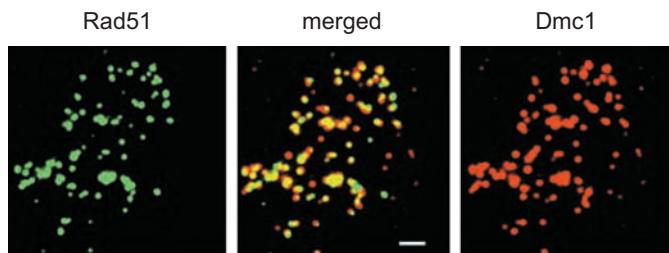
Eukaryotes encode two well-characterized homologs of the bacterial RecA protein: Rad51 and Dmc1. Both proteins function in meiotic recombination. Whereas Rad51 is widely expressed in cells dividing mitotically and meiotically, Dmc1 is expressed only as cells enter meiosis.

Strand exchange during meiosis occurs between a particular type of homologous DNA partner. Recall that meiotic recombination occurs at a time when there are four complete, double-stranded DNA molecules representing each chromosome: the two homologs each of which have been copied to generate two sister chromatids (see Fig. 11-20). Although the two homologs likely contain small sequence differences and carry distinct alleles for various genes, the majority of the DNA sequence among these four copies of the chromosome will be identical. Interestingly, Dmc1-dependent recombination is preferentially between the *non-sister* homologous chromatids, rather than between the sisters (see Fig. 11-20). Although the mechanistic basis of this selectivity is unknown, there is a clear biological rationale: Meiotic recombination promotes interhomolog connections to assist alignment of the chromosomes for division.

### Many Proteins Function Together to Promote Meiotic Recombination

As we have described, proteins involved in the critical stages of DSB formation, DNA processing to generate 3' ssDNA tails, and strand exchange during meiotic recombination have been identified and characterized. Genetic experiments indicate that many additional proteins also participate in this process. In addition, many proteins appear to interact with the known recombination enzymes, and it seems likely that these proteins function in the context of a large multicomponent complex. These large protein–DNA complexes, known as **recombination factories**, can be visualized in cells. For example, the colocalization of Rad51 and Dmc1 to these factories during meiosis is shown in Figure 11-24.

Various other proteins have been shown to be involved with Rad51 to help promote recombination and DSB repair. Rad52 is another essential recombination protein that interacts with Rad51. Rad52 functions to promote assembly of Rad51 DNA filaments, the active form of Rad51. It does this by antagonizing the action of RPA, the major ssDNA-binding protein present in eukaryotic cells. In this respect, Rad52 shares an activity with the *E. coli* RecBCD protein, which, as we learned, helps RecA load onto



**FIGURE 11-24** Colocalizations of the Rad51 and Dmc1 proteins to “recombination factories” in cells undergoing meiosis. Proteins were detected by immunostaining with fluorescently labeled antibodies to Rad51 (green) and Dmc1 (red). When the two proteins colocalize, the merged image appears yellow. (Reprinted, with permission, from Shinohara M. et al. 2000. *Proc. Natl. Acad. Sci.* **97**: 10814–10819, Fig. 1A. © National Academy of Sciences.)

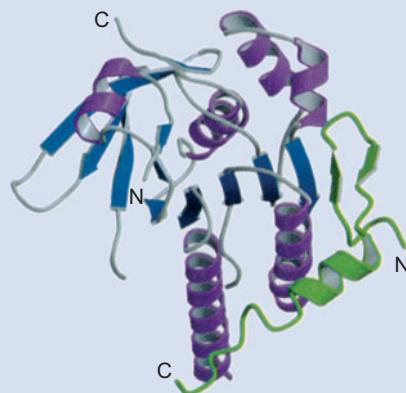
### MEDICAL CONNECTIONS

#### Box 11-2 The Product of the Tumor Suppressor Gene *BRCA2* Interacts with Rad51 Protein and Controls Genome Stability

The *BRCA2* gene is important for maintaining genome stability. In humans, mutations in *BRCA2* are thought to be responsible for half the familial breast cancers. This cancer predisposition appears attributable, at least in part, to a direct role of the *BRCA2* protein in Rad51-mediated DSB repair. When cells are subjected to agents that damage DNA, Rad51 foci assemble as an apparent prerequisite to activation of the repair functions. Cells with defects in *BRCA2* fail to assemble these foci in the nuclei of the damaged cells and have a corresponding defect

in repair of DNA breaks. *BRCA2* makes direct protein–protein contacts with Rad51 (see Box 11-2 Fig. 1), and these interactions are likely important for recruiting Rad51 to the proper cellular location for repair, as well as modulating the activity of the protein. The strong phenotype associated with the *BRCA2* mutations therefore illustrates the central importance of DSB repair and Rad51-dependent homologous recombination in eukaryotes, including humans.

**BOX 11-2 FIGURE 1** Structure of the complex between Rad51 and BRCA2 repeat motif. Various biochemical and structural studies have shown that specific regions within *BRCA2*, conserved repeat sequences known as the BRC motifs, are the major sites of interaction with Rad51. One of these motifs, BRC repeat 4 (BRC4), has been shown to bind Rad51 with high affinity. Structural analysis has revealed more precisely how BRC4 forms a complex with Rad51. In this view, the  $\alpha$  helices of Rad51 are shown in purple and the  $\beta$  strands in blue, whereas the peptide BRC repeat sequence is shown in green. The amino and carboxyl termini are marked for each sequence. (Reprinted, with permission, from Pellegrini L. et al. 2002. *Nature* **420**: 287–293, Fig. 1a. © Macmillan.)



ssDNA that would otherwise have been bound by SSB. Rad52 protein also promotes the annealing and base pairing of complementary ssDNA molecules, and this activity may also play a role in the strand-pairing reactions that occur during initiation of recombination. The product of the *BRCA2* gene also participates in Rad51-mediated DSB repair (see Box 11-2, The Product of the Tumor Suppressor Gene *BRCA2* Interacts with Rad51 Protein and Controls Genome Stability).

By analogy with bacteria, we expect that eukaryotic cells encode proteins that promote the branch migration and Holliday junction resolution steps of recombination. In fact, enzymes capable of promoting these reactions are being identified and characterized. A complex containing a Rad51-like protein, called Rad51C, and a second protein, called XRCC3, has been found to contain Holliday junction resolvase activity. Similarly, members of a family of enzymes—the RecQ helicases—play critical roles in homologous recombination during DSB repair and are likely also involved in meiosis. In humans, for example, an alternative process to resolve a double Holliday junction involves a RecQ helicase acting in concert with a topoisomerase. This mechanism, called double-junction dissolution, prevents the exchange of flanking sequences. Three of these helicases found in humans (BLM, WRN, RTS/RECQL) are associated with Bloom, Werner, and Rothmund–Thomson syndromes, respectively, which cause predisposition to premature aging and/or tumorigenesis (see Box 11-3, Proteins Associated with Premature Aging and Cancer Promote an Alternative Pathway for Holliday Junction Processing).

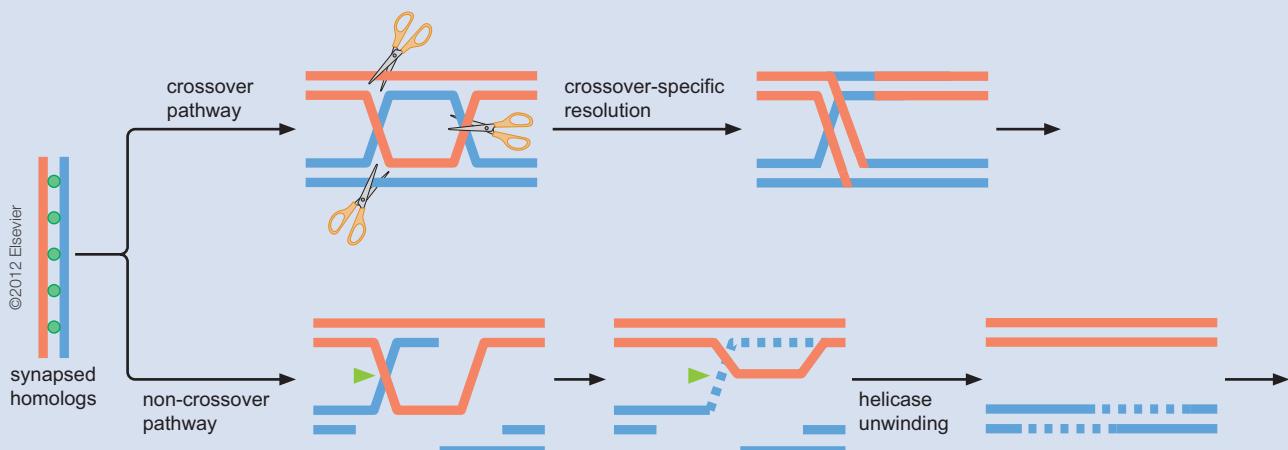
As we have seen, meiotic recombination aligns homologous chromosomes and promotes genetic exchange between them. These recombination

## ► MEDICAL CONNECTIONS

**Box 11-3** Proteins Associated with Premature Aging and Cancer Promote an Alternative Pathway for Holliday Junction Processing

The RecQ DNA helicases, conserved from bacteria to humans, play important roles in early and late stages of homologous recombination. Specifically, these helicases can process and edit recombination intermediates, often resulting in the collapse of joint molecules before establishment of the double Holliday junction intermediate. As a result, the helicases promote non-

crossover recombination at the expense of the crossover class of events (Box 11-3 Fig. 1). In humans, seriously premature aging and assorted cancers are associated with loss-of-function mutations in the genes encoding three of these helicases—WRN, BLM, and RTS/RECQL (see Box 11-3 Table 1).



**BOX 11-3 FIGURE 1** Crossover and non-crossover pathways for recombination. In the crossover pathway (upper pathway shown here), the Holliday junction resolvase (shown as scissors) assembles at the junction and cleaves asymmetrically to produce the crossover products. In contrast, in the non-crossover pathway (lower pathway), the RecQ-family helicase (noted by green arrow) promotes synthesis-dependent strand annealing and resolution. The mechanism of action appears to be that the helicase activity of these enzymes takes apart the joint molecules initially made by the strand exchange proteins. This action can also slide the D-loop along the DNA and allow the invading strand to be extended by DNA polymerase before collapse of the joint molecule. (Adapted and modified, with permission, from Zakharyevich K. et al. 2012. *Cell* 149: 334–347, Fig. 7. © Elsevier.)

**BOX 11-3 TABLE 1** Clinical Features of RecQ Disorders

Syndrome (Gene)	Main Clinical Features	Cancer Predisposition
Bloom syndrome (BLM)	Dwarfism, beaked nose, narrow face, pigmentation, redness, and dilated blood vessels in skin, mental retardation, type II diabetes, immunodeficiency, lung problems, low or no fertility	Early onset with normal distribution of tissue and type
Werner syndrome (WRN)	Bilateral cataracts, hoarseness, skin alterations, thin limbs, premature gray/loss of hair, pinched facial features, short stature, osteoporosis, hypogonadism, diabetes, soft tissue calcification	Early onset of primary sarcomas and mesenchymal tumors
Rothmund–Thomson syndrome (RECQL)	Poikiloderma, juvenile cataracts, growth retardation, skeletal dysplasia, sparse scalp hair, hypogonadism	Early onset of osteosarcomas

Adapted, with permission, from Bernstein K.A. et al. 2010. *Annu. Rev. Genet.* 44: 393–417, Table 1, p. 395. © Annual Reviews.

reactions often lead to crossing over between the parental chromosomes. Recall, however, that depending on how the Holliday junctions in the recombination intermediates are resolved, recombination via the DSB-repair pathway can also give rise to noncrossover products (see above). These events may provide the essential chromosome-pairing function needed for a successful meiotic division, yet leave no detectable change in the genetic makeup of the chromosomes.

But even noncrossover recombination can have genetic consequences, such as giving rise to a **gene conversion** event. Gene conversion happens

when an allele of a gene is lost and replaced by an alternative allele. Examples of how gene conversion can occur both in mitotically growing cells and during meiosis are described in the following sections.

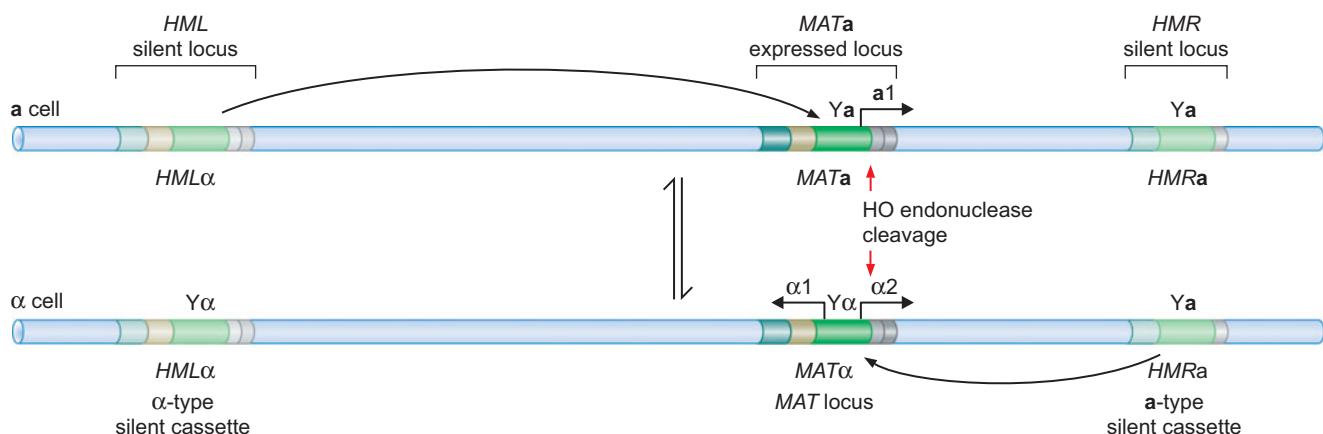
## MATING-TYPE SWITCHING

In addition to promoting DNA pairing, DNA repair, and genetic exchange, homologous recombination can also serve to change the DNA sequence at a specific chromosomal location. This type of recombination is sometimes used to regulate gene expression. For example, recombination controls the mating type of the budding yeast *S. cerevisiae* by switching which mating-type genes are present at a specific location that is being expressed in that organism's genome.

*S. cerevisiae* is a single-cell eukaryote that can exist as any of three different cell types (see Appendix 1). Haploid *S. cerevisiae* cells can be either of two mating types, **a** or **α**. In addition, when an **a** cell and an **α** cell come into close proximity, they can fuse (i.e., “mate”) to form an **a/α** diploid cell. The **a/α** cell may then go through meiosis to form two haploid **a** cells and two haploid **α** cells.

The mating-type genes encode transcriptional regulators. These regulators control expression of target genes whose products define each cell type. The mating-type genes expressed in a given cell are those found at the mating-type locus (**MAT locus**) in that cell (Fig. 11-25). Thus, in **a** cells, the **a1** gene is present at the **MAT** locus, whereas in **α** cells, the **α1** and **α2** genes are present at the **MAT** locus. In the diploid cell, both sets of mating-type control genes are expressed. The regulators encoded by the mating-type genes, together with others found in all three cell types, act in various combinations to ensure that the correct pattern of genes is expressed in each cell type (see Chapter 19).

Cells can switch their mating type by recombination as we now describe. In addition to the **a** or **α** genes present at the **MAT** locus in each cell, there is an additional copy of both the **a** and **α** genes present (but not expressed)



**FIGURE 11-25** Genetic loci encoding mating-type information. Although chromosome III carries three mating-type loci, only the genes at the **MAT** locus are expressed. **HML** encodes a silent copy of the **α** genes, whereas **HMR** encodes a silent copy of the **a** genes. When recombination occurs between **MAT** and **HML**, **a** cells switch to **α** cells. When recombination occurs between **MAT** and **HMR**, **α** cells switch to **a** cells. (Adapted, with permission, from Haber J.E. 1998. *Annu. Rev. Genet.* 32: 561–599, Fig. 3. © Annual Reviews.)

elsewhere in the genome. These additional silent copies are found at loci called *HMR* and *HML* (Fig. 11-25).

These *HMR* and *HML* loci are therefore known as **silent cassettes**. Their function is to provide a “storehouse” of genetic information that can be used to switch a cell’s mating type. This switch requires the transfer of genetic information from the *HM* sites to the *MAT* locus via homologous recombination.

### Mating-Type Switching Is Initiated by a Site-Specific Double-Strand Break

Mating-type switching is initiated by the introduction of a DSB at the *MAT* locus. This reaction is performed by a specialized DNA-cleaving enzyme, called the **HO endonuclease**. Expression of the HO gene is tightly regulated to ensure that switching occurs only when it should. The mechanisms responsible for this regulation are discussed in Chapter 19. HO is a sequence-specific endonuclease; the only sites in the yeast chromosome that carry HO recognition sequences are the mating-type loci. HO cutting introduces a staggered break in the chromosome. In contrast to Spo11 cleavage, HO simply hydrolyzes the DNA and does not remain covalently linked to the cut strands.

5'-to-3' resection of the DNA at the site of the HO-induced break occurs by the same mechanism used during meiotic recombination. Thus, resection depends on the MRX–protein complex and is specific for the strands that terminate with 5' ends. In contrast, the strands terminating with 3' ends are very stable and not subjected to nuclease digestion. Once the long 3' ssDNA tails have been generated, this DNA becomes coated by the Rad51 and Rad52 proteins (as well as other proteins that help the assembly of the recombinogenic protein–DNA complex). These Rad51 protein–coated strands then search for homologous chromosomal regions to initiate strand invasion and genetic exchange.

Mating-type switching is unidirectional. That is, sequence information (although not the actual DNA segment) is “moved” to the *MAT* locus, from *HMR* and *HML*, but information never “goes” in the other direction. Thus, the cut *MAT* locus is always the “recipient” partner during recombination, and the *HMR* and *HML* sites remain unchanged by the recombination process. This directionality stems from the fact that HO endonuclease cannot cleave its recognition sequence at either *HML* or *HMR* because the chromatin structure renders these sites inaccessible to this enzyme.

The Rad51-coated 3' ssDNA tails from the *MAT* locus “choose” the DNA at either the *HMR* or *HML* locus for strand invasion. If the DNA sequence at *MAT* is **a**, then invasion will occur with *HML*, which carries the “storage” copy of the  **$\alpha$**  sequences. In contrast, if the  **$\alpha$**  genes are present at *MAT*, then invasion occurs with *HMR*, the locus that carries the stored **a** sequences. After recombination, the genetic information that was at the chosen *HM* loci is present at the *MAT* loci as well. This genetic change occurs without a reciprocal swap of information from *MAT* to the *HR* loci. This type of nonreciprocal recombination event is a specialized example of gene conversion.

### Mating-Type Switching Is a Gene Conversion Event and Not Associated with Crossing Over

Although the DSB-repair pathway could explain the mechanism of mating-type switch recombination, substantial experimental evidence indicates that after the strand invasion step, this recombination pathway diverges from the DSB-repair mechanism. One hint that the mechanism is distinct is that the crossover class of recombination products is never observed

during mating-type switching. Recall that in the DSB-repair pathway, resolution of the Holliday junction intermediates gives two classes of products: the splice, or crossover class, and the patch, or noncrossover, class (see Fig. 11-3). According to the DSB-repair model, these two types of products are predicted to occur at a similar frequency, yet, in mating-type switching, crossover products are never observed. Therefore, models for recombination that do not involve resolution of Holliday junction intermediates better explain mating-type switching.

To explain gene conversion without crossing over, a new recombination model termed **synthesis-dependent strand annealing (SDSA)** has been proposed. Figure 11-26 shows how mating-type switching can occur using this mechanism. The initiating event is, as we saw above, the introduction of a DSB at the recombination site (Fig. 11-26a). After 5'-to-3' resection and strand invasion (Fig. 11-26b,c), the invading 3' end serves as the primer to initiate new DNA synthesis at a region of homology flanking Ya and continue copying the Ya sequence (Fig. 11-26c,d). Remarkably, in contrast to what occurs during the DSB-repair pathway, a complete replication fork is assembled at this site.

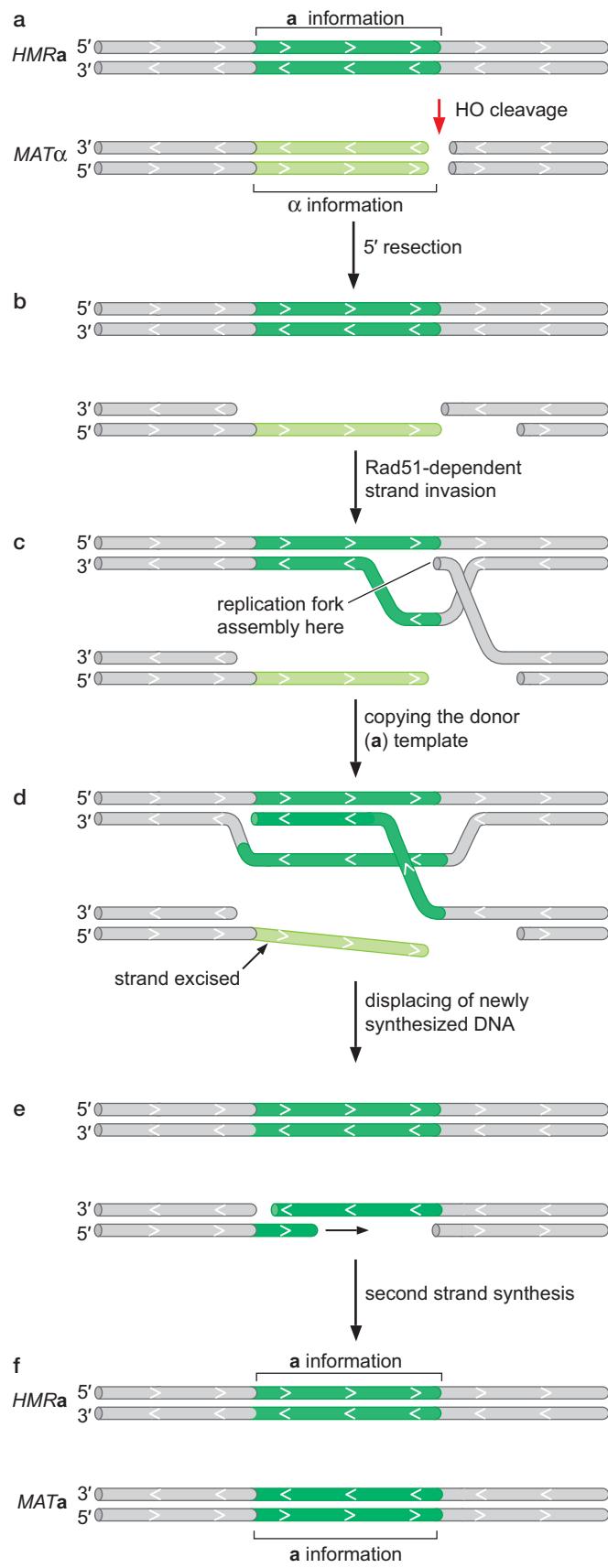
In contrast to normal DNA replication, however, the newly synthesized strand is displaced from its template and anneals with the second resected 5' end. Once this annealing step occurs, the corresponding long 3' tail (bottom strand in Fig. 11-26d) is clipped off by an endonuclease, and the new 3' end is used to prime and extend and copy the second strand of Ya sequences (Fig. 11-26e,f). As a result, a new double-stranded DNA segment is synthesized, joined to the DNA site that was originally cut by HO, and resected by MRX. This new segment has the sequence of the DNA segment used as the template (*HMRa* in Fig. 11-26).

Thus, the newly synthesized DNA—an exact copy of the information in the partner DNA molecule—replaces the information that was originally present. This mechanism nicely explains how gene conversion occurs without the need to cleave a Holliday junction. With this model, the absence of crossover products during mating-type recombination is no longer a mystery.

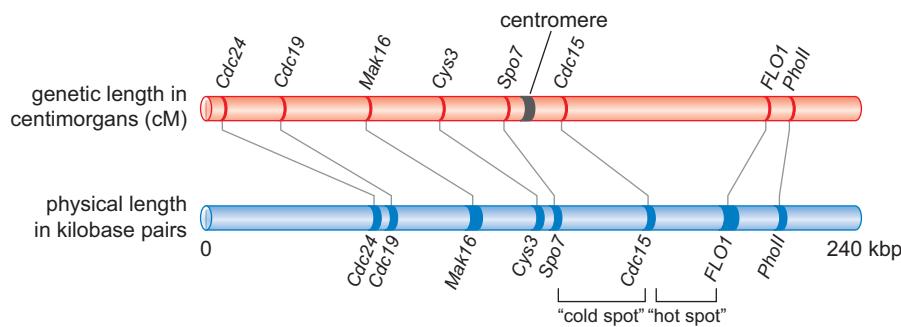
## GENETIC CONSEQUENCES OF THE MECHANISM OF HOMOLOGOUS RECOMBINATION

As we discussed at the beginning of this chapter, initial models for the mechanism of homologous recombination were formulated largely to explain the genetic consequences of the process. Now that the basic steps involved in recombination are understood, it is useful to review how the process of homologous recombination alters DNA molecules and thereby generates specific genetic changes.

A central feature of homologous recombination is that it can occur between any two regions of DNA, regardless of the sequence, provided these regions are sufficiently similar. We now understand why this is true; none of the steps in homologous recombination require recognition of a highly specific DNA sequence. For steps that have some sequence preference (such as the transformation of RecBCD by Chi sites and DNA cleavage by RuvC protein), the preferred sequences are very common. The committed step during recombination between two DNA molecules occurs when a strand-exchange protein of the RecA family successfully pairs the molecules, a process dictated only by the normal capacity of DNA strands to form proper base pairs.



**FIGURE 11-26** Recombination model for mating-type switching: Synthesis-dependent strand annealing (SDSA). The sequence of steps leading to gene conversion at the *MAT* locus is shown (see text for details). The *HMR* and *MAT* regions are shown in green; the region of *HMR* encoding the *a* information is represented in dark green; and the region of *MAT* encoding the  $\alpha$  information is shown in lime green. Upon completion of the process of SDSA, the  $\alpha$  region originally present at *MAT* has been replaced by (converted to) the *a* information present in the *HMR* region.



**FIGURE 11-27** Comparison of the genetic and physical maps of a typical region of a yeast chromosome. Markers show the location of various genes. Notice in the region between *Spo7* and *Cdc15* that the genetic map is contracted because of a low frequency of crossing over. In contrast, in the region between *Cdc15* and *FLO1*, the genetic map is expanded because of a high frequency of crossing over. (Adapted, with permission, from Alberts B. et al. 2002. *Molecular biology of the cell*, 4th ed., p. 1138, Fig. 20-14. © Garland Science/Taylor & Francis LLC.)

A corollary of the fact that recombination is generally independent of sequence is that the frequency of recombination between any two genes is generally proportional to the distance between those genes. This proportionality is observed because regions of DNA are, in general, equally likely to be used to initiate a successful recombination event. This fundamental aspect of homologous recombination is what makes it possible to use recombination frequencies to generate useful genetic maps that display the order and spacing of genes along a chromosome.

Distortions in genetic maps compared with physical maps occur when a region of DNA does not have the “average” probability of participating in recombination (Fig. 11-27). Regions with a higher than average probability are “hot spots,” whereas regions that participate less commonly than an average segment are “cold.” Therefore, two genes that have a hot spot between them appear in a genetic map to be farther apart than is true in a physical map of the same region. In contrast, genes separated by a “cold” interval appear by genetic mapping to be closer together than is true from their physical distance. We have encountered two examples for the molecular explanation of hot and cold spots in chromosomes. Regions near Chi sites and Spo11 cleavage sites have a higher than average probability of initiating recombination and are “hot,” whereas regions having few such sites are correspondingly “cold.”

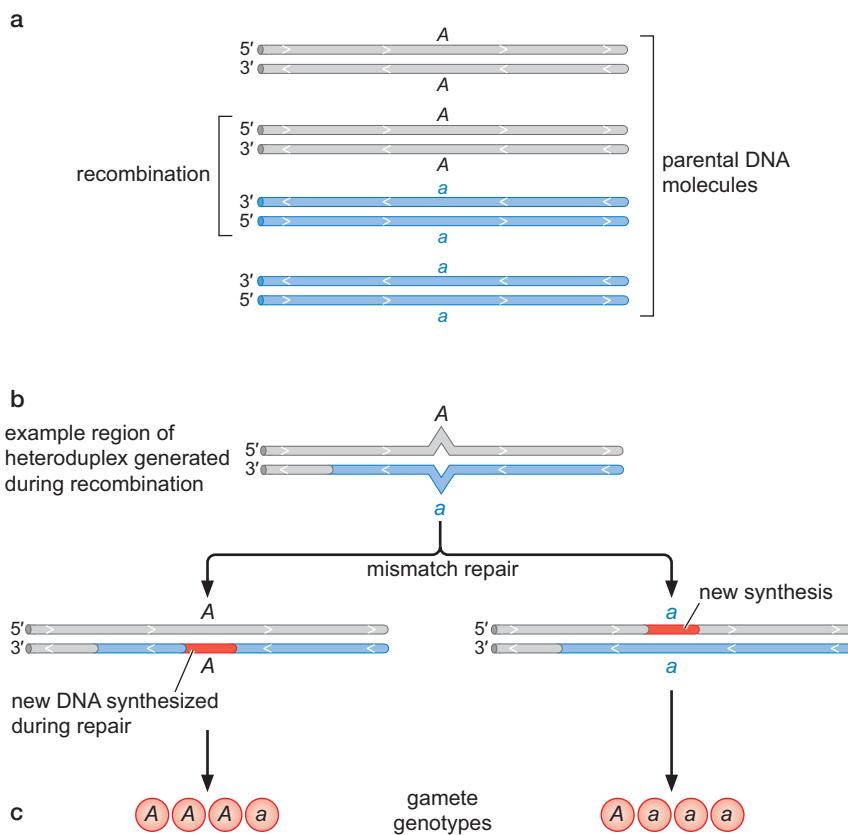
### One Cause of Gene Conversion Is DNA Repair during Recombination

Another genetic consequence of homologous recombination is **gene conversion**. We have introduced the concept of gene conversion during the specialized recombination events responsible for mating-type switching in yeast. However, gene conversion is also commonly observed during normal homologous recombination events, such as those responsible for genetic exchange in bacteria and for pairing chromosomes during meiosis.

To illustrate gene conversion during meiotic recombination, consider a cell undergoing meiosis that has the *A* allele on one homolog and the *a* allele on the other. After DNA replication, four copies of this gene are present, and the genotype would be *AAaa*. In the absence of gene conversion, two gametes carrying the *A* allele and two gametes carrying the *a* allele would be generated. If, instead, gametes with genotype *A a a a* (or *A A A a*) are formed, then a gene conversion event has occurred, in which one copy of the *A* gene has been converted into *a* (or vice versa). How might this conversion arise?

There are two ways that gene conversion can occur during the DSB-repair pathway. First, consider what would happen if the *A* gene was very close to the site of the DSB. In this case, when the 3' ssDNA tails invade the homologous duplexes and are elongated, they may copy the *a* information, which

**FIGURE 11-28** Mismatch repair of heteroduplex DNA within recombination intermediates can give rise to gene conversion.



could replace the *A* information in the product chromosome upon completion of recombination (see Fig. 11-4d).

The second mechanism of gene conversion involves the repair of base-pair mismatches that occur in the recombination intermediates. For example, if either strand invasion or branch migration includes the *A/a* gene, a segment of heteroduplex DNA carrying the *A* sequence on one strand and the *a* sequence on the other strand would be formed (Fig. 11-28; see also Fig. 11-2d, inset). This region of DNA carrying base-pair mismatches could be recognized and acted on by the cellular mismatch repair enzymes (which we discussed in Chapter 10). These enzymes are specialized for fixing base-pair mismatches in DNA. When they detect a mismatched base pair, these enzymes excise a short stretch of DNA from one of the two strands. A repair DNA polymerase then fills in the gap, now with the properly base-paired sequence. When working on recombination intermediates, the mismatch repair enzymes will likely choose randomly which strand to repair. Therefore, after their action, both strands will carry the sequence encoding either the *A* information or the *a* information (depending on which strand was “fixed” by the repair enzymes), and gene conversion will be observed.

## SUMMARY

Homologous recombination occurs in all organisms, allowing for genetic exchange, the reassortment of genes along chromosomes, and the repair of broken DNA strands and collapsed replication forks. The recombination process in-

volves the breaking and rejoining of DNA molecules. The double-strand repair pathway of homologous recombination well describes many recombination events. By this model, initiation of exchange requires that one of the two homolo-

gous DNA molecules has a double-strand break. The broken DNA ends are processed by DNA-degrading enzymes to generate single-stranded DNA segments. These single-strand regions participate in DNA pairing with the homologous partner DNA. Once pairing occurs, the two DNA molecules are joined by a branched structure in the DNA called a Holliday junction. Cutting the DNA at the Holliday junction resolves the junction and terminates recombination. Holliday junctions can be cut in two alternative ways. One way generates crossover products, in which regions from two parental DNA molecules are now covalently joined. The alternative way of cleaving the junction generates a “patch” of recombined DNA but does not result in crossing over.

Cells encode enzymes that catalyze all of the steps in homologous recombination. Key enzymes are the strand-exchange proteins. Of these, *E. coli* RecA is the premier example; RecA-like proteins are found in all organisms. RecA-like strand-exchange proteins promote the search for homologous sequences between two DNA molecules and the exchange of DNA strands within the recombination intermediate. RecA functions as a large protein–DNA complex, known as the RecA filament. Eukaryotic cells encode two strand-exchange proteins, called Rad51 and Dmc1. Other important recombination enzymes are the DNA-cleaving enzymes that generate

double-strand breaks in DNA to initiate recombination; these proteins appear to be found only in eukaryotes and include Spo11 and HO. Nucleases that process the DNA at the break site to generate the required single-strand regions include the RecBCD enzyme in prokaryotes and the MRX–enzyme complex in eukaryotes. Additional enzymes promote the movement (branch migration) and cleavage (resolution) of Holliday junctions.

During meiosis, recombination is essential for the proper homologous pairing of chromosomes before the first nuclear division. Therefore, recombination is highly regulated to ensure that it occurs on all chromosomes. The Spo11 DNA-cutting enzyme and the Dmc1 strand-exchange protein are both specifically involved in these recombination reactions. Homologous recombination is also sometimes used to control gene expression. The mating-type switching of yeast is an excellent example of this type of regulation; it is also an example of gene conversion. Analysis of the mechanism of mating-type switching has generated a new class of models to describe some homologous recombination events called synthesis-dependent strand annealing. This mechanism gives rise to the gene-conversion-type genetic exchange products but does not result in crossing over.

## BIBLIOGRAPHY

### Books

- Brown T.A. 2007. *Genomes*, 3rd ed. Garland Science, New York.  
 Griffiths A.J.F., Miller J.H., Suzuki D.T., Lewontin R.C., and Gelbart W.M. 2000. *An introduction to genetic analysis*, 7th ed. W.H. Freeman, New York.

### Recombination in Bacteria

- Chen Z., Yang H., and Pavletich N.P. 2008. Mechanism of homologous recombination from the RecA-ssDNA/dsDNA structures. *Nature* **453**: 489–494.  
 Court D.L., Sawitzke J.A., and Thomason L.C. 2002. Genetic engineering using homologous recombination. *Annu. Rev. Genet.* **36**: 361–388.  
 Cox M.M. 2001. Recombinational DNA repair of damaged replication forks in *Escherichia coli*: Questions. *Annu. Rev. Genet.* **35**: 53–82.  
 Kowalczykowski S.C., Dixon D.A., Eggleston A.K., Lauder S.D., and Rehrauer W.M. 1994. Biochemistry of homologous recombination in *Escherichia coli*. *Microbiol. Rev.* **58**: 401–465.  
 Lusetti S.L. and Cox M.M. 2002. The bacterial RecA protein and the recombinatorial DNA repair of stalled replication forks. *Annu. Rev. Biochem.* **71**: 71–100.  
 Smith G.R. 2001. Homologous recombination near and far from DNA breaks: Alternative roles and contrasting views. *Annu. Rev. Genet.* **35**: 243–274.

### Recombination in Eukaryotes

- Bernstein K.A., Gangloff S., and Rothstein R. 2010. The RecQ DNA helicases in DNA repair. *Annu. Rev. Genet.* **44**: 393–417.  
 Eichler E.E. and Sankoff D. 2003. Structural dynamics of eukaryotic chromosome evolution. *Science* **301**: 793–797.

Keeney S. 2001. Mechanism and control of meiotic recombination initiation. *Curr. Top. Dev. Biol.* **52**: 1–53.

Page S.L. and Hawley R.S. 2003. Chromosome choreography: The meiotic ballet. *Science* **301**: 785–789.

Pâques F. and Haber J.E. 1999. Multiple pathways of recombination induced by double-strand breaks in *Saccharomyces cerevisiae*. *Microbiol. Mol. Biol. Rev.* **63**: 349–404.

Pastink A., Eeken J.C., and Lohman P.H. 2001. Genomic integrity and the repair of double-strand DNA breaks. *Mutat. Res.* **480–481**: 37–50.

Prado F., Cortes-Ledesma F., Huertas P., and Aguilera A. 2003. Mitotic recombination in *Saccharomyces cerevisiae*. *Curr. Genet.* **42**: 185–198.

Stracker T.H. and Petrini J.H.J. 2011. The MRE11 complex: Starting from the ends. *Nat. Rev.* **12**: 90–103.

Symington L.S. 2002. Role of RAD52 epistasis group genes in homologous recombination and double-strand break repair. *Microbiol. Mol. Biol. Rev.* **66**: 630–670.

van den Bosch M., Lohman P.H., and Pastink A. 2002. DNA double-strand break repair by homologous recombination. *Biol. Chem.* **383**: 873–892.

West S.C. 2003. Molecular views of recombination proteins and their control. *Nat. Rev. Mol. Cell Biol.* **4**: 435–445.

### Mating-Type Switching in Yeast

- Haber J.E. 2002. Switching of *Saccharomyces cerevisiae* mating-type genes. In *Mobile DNA II* (ed. N.L. Craig, et al.), pp. 927–952. ASM Press, Washington, D.C.  
 Haber J.E. 2012. Mating-type genes and MAT switching in *Saccharomyces cerevisiae*. *Genetics* **191**: 33–64.

**QUESTIONS**

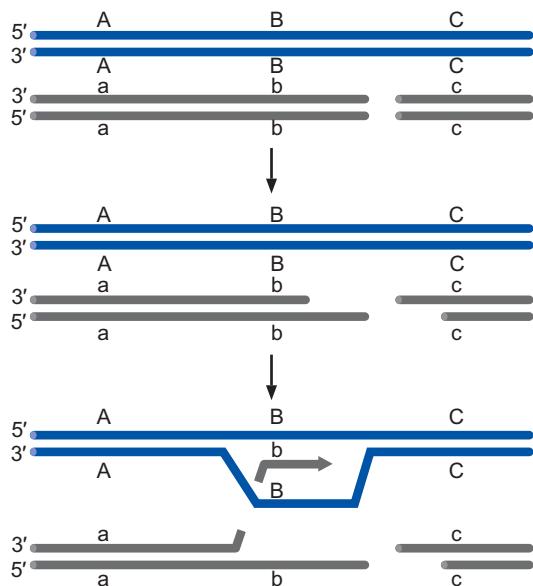
For answers to even-numbered questions, see Appendix 2: Answers.

**Question 1.** Explain two ways in which DNA replication can introduce double-strand breaks (DSBs) in a DNA template.

**Question 2.** You are considering two alleles of one specific gene. Describe what feature with respect to the DNA distinguishes one allele from the other. Are the two alleles homologous?

**Question 3.** Following formation of a DSB, enzymes process the double-stranded DNA to form single-stranded DNA as shown in Figure 11-4b. Does it matter if resection occurs in the 5' to 3' direction or the 3' to 5' direction? Explain why or why not.

**Question 4.** The picture below depicts the first three steps of DSB repair homologous recombination with some errors. List the error(s), why each error is a problem, and how this error should (or could) be corrected.

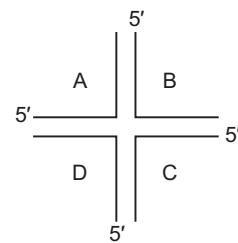


**Question 5.** What is meant by the term heteroduplex DNA?

**Question 6.** List the different enzymatic activities that RecBCD catalyzes and describe the significance of each activity in the steps of homologous recombination (via the DSB-repair pathway).

**Question 7.** Explain why RecA-dependent strand exchange cannot occur between two homologous, double-stranded, covalently closed, circular DNAs (cccDNAs) but can occur between the two double-stranded DNA molecules pictured in Figure 11-8c.

**Question 8.** Using the Holliday junction DNA substrate pictured below (5'-<sup>32</sup>P end labeled), propose an assay that researchers may have used to determine RuvA protein binding to the DNA substrate. Propose a modification to the DNA substrate to use as a negative control that demonstrates an aspect of the specificity of RuvA binding. (A, B, C, and D are labels for the unique ssDNA fragments that were used to assemble this substrate.)



**Question 9.** Explain the most significant role of homologous recombination in eukaryotes that is not found in prokaryotes.

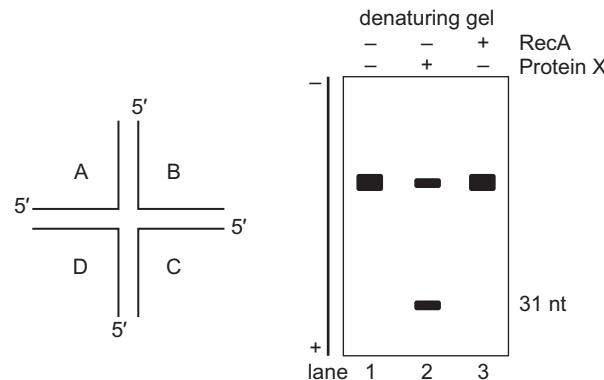
**Question 10.** Briefly describe how cells generate the DSB required to initiate meiotic homologous recombination in eukaryotes.

**Question 11.** Define gene conversion and give an example of a mechanism explaining how gene conversion occurs.

**Question 12.** Compare and contrast synthesis-dependent strand annealing (SDSA) used in mating-type switching with DSB-repair homologous recombination.

**Question 13.** Explain why DSB-repair homologous recombination can occur between any two DNA molecules that share homology rather than only between two DNA molecules that carry a specific sequence.

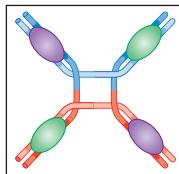
**Question 14.** Researchers characterized a human protein with a role in Holliday junction resolution. They purified and used the protein in an assay with the DNA substrate shown below on the left. Each of the four strands of DNA (A, B, C, D) is 60 nucleotides (nt) long, and only the DNA strand marked A is end-labeled with 5'-<sup>32</sup>P. The researchers set up three reactions with the DNA substrate, incubating (1) with no protein, (2) with the purified protein (Protein X), and (3) with RecA (as a control). The products from each reaction were run in separate lanes of a denaturing polyacrylamide gel. An autoradiogram of the gel is shown below on the right.



- A. Given these data, propose a function for Protein X.
- B. Based on your answer in part A, this protein functions similarly to what *E. coli* protein?

Data adapted from Rass et al. (2010. *Genes Dev.* **24**: 1559–1569).

CHAPTER 12



# Site-Specific Recombination and Transposition of DNA

## OUTLINE

DNA IS A VERY STABLE MOLECULE. DNA replication, repair, and homologous recombination, as we have learned in the previous chapters, all occur with high fidelity. These processes serve to ensure that the genomes of an organism are nearly identical from one generation to the next. Importantly, however, there are also genetic processes that rearrange DNA sequences and thus lead to a more dynamic genome structure. These processes are the subject of this chapter.

Two classes of genetic recombination—**conservative site-specific recombination (CSSR)** and **transpositional recombination** (generally called **transposition**)—are responsible for many important DNA rearrangements. CSSR is recombination between two defined sequence elements (Fig. 12-1). Transposition, in contrast, is recombination between specific sequences and nonspecific DNA sites. The biological processes promoted by these recombination reactions include the insertion of viral genomes into the DNA of the host cell during infection, the inversion of DNA segments to alter gene structure, and the movement of **transposable elements**—often called “jumping” genes—from one chromosomal site to another.

The impact of these DNA rearrangements on chromosome structure and function is profound. In many organisms, transposition is the major source of spontaneous mutation, and nearly half the human genome consists of sequences derived from transposable elements (although most elements are currently inactive). Furthermore, as we shall see, both viral infection and development of the vertebrate immune system depend critically on these specialized DNA rearrangements.

Conservative site-specific recombination and transposition share key mechanistic features. Proteins known as **recombinases** recognize specific sequences where recombination will occur within a DNA molecule. The recombinases bring these specific sites together to form a protein–DNA complex bridging the DNA sites, known as the **synaptic complex**. Within the synaptic complex, the recombinase catalyzes the cleavage and rejoicing of the DNA molecules either to invert a DNA segment or to move a segment to a new site. One recombinase protein is usually responsible for all of these steps. Both types of recombination are also carefully controlled such that

Conservative Site-Specific Recombination, 378



Biological Roles of Site-Specific Recombination, 386



Transposition, 393



Examples of Transposable Elements and Their Regulation, 406

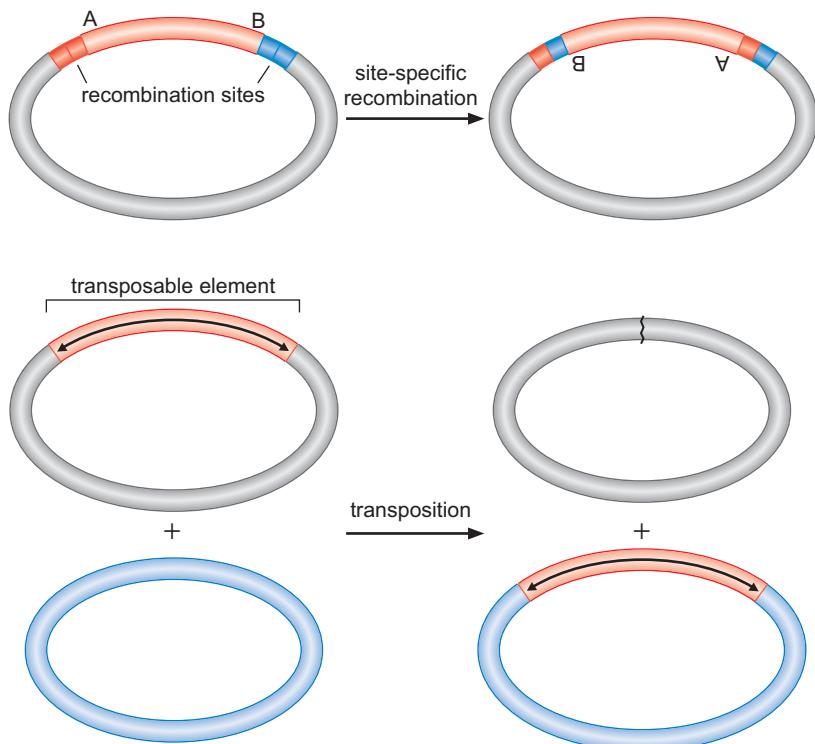


V(D)J Recombination, 416



Visit Web Content for Structural Tutorials and Interactive Animations

**FIGURE 12-1** Two classes of genetic recombination. (Top panel) Example of site-specific recombination. Here, recombination between the red and blue recombination sites inverts the DNA segment carrying the A and B genes. (Bottom panel) Example of transposition in which the red transposable element excises from the gray DNA and inserts into an unrelated site in the blue DNA.



the danger to the cell of introducing breaks in the DNA, and rearranging DNA segments in an unintended manner, is minimized. As we shall see, however, the two types of recombination also have key mechanistic differences.

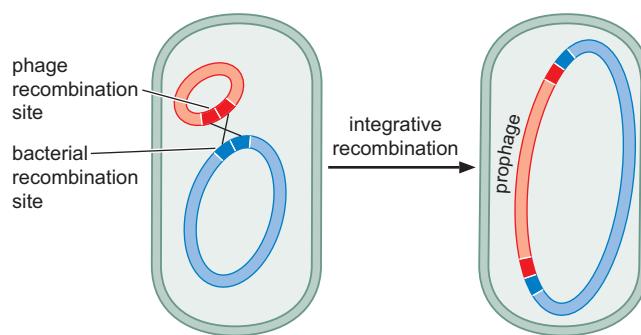
In the following sections, the simpler site-specific recombination reactions are introduced first, followed by a discussion of transposition. Each of these sections is organized to describe general features of the mechanism first and then to provide some specific examples.

## CONSERVATIVE SITE-SPECIFIC RECOMBINATION

### Site-Specific Recombination Occurs at Specific DNA Sequences in the Target DNA

CSSR is responsible for many reactions in which a defined segment of DNA is rearranged. A key feature of these reactions is that the segment of DNA that will be moved carries specific short sequence elements, called **recombination sites**, where DNA exchange occurs. An example of this type of recombination is the integration of the bacteriophage  $\lambda$  genome into the bacterial chromosome (Fig. 12-2 and Appendix 1).

During  $\lambda$  integration, recombination always occurs at exactly the same nucleotide sequence within two recombination sites, one on the phage DNA and the other on the bacterial DNA. Recombination sites carry two classes of sequence elements: sequences specifically bound by the recombinases and sequences where DNA cleavage and rejoining occur. Recombination sites are often quite short, 20 bp or so, although they may be much longer and carry additional sequence motifs and protein-binding sites.



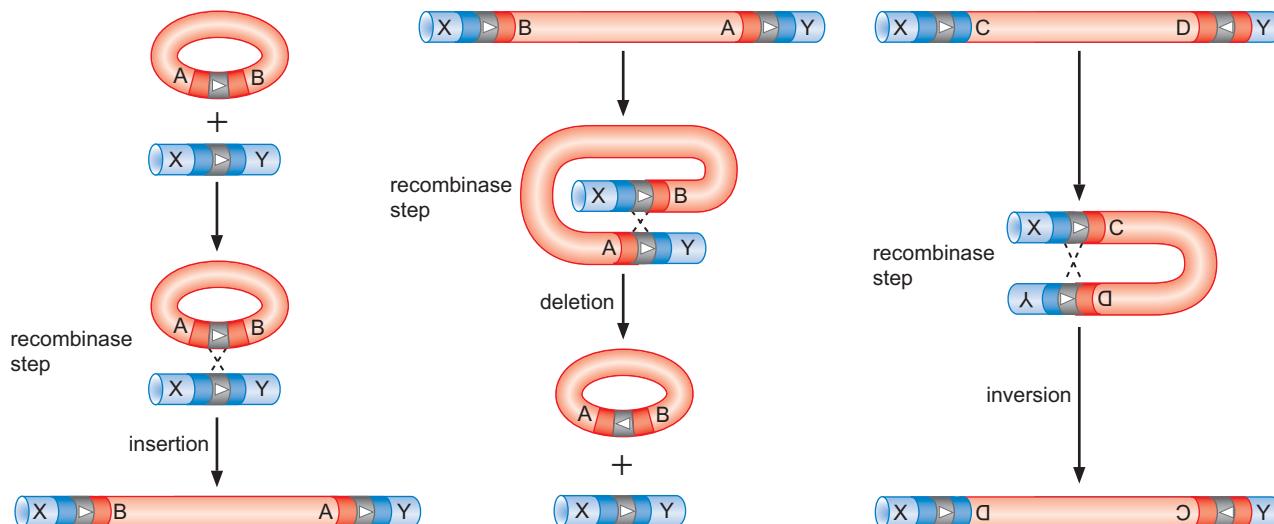
**FIGURE 12-2** Integration of the  $\lambda$  genome into the chromosome of the host cell. DNA exchange occurs specifically between the recombination sites on the two DNA molecules. The relative lengths of the  $\lambda$  and cellular chromosomes are not shown to scale.

Examples of the more complex recombination sites are discussed when we consider specific recombination examples.

CSSR can generate three different types of DNA rearrangements (Fig. 12-3): (1) insertion of a segment of DNA into a specific site (as occurs during bacteriophage  $\lambda$  DNA integration), (2) deletion of a DNA segment, or (3) inversion of a DNA segment. Whether recombination results in DNA insertion, deletion, or inversion depends on the organization of the recombinase recognition sites on the DNA molecule or molecules that participate in recombination.

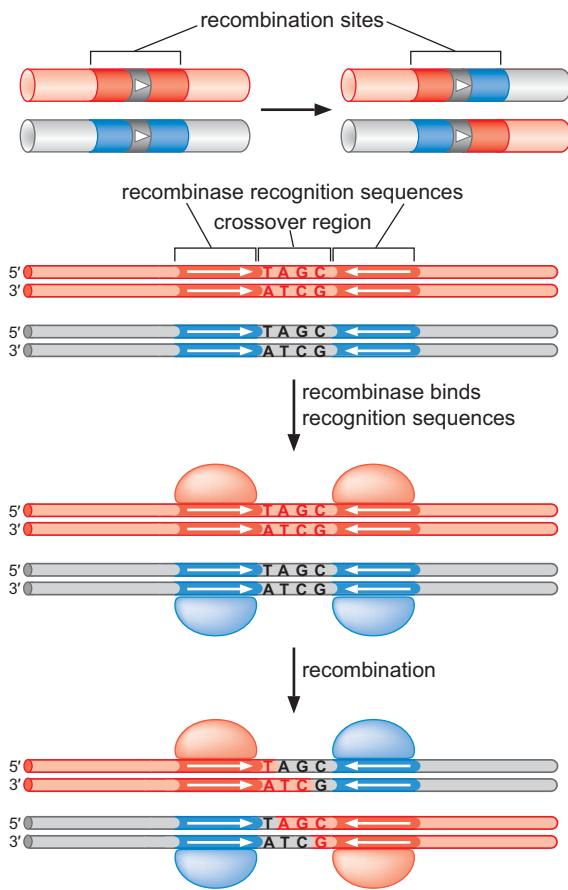
To understand how the organization of recombination sites determines the type of DNA rearrangement, we must look at the sequence elements within the recombination sites in more detail (Fig. 12-4). Each recombination site is organized as a pair of **recombinase recognition sequences**, positioned symmetrically. These recognition sequences flank a central short asymmetric sequence, known as the **crossover region**, where DNA cleavage and rejoining occur.

Because the crossover region is asymmetric, a given recombination site always has a defined polarity. The orientation of two sites present on a single



**FIGURE 12-3** Three types of CSSR recombination. In each case, it is the red segment of DNA that is moved or rearranged during recombination. A, B, C, D, X, and Y denote genes that lie within the different segments of DNA. Darker red and blue boxes represent the recombinase-recognition sequences, and black arrows show the crossover regions. Together these sequence elements form the recombination sites.

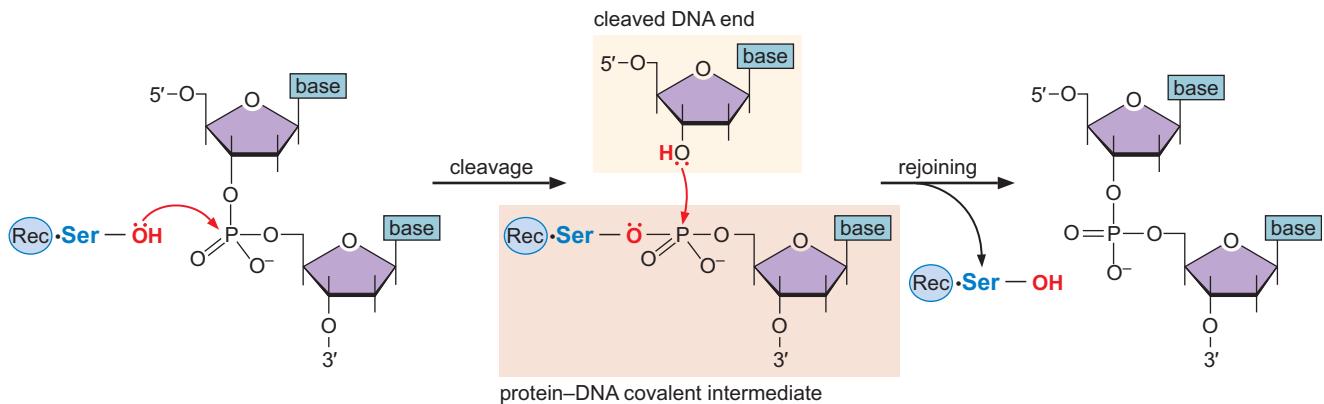
**FIGURE 12-4** Structures involved in CSSR. The pair of symmetric recombinase recognition sequences flanks the crossover region where recombination occurs. The subunits of the recombinase bind these recognition sites. Note that the sequence of the crossover region is not palindromic, resulting in an intrinsic asymmetry to the recombination sites. (Adapted, with permission, from Craig N. et al. 2002. *Mobile DNA II*, p. 4, Fig. 1. © ASM Press.)



DNA molecule will be related to each other in either an **inverted repeat** or a **direct repeat** manner. Recombination between a pair of inverted sites will invert the DNA segment between the two sites (Fig. 12-3, right panel). In contrast, recombination using the identical mechanism but occurring between sites organized as direct repeats deletes the DNA segment between the two sites (Fig. 12-3, middle panel). Finally, insertion specifically occurs when recombination sites on two different molecules are brought together for DNA exchange (Fig. 12-3, left panel). Examples of each of these three types of rearrangements are considered later, after a general discussion of the recombinases.

### Site-Specific Recombinases Cleave and Rejoin DNA Using a Covalent Protein–DNA Intermediate

There are two families of conservative site-specific recombinases: the **serine recombinases** and the **tyrosine recombinases**. Fundamental to the mechanism used by both families is that when they cleave the DNA, a covalent protein–DNA intermediate is generated. For the serine recombinases, the side chain of a serine residue within the protein's active site attacks a specific phosphodiester bond in the recombination site (Fig. 12-5). This reaction introduces a single-strand break in the DNA and simultaneously generates a covalent linkage between the serine and a phosphate at this DNA cleavage site. Likewise, for the tyrosine recombinases, it is the side chain of the active-site tyrosine that attacks and then becomes joined to the DNA. Table 12-1 classifies several important recombinases by family and biological function.



**FIGURE 12-5** Covalent-intermediate mechanism used by the serine and tyrosine recombinases. Here, an OH group from an active-site serine is shown to attack the phosphate and thereby introduce a single-strand break at the site of recombination. The liberated OH group on the broken DNA can then reattack the protein–DNA covalent bond to reverse this cleavage reaction, reseal the DNA, and release the protein. (Blue) The recombinase, labeled Rec.

The covalent protein–DNA intermediate conserves the energy of the cleaved phosphodiester bond within the protein–DNA linkage. As a result, the DNA strands can be rejoined by reversal of the cleavage process. For reversal, an OH group from the cleaved DNA attacks the covalent bond that links the protein to the DNA. This process covalently seals the DNA break and regenerates the free (non–DNA bound) recombinase (see Fig. 12-5).

It is this mechanistic feature that contributes “conservative” to the CSSR name: it is called conservative because every DNA bond that is broken during the reaction is resealed by the recombinase. No external energy, such as that released by ATP hydrolysis, is needed for DNA cleavage and joining by these proteins. This cleavage mechanism, with its covalent intermediate, is not unique to the recombinases. Both DNA topoisomerases (see Chapter 4) and Spo11, the protein that introduces double-strand breaks into DNA to initiate homologous recombination during meiosis (see Chapter 11), use this mechanism.

**TABLE 12-1** Recombinases by Family and by Function

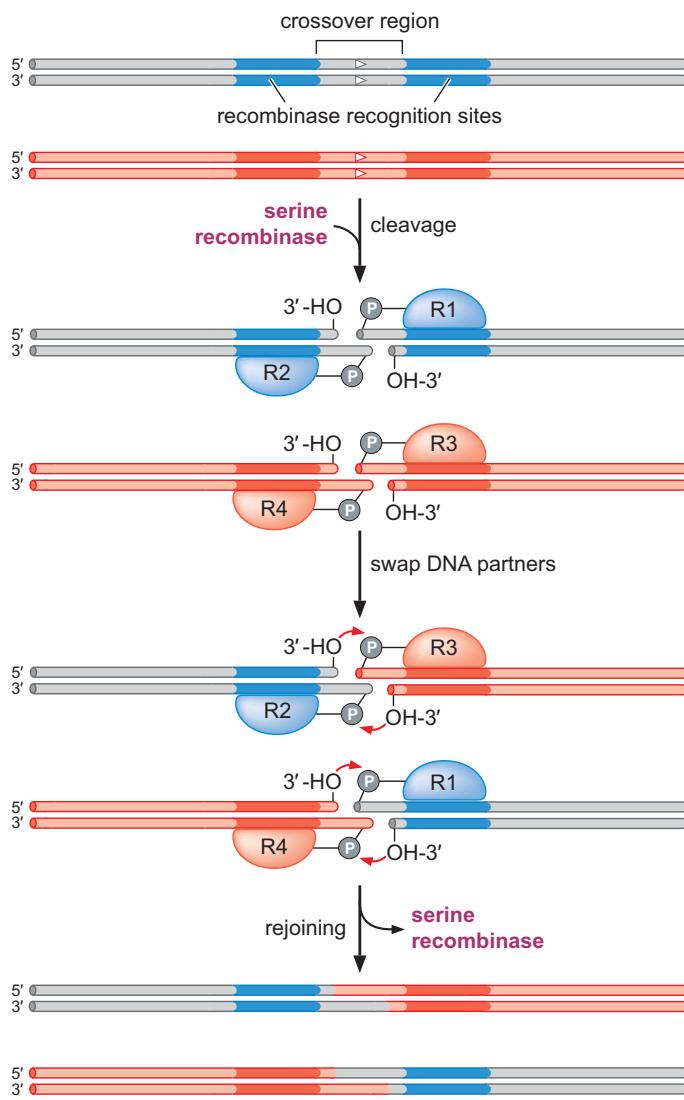
Recombinase	Function
<b>Serine family</b>	
<i>Salmonella</i> Hin invertase	Inverts a chromosomal region to flip a gene promoter by recognizing <i>hix</i> sites. Allows expression of two distinct surface antigens.
Transposon Tn3 and $\gamma\delta$ resolvases	Promotes a DNA deletion reaction to resolve the DNA fusion event that results from replicative transposition. Recombination sites are called <i>res</i> sites.
<b>Tyrosine family</b>	
Bacteriophage $\lambda$ integrase	Promotes DNA integration and excision of the $\lambda$ genome into, and out of, a specific sequence on the <i>E. coli</i> chromosome. Recombination sites are called <i>att</i> sites.
Phage P1 Cre	Promotes circularization of the phage DNA during infection by recognizing sites (called <i>lox</i> sites) on the phage DNA.
<i>Escherichia coli</i> XerC and XerD	Promotes several DNA deletion reactions that convert dimeric circular DNA molecules into monomers. Recognizes both plasmid-borne sites ( <i>cer</i> ) and chromosomal sites ( <i>dif</i> ).
Yeast FLP	Inverts a region of the yeast $2\mu$ plasmid to allow for a DNA amplification reaction called rolling circle replication. Recombination sites are called <i>frt</i> sites.

### Serine Recombinases Introduce Double-Strand Breaks in DNA and Then Swap Strands to Promote Recombination

CSSR always occurs between two recombination sites. As we have seen above, these sites may be on the same DNA molecule (for inversion or deletion) or on two different molecules (for integration). Each recombination site is made up of double-stranded DNA. Therefore, during recombination, four single strands of DNA (two from each duplex) must be cleaved and then rejoined—now with a different partner strand—to generate the rearranged DNA.

The serine recombinases cleave all four strands before strand exchange occurs (Fig. 12-6). One molecule of the recombinase protein promotes each of these cleavage reactions; therefore, four subunits of the recombinase are required.

These double-stranded DNA breaks in the parental DNA molecules generate four double-stranded DNA segments (marked by the proteins bound to them as R1, R2, R3, and R4 in Fig. 12-6). For recombination to occur, the R2 segment of the top DNA molecule must recombine with the R3 segment of the bottom DNA molecule. Likewise, the R1 segment of the top molecule



**FIGURE 12-6** Recombination by a serine recombinase. Each of the four DNA strands is cleaved within the crossover region by one subunit of the protein. These subunits are labeled R1, R2, R3, and R4. Cleavage of the two individual strands of one duplex is staggered by two bases. This two-base region forms a hybrid duplex in the recombinant products. The recombination sites are similar to those shown in Figure 12-4.

must recombine with the R4 segment of the bottom DNA molecule. Once this DNA “swap” has occurred, the 3'-OH ends of each of the cleaved DNA strands can attack the recombinase–DNA bond in their new partner segment. As discussed above, this reaction liberates the recombinase and covalently seals the DNA strands to generate the rearranged DNA product.

### Structure of the Serine Recombinase–DNA Complex Indicates that Subunits Rotate to Achieve Strand Exchange

The structure of a serine recombinase–DNA complex in the process of recombination provides a snapshot of how the exchange of DNA strands is physically coordinated. The complex contains four subunits of the recombinase, and two cleaved, double-stranded DNA molecules. The covalent linkage between the active-site serine in each recombinase subunit and 5'-phosphates in the DNA of each recombination half-site is clearly visible. Each of these linkages, in turn, leaves a free 3'-OH DNA end that can participate in strand exchange.

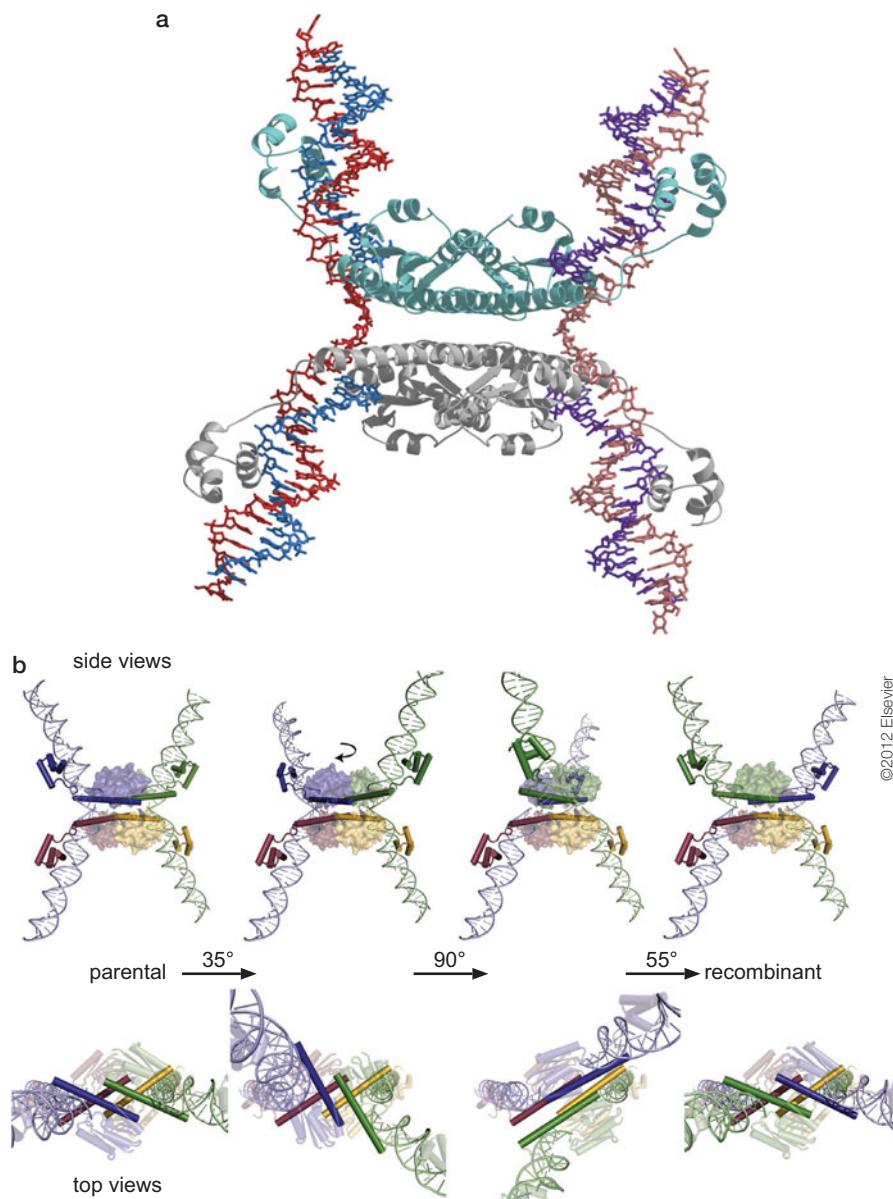
The most dramatic feature of the structure is the large, flat interface between the “top” and “bottom” recombinase dimers (Fig. 12-7a). This structure is largely hydrophobic, and slippery, providing little barrier to impede rotation of the top and bottom halves of the complex around each other. However, some regions of complementary positive and negative charge can serve to stabilize the structure specifically in the initial and the 180° rotated orientation. Thus, analysis of this complex strongly supports the model that the mechanism of recombination is (1) DNA cleavage to form the covalent enzyme–DNA intermediate; (2) an 180° rotation of the dimers in the protein–DNA complex; and (3) attack of the 3'-OH DNA ends on the resolvase-DNA linkages to join the strands in the new, recombined configuration. As more structural and mechanistic experiments have been completed, further insight into this dramatic protein rotation has emerged (Fig. 12-7b).

### Tyrosine Recombinases Break and Rejoin One Pair of DNA Strands at a Time

In contrast to the serine recombinases, the tyrosine recombinases cleave and rejoin two DNA strands first and only then cleave and rejoin the other two strands (Fig. 12-8). Consider two DNA molecules with their recombination sites aligned. Here also, four molecules of the recombinase are needed, one to cleave each of the four individual DNA strands. To start recombination, the subunits of recombinase bound to the left recombinase-binding sites (marked as R1 and R3 in Fig. 12-8a) each cleave the top strand of the DNA molecule to which they are bound. This cleavage occurs at the first nucleotide of the crossover region. Next, the right top strand from the top (gray) DNA molecule and the right top strand from the bottom (red) DNA molecule “swap” partners. These two DNA strands are then joined, now in the recombined configurations. This “first-strand” exchange reaction generates a branched DNA intermediate known as a Holliday junction (see Chapter 11) (Fig. 12-8b).

Once the first-strand exchange is complete, two more recombinase subunits (those marked R2 and R4) cleave the bottom strands of each DNA molecule (Fig. 12-8c). These strands again switch partners and then are joined by the reversal of the cleavage reaction. This “second-strand” exchange reaction “undoes” (i.e., resolves; see Chapter 11) the Holliday junction, to yield

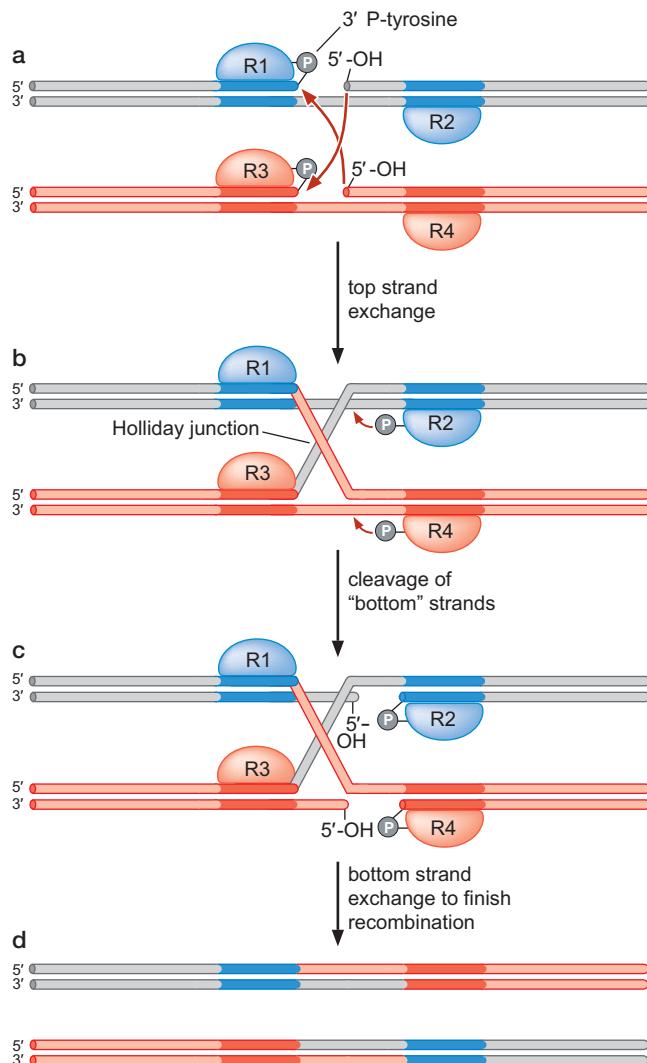
**FIGURE 12-7** The structure of serine recombinase. (a) The structure shows the large, flat tetramer interface that is the site of rotation. This recombinase tetramer is constructed from a dimer of the blue-green subunits (top) and a second dimer of the gray subunits (bottom). Each subunit in this tetramer is bound site-specifically to one of the four “DNA arms” in the structure (note DNA contacts provided by  $\alpha$  helices in the protein subunits on the outside faces of the DNA molecules). For the full recombination reaction, all four DNA strands must be cleaved (as in Fig. 12-6). In this structure, however, only two are cleaved. (Li W. et al. 2005. *Science* **309**: 1210. PDB Code: 1ZR4.) Image prepared with MolScript, BobScript, and Raster3D. (b) In the upper sequence, the side views are shown looking into the tetramer interface that is the site of rotation; in the lower sequence, the structures are rotated by 90° to show the top views, revealing the angles of alignment of the subunit pairs of the tetramers (shown as colored bars). In this sequence, the structure is shown in the different rotational conformations of serine recombinase tetramers during the process of DNA cleavage. The starting tetramer conformation on the left (the parental form) is poised for cleavage; the  $\alpha$  helices from each rotating subunit pair are oriented at an  $\sim 50^\circ$  angle. The first clockwise rotation of  $\sim 35^\circ$  generates the conformer in which the helices of the subunit pairs are now at  $\sim 85^\circ$  crossing angle. An additional  $\sim 90^\circ$  rotation generates the aligned conformer, and a final rotation of  $\sim 55^\circ$  completes one round of subunit exchange and aligns the DNA strands for ligation in the recombinant configuration. (Image modified, with permission, from Johnson R.C. and McLean M.M. 2011. *Structure* **19**: 751–753; Fig. 2, p. 752. © Elsevier.)



the rearranged DNA products. In the next section, we discuss how these chemical steps occur in the context of the recombinase protein–DNA complex.

### Structures of Tyrosine Recombinases Bound to DNA Reveal the Mechanism of DNA Exchange

The mechanism of site-specific recombination is best understood for the tyrosine recombinases. Several structures of members of this protein class have been solved, and these structures reveal the recombinases “caught in the act” of recombination. One beautiful example is the structure of the Cre recombinase bound to two different configurations of the recombinating DNA (see Structural Tutorial 12.1). Insights into the mechanisms derived from these structures are explained later. Cre is an enzyme encoded by phage P1, which functions to circularize the linear phage genome



**FIGURE 12-8** Recombination by a tyrosine recombinase. Here, the R1 and R3 subunits cleave the DNA in the first step (a); in the example shown, the protein becomes linked to the cut DNA by a 3' P-tyrosine bond. (b) Exchange of the first pair of strands occurs when the two 5'-OH groups at the break sites each attacks the protein-DNA bond on the other DNA molecule. (c,d) The second-strand exchange occurs by the same mechanism, using the R2 and R4 subunits. (Adapted from Craig N. et al. 2002. *Mobile DNA II*, Color Plate 1, Chapter 2. © ASM Press.)

during infection. The recombination sites on the DNA, where Cre acts, are called *lox* sites. *Cre-lox* is a simple example of recombination by the tyrosine recombinase family; only Cre protein and the *lox* sites are needed for complete recombination. Cre is also widely used as a tool in genetic engineering (see Box 12-1, Application of Site-Specific Recombination to Genetic Engineering).

The *Cre-lox* structures reveal that recombination requires four subunits of Cre, with each molecule bound to one binding site on the substrate DNA molecules (Fig. 12-9). The conformation of the DNA is generally a square planar four-way junction (see discussion of Holliday junctions in Chapter 11) with each “arm” of this junction bound by one subunit of Cre. Although at first glance the structures appear to have fourfold symmetry, this is not really the case. Cre exists in two distinct conformations with one pair of subunits in conformation 1, shown in green, and the other pair in conformation 2, shown in purple (Fig. 12-9b). Only in one of these conformations (the green subunits in the figure) can Cre cleave and rejoin DNA. Thus, only one pair of subunits is in the active conformation at a time. The pair of subunits in this active conformation switches as the reaction progresses. This switching is critical for controlling the progress of recombination and ensuring the sequential “one strand at a time” exchange mechanism.

## ► MEDICAL CONNECTIONS

### Box 12-1 Application of Site-Specific Recombination to Genetic Engineering

Because some site-specific recombination systems are so simple, they have become widely used as tools in experimental genetics. Cre recombinase and its close relative FLP recombinase are both used experimentally to delete genes in eukaryotic organisms (also see the example in Appendix 1).

An example of the usefulness of this strategy becomes clear when we consider the following hypothetical example. A researcher is interested in the role of a specific gene in the development of lung cancer and wishes to study this process using the mouse as a model organism (see Appendix 1). When the gene of interest is disrupted or “knocked out” (see Fig. A-27), however, the mice all die during early embryogenesis. Apparently, the gene is required very early in development. How can its role in lung cancer be studied in the adult animal?

Site-specific recombination can often provide the answer. Using routine methods, researchers can introduce recombi-

nation sites recognized by Cre (or FLP) flanking the gene of interest. These sites will have no effect on the gene’s function, unless the recombinase is also present. Therefore, the Cre protein (or FLP protein) can be introduced into the same organism, under the control of a promoter that can be carefully regulated (see Chapter 19). The mice can therefore be allowed to develop in the absence of the recombinase, but then after birth, Cre expression can be “turned on.” The presence of the recombinase causes deletion of the gene of interest. In this case, the propensity of the Cre-treated mice (in which the gene is deleted) for lung cancer can now be compared with their “normal” littermates, in which the gene of interest is still intact. Thus, recombination using Cre allows the potential functions of the genes to be uncovered in different stages of development.

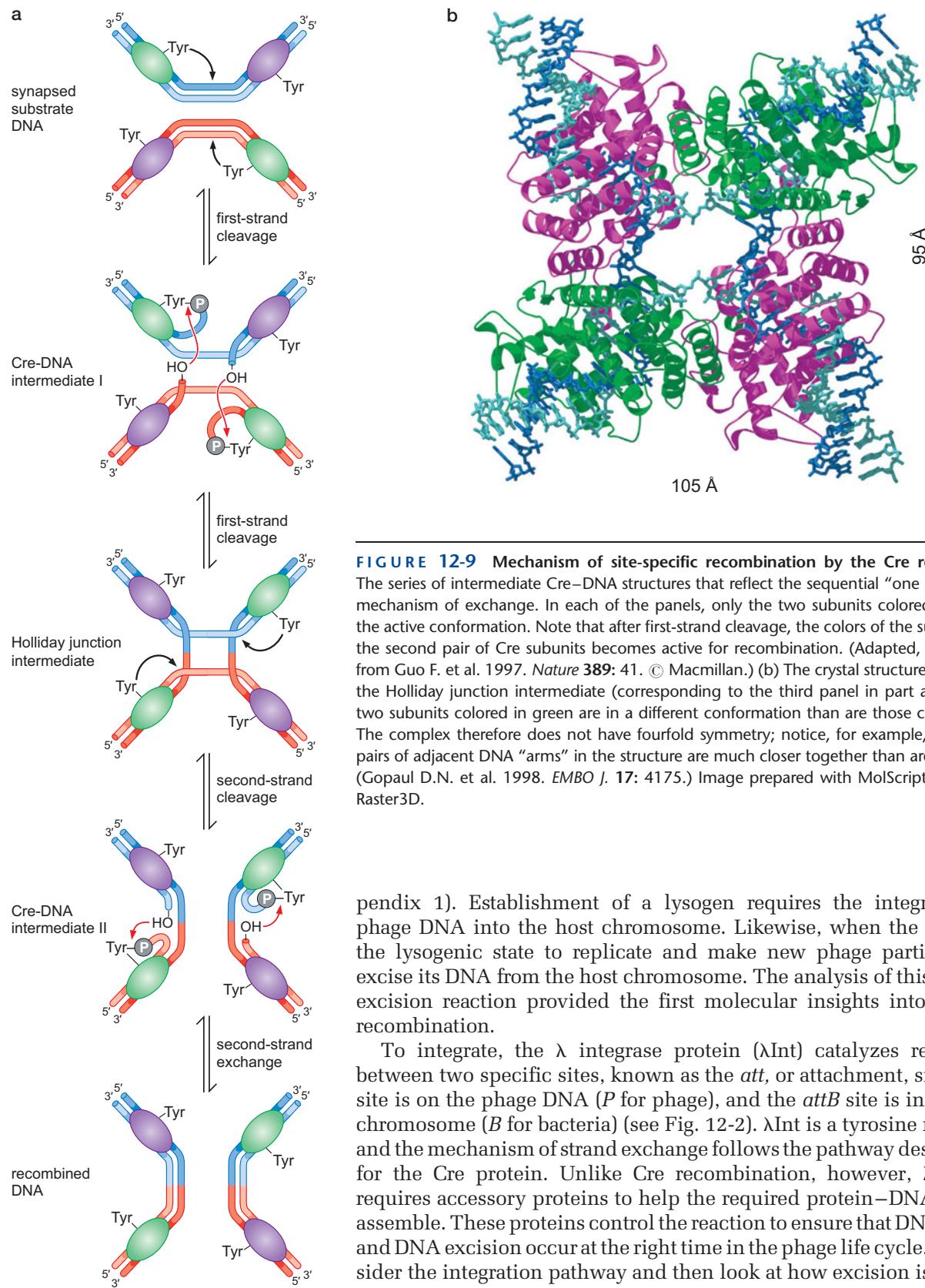
## BIOLOGICAL ROLES OF SITE-SPECIFIC RECOMBINATION

Cells and viruses use conservative site-specific recombination for a wide variety of biological functions. Some of these functions are discussed in the following sections. Many phage insert their DNA into the host chromosome during infection using this recombination mechanism. In other cases, site-specific recombination is used to alter gene expression. For example, inversion of a DNA segment can allow two alternative genes to be expressed. Site-specific recombination is also widely used to help maintain the structural integrity of circular DNA molecules during cycles of DNA replication, homologous recombination, and cell division.

A comparison of site-specific recombination systems reveals some general themes. All reactions depend critically on the assembly of the recombinase protein on the DNA and the bringing together of the two recombination sites. For some recombination reactions, this assembly is very simple, requiring only the recombinase and its DNA recognition sequences as just described for Cre. In contrast, other reactions require accessory proteins. These accessory proteins include so-called **architectural proteins** that bind specific DNA sequences and bend the DNA. They organize DNA into a specific shape and thereby stimulate recombination. Architectural proteins can also control the direction of a recombination reaction, for example, to ensure that integration of a DNA segment occurs while preventing the reverse reaction—DNA excision. Clearly, this type of regulation is essential for a logical biological outcome. Finally, we will also see that recombinases can be regulated by other proteins to control when a particular DNA rearrangement takes place and coordinate it with other cellular events.

### **λ Integrase Promotes the Integration and Excision of a Viral Genome into the Host-Cell Chromosome**

When bacteriophage  $\lambda$  infects a host bacterium, a series of regulatory events results either in establishment of the quiescent **lysogenic state** or in phage multiplication, a process called **lytic growth** (see Chapter 18 and Ap-



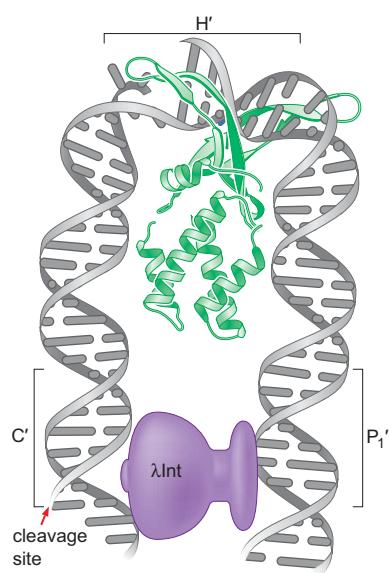
**FIGURE 12-9** Mechanism of site-specific recombination by the Cre recombinase. (a) The series of intermediate Cre–DNA structures that reflect the sequential “one strand at a time” mechanism of exchange. In each of the panels, only the two subunits colored in green are in the active conformation. Note that after first-strand cleavage, the colors of the subunits switch as the second pair of Cre subunits becomes active for recombination. (Adapted, with permission, from Guo F. et al. 1997. *Nature* 389: 41. © Macmillan.) (b) The crystal structure of Cre bound to the Holliday junction intermediate (corresponding to the third panel in part a). Note that the two subunits colored in green are in a different conformation than are those colored in purple. The complex therefore does not have fourfold symmetry; notice, for example, that two of the pairs of adjacent DNA “arms” in the structure are much closer together than are the other pairs. (Gopaul D.N. et al. 1998. *EMBO J.* 17: 4175.) Image prepared with MolScript, BobScript, and Raster3D.

pendix 1). Establishment of a lysogen requires the integration of the phage DNA into the host chromosome. Likewise, when the phage leaves the lysogenic state to replicate and make new phage particles, it must excise its DNA from the host chromosome. The analysis of this integration/excision reaction provided the first molecular insights into site-specific recombination.

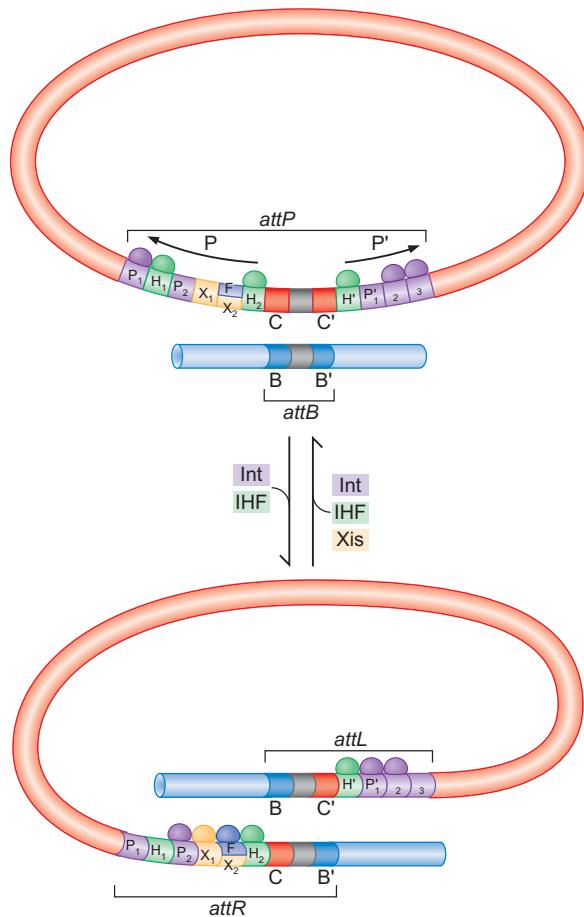
To integrate, the  $\lambda$  integrase protein ( $\lambda$ Int) catalyzes recombination between two specific sites, known as the *att*, or attachment, sites. The *attP* site is on the phage DNA (*P* for phage), and the *attB* site is in the bacterial chromosome (*B* for bacteria) (see Fig. 12-2).  $\lambda$ Int is a tyrosine recombinase, and the mechanism of strand exchange follows the pathway described above for the Cre protein. Unlike Cre recombination, however,  $\lambda$  integration requires accessory proteins to help the required protein–DNA complex to assemble. These proteins control the reaction to ensure that DNA integration and DNA excision occur at the right time in the phage life cycle. We first consider the integration pathway and then look at how excision is triggered.

Important to the regulation of  $\lambda$  integration is the highly asymmetric organization of the *attP* and *attB* sites (Fig. 12-10). Both sites carry a central core segment (~30 bp). These core recombination sites each consist of two  $\lambda$ Int-binding sites and a crossover region where strand exchange occurs (as

**FIGURE 12-10 Recombination sites involved in  $\lambda$  integration and excision showing the important sequence elements.** C, C', B, and B' are the core  $\lambda$ Int-binding sites. The additional protein-binding sites are on  $attP$  and flank the C and C' sites. These regions are called the “arms”; the sequences on the left are called the P arm, and those on the right are called the P' arm. The small purple boxes labeled P<sub>1</sub>, P<sub>2</sub>, and P'<sub>1</sub>' are the arm  $\lambda$ Int-binding sites. Sites marked H are the integration host factor (IHF)-binding sites, and sites marked X are the sites that bind Xis. F is the site bound by Fis, another architectural protein not discussed further here. (Gray) The crossover regions. For clarity,  $\lambda$ Int is not shown bound to the core sites. Note that not all protein-binding sites are filled during either integrative or exciseive recombination. After recombination, the P arm is part of  $attL$ , whereas the P' arm becomes part of  $attR$ .



**FIGURE 12-11 Model for IHF bending DNA to bring DNA-binding sites together.** The  $\lambda$ Int and IHF-binding sites from the P' arm of  $attP$  are shown. IHF binding to the H' site bends the DNA to allow one molecule of  $\lambda$ Int to bind both the P'<sub>1</sub> and C' sites. The break in the DNA within the H' site reflects a nick that was present in the DNA used for structural analysis of the IHF–DNA complex. (Adapted, with permission, from Rice P. et al. 1996. *Cell* 87:1295–1306, Fig. 8. © Elsevier.)



described above). Whereas  $attB$  consists only of this central core region,  $attP$  is much longer (240 bp) and carries several additional protein-binding sites.

Flanking each side of the core region of  $attP$  are DNA regions known as the “arms.” These arms carry a variety of protein-binding sites, including additional sites bound by  $\lambda$ Int (labeled as P<sub>1</sub>, P<sub>2</sub>, and P'<sub>1</sub>' in Fig. 12-10).  $\lambda$ Int is an unusual protein because it has two domains involved in sequence-specific DNA binding: one domain binds to the arm recombinase recognition sites, and the other binds to the core recognition sites. In addition, the arms of  $attP$  carry sites bound by several architectural proteins. Binding of these proteins governs the directionality and efficiency of recombination.

Integration requires  $attB$ ,  $attP$ ,  $\lambda$ Int, and an architectural protein called **integration host factor (IHF)**. IHF is a sequence-dependent DNA-binding protein that introduces large bends ( $>160^\circ$ ) in DNA (Fig. 12-11). The arms of  $attP$  carry three IHF-binding sites (labeled H<sub>1</sub>, H<sub>2</sub>, and H' in Fig. 12-10). The function of IHF is to bring together the  $\lambda$ Int sites on the DNA arms (where  $\lambda$ Int binds strongly) with the sites present at the central core (where  $\lambda$ Int binds only weakly). Thus, bending of the DNA, mediated by IHF, allows  $\lambda$ Int to find the weak core sites and to catalyze recombination.

When recombination is complete, the circular phage genome is stably integrated into the host chromosome. As a result, two new hybrid sites are generated at the junctions between the phage and the host DNA. These sites are called  $attL$  (left) and  $attR$  (right) (see Fig. 12-10). Both of these sites contain the core region, but the two arm regions are now separated from each other (see the location of the P and P' regions in Fig. 12-10). Thus, neither of the two core regions in this new arrangement is competent to assemble

an active  $\lambda$ Int recombinase complex via the mechanism that was used to generate this complex during integration; the DNA sites important for assembly are simply not in the right place.

### Bacteriophage $\lambda$ Excision Requires a New DNA-Bending Protein

How does  $\lambda$  excise? An additional architectural protein, this one phage-encoded, is essential for excisive recombination. This protein, called Xis (for “excise”), binds to specific DNA sequences and introduces bends in the DNA. In this manner, Xis is similar in function to IHF. Xis recognizes two sequence motifs present in one arm of *attR* (and also present in *attP*—marked  $X_1$  and  $X_2$  in Fig. 12-10). Binding these sites introduces a large bend ( $>140^\circ$ ), and together Xis,  $\lambda$ Int, and IHF stimulate excision by assembling an active protein–DNA complex at *attR*. This complex then interacts productively with proteins assembled at *attL* and recombination occurs.

In addition to stimulating excision (recombination between *attL* and *attR*), DNA binding by Xis also inhibits integration (recombination between *attP* and *attB*). The DNA structure created upon Xis binding to *attP* is incompatible with proper assembly of  $\lambda$ Int and IHF at this site. Xis is a phage-encoded protein and is only made when the phage is triggered to enter lytic growth. Xis expression is described in detail in Chapter 18. Its dual action as a stimulatory co-factor for excision and an inhibitor of integration ensures that the phage genome will be free, and remain free, from the host chromosome when Xis is present.

### The Hin Recombinase Inverts a Segment of DNA Allowing Expression of Alternative Genes

The *Salmonella* Hin recombinase inverts a segment of the bacterial chromosome to allow expression of two alternative sets of genes. Hin recombination is an example of a class of recombination reactions, relatively common in bacteria, known as programmed rearrangements. These reactions often function to “preadapt” a portion of a population to a sudden change in the environment. In the case of Hin inversion, recombination is used to help the bacteria evade the host immune system, as we now explain.

The genes that are controlled by the inversion process encode two alternative forms of flagellin (called the H1 and H2 forms), the protein component of the flagellar filament. Flagella are on the surface of the bacteria and are thus a common target for the immune system (Fig. 12-12). By using Hin to switch between these alternative forms, at least some individuals in the bacterial population can avoid recognition of this surface structure by the immune system.

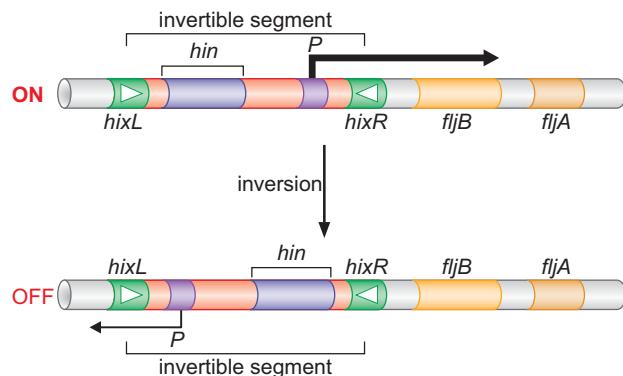
The chromosomal region inverted by Hin is  $\sim 1000$  bp and is flanked by specific recombination sites called *hixL* (on the left) and *hixR* (on the right) (Fig. 12-13). These sequences are in inverted orientation with respect to one another. Hin, a serine recombinase, promotes inversion using the basic mechanism described above for this enzyme family. The invertible segment carries the gene encoding Hin, as well as a promoter, which in one orientation is positioned to express the genes located outside of the invertible segment directly adjacent to the *hixR* site. When the invertible segment is in the “ON” orientation, these adjacent genes are expressed, whereas when the segment is flipped into the “OFF” orientation, the genes cannot be transcribed, because they lack a functional promoter.

The two genes under control of this “flipping” promoter are *fliB*, which encodes the H2 flagellin, and *fliA*, which encodes a transcriptional repressor



**FIGURE 12-12** Micrograph of bacteria (*Salmonella*) showing flagella. The color-enhanced scanning electron micrograph shows *Salmonella typhimurium* (red) invading cultured human cells. The hair-like protrusions on the bacteria are the flagella. (Courtesy of the Rocky Mountain Laboratories, NIAID, NIH.)

**FIGURE 12-13** DNA inversion by the Hin recombinase of *Salmonella*. Inversion of the DNA segment between the *hix* sites flips a promoter (*P*) to give two alternative patterns of flagellin gene expression.

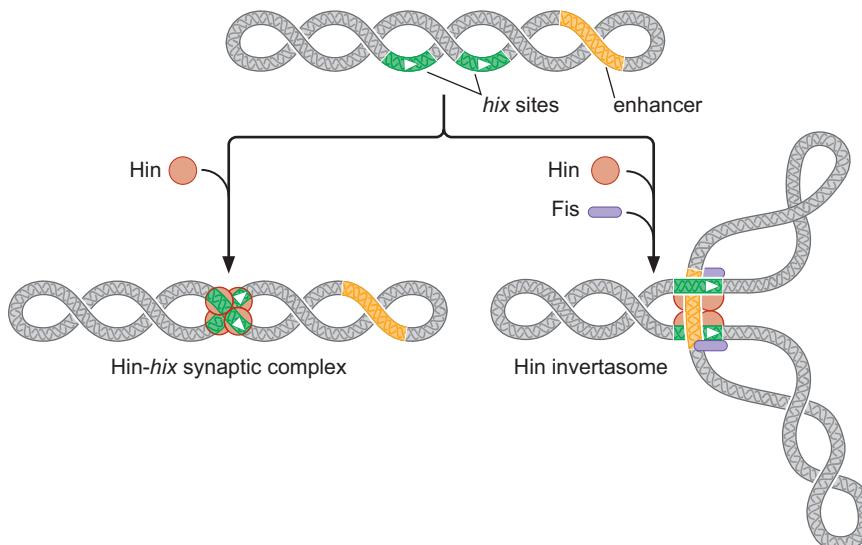


of the gene for the H1 flagellin. The H1 flagellin gene is located at a distant site. Thus, in the ON orientation, H2 flagellin and the H1 repressor are expressed. These cells have exclusively H2-type flagella on their surface. In the OFF orientation, however, neither H2 nor the H1 repressor is synthesized, and the H1-type flagella are present.

### Hin Recombination Requires a DNA Enhancer

Hin recombination requires a sequence in addition to the *hix* sites. This short (~60 bp) sequence is an enhancer that stimulates the rate of recombination ~1000-fold. Like enhancer sequences that stimulate transcription (see Chapter 19), this sequence can function even when located quite a distance from the recombination sites. Enhancer function requires the bacterial **Fis** protein (named because it was discovered as a **factor for inversion stimulation**). Like IHF, Fis is a site-specific DNA-bending protein. In addition, it makes protein–protein contacts with Hin that are important for recombination.

The Fis–enhancer complex activates the catalytic steps of recombination. Hin can actually assemble and pair the *hix* recombination sites to form a synaptic complex in the absence of the Fis–enhancer complex (Fig. 12-14). This contrasts with the role of IHF in  $\lambda$  integration, where the accessory protein is essential for assembly of the recombinase–DNA complex. For Fis activation of Hin, the three DNA sites (*hixL*, *hixR*, and en-



**FIGURE 12-14** Complexes formed during Hin-catalyzed recombination. Hin protein alone recognizes and pairs the two *hix* sites. When Fis protein is also present, the three-segment complex can form. This complex is called the invertasome and is the most active complex for promoting recombination. (Adapted, with permission, from Craig N. et al. 2002. *Mobile DNA II*, p. 246, Fig. 9. © ASM Press.)

hancer) need to come together. Formation of this three-way complex is greatly facilitated by negative DNA supercoiling (see Chapter 4), which stabilizes the association of the distant DNA sites. Another bacterial architectural protein, HU, also facilitates assembly of this invertasome complex. HU is a close structural homolog of IHF, yet in contrast to IHF, it binds DNA in a sequence-independent manner.

What is the biological rationale for control of Hin inversion by the Fis–enhancer complex? The principal function is to ensure that recombination occurs only between *hix* sites that are present on the same DNA molecule. This selectivity ensures that the invertible segment is flipped frequently but also that intermolecular DNA rearrangements, which could disrupt the integrity of the bacterial chromosome, are avoided.

In contrast to integration and excision of bacteriophage  $\lambda$ , Hin-catalyzed inversion is not highly regulated. Rather, inversion occurs stochastically, such that within a population of cells, there will always be some cells that carry the invertible segment in each orientation.

### Recombinases Convert Multimeric Circular DNA Molecules into Monomers

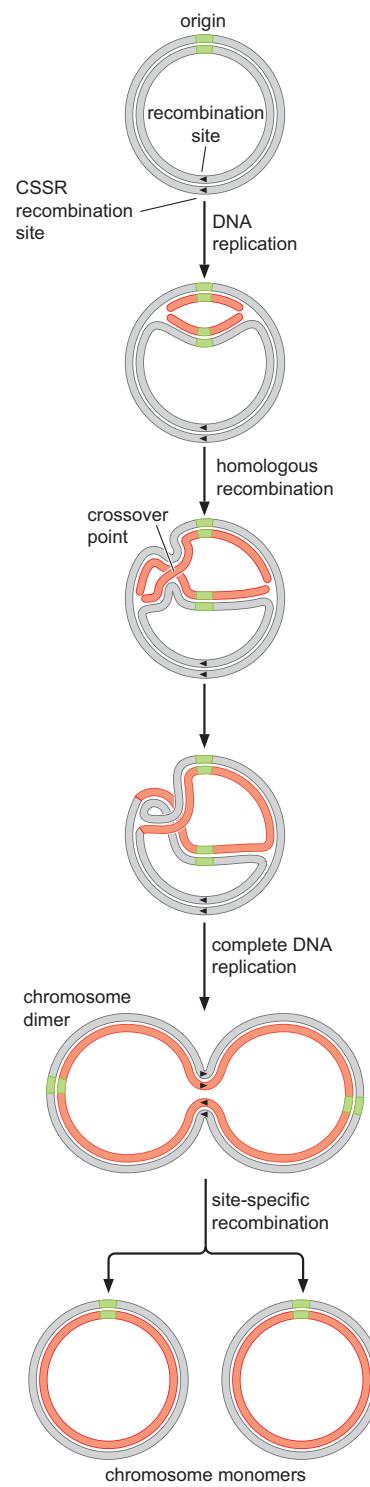
Site-specific recombination is critical to the maintenance of circular DNA molecules within cells. The chromosomes of most bacteria are circular, as are most plasmids in both prokaryotic and eukaryotic cells. Some viral genomes are also circular. An intrinsic problem with circular DNA molecules is that they sometimes form dimers and even higher multimeric forms during the process of homologous recombination. Site-specific recombination can be used to convert these DNA multimers back into monomers.

Consider what happens when a DNA crossover occurs between two identical circular molecules. This process is shown occurring between two copies of a bacterial chromosome during replication in Figure 12-15 (for a discussion of homologous recombination, see Chapter 11). A single homologous recombination event can generate one large circular chromosome with two copies of all of the genes (i.e., a dimeric chromosome). At the time of cell division, this dimer poses a major problem, because there will be only one rather than two DNA molecules to be segregated into the two daughter cells.

Because of this multimerization problem, many circular DNA molecules carry sequences recognized by site-specific recombinases. Proteins that function at these sequences are called **resolvases** because they “resolve” dimers (and larger multimers) into monomers. Clearly, it is essential that these proteins specifically catalyze resolution (a DNA deletion reaction) but not the reverse reaction (conversion of monomers to dimers), which would only make the multimerization problem worse! Specific mechanisms are in place to enforce this directional selectivity on the recombination process (see Box 12-2, The Xer Recombinase Catalyzes the Monomerization of Bacterial Chromosomes and of Many Bacterial Plasmids).

### There Are Other Mechanisms to Direct Recombination to Specific Segments of DNA

Although we have limited our discussion to conservative site-specific recombination, other recombination events occur at specific sequences and serve similar biological functions. Some of these reactions, for example, mating-type switching in yeast, occur by a targeted gene-conversion event,

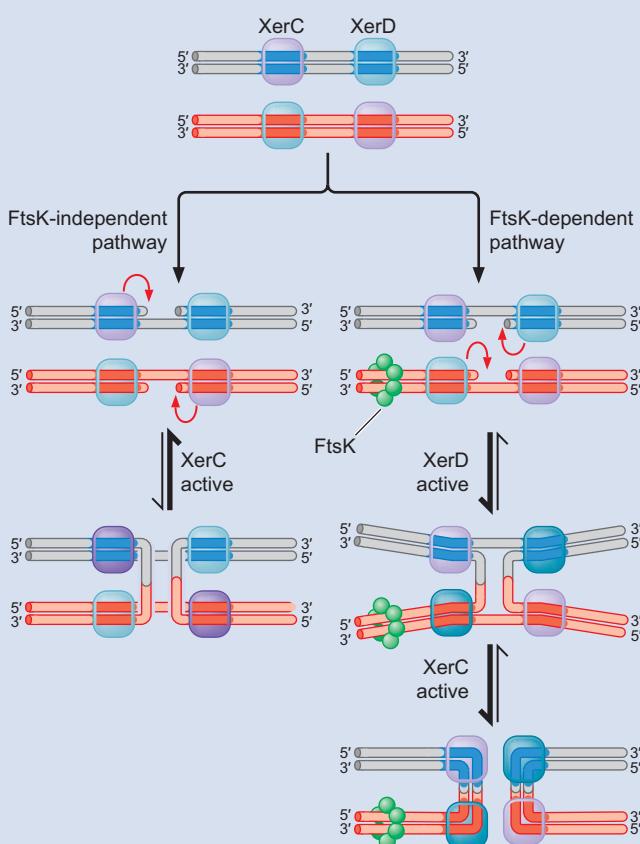


**FIGURE 12-15** Circular DNA molecules can form multimers. Homologous recombination between the two daughter DNA molecules during DNA replication generates a dimeric chromosome (or plasmid). Site-specific recombination by the XerCD recombinase is then needed to generate the monomeric DNA molecules needed for cell division.

## ► ADVANCED CONCEPTS

**Box 12-2** The Xer Recombinase Catalyzes the Monomerization of Bacterial Chromosomes and of Many Bacterial Plasmids

Xer is a member of the tyrosine recombinase family, and its mechanism for promoting recombination is very similar to that described above for Cre. Xer is a heterotetramer, containing two subunits of a protein called XerC and two subunits of a protein called XerD. Both XerC and XerD are tyrosine recombinases, but they recognize different DNA sequences. Therefore, the recombination sites used by the Xer recombinase must carry recognition sequences for each of these proteins. The



**BOX 12-2 FIGURE 1** Pathways for Xer-mediated recombination at *dif*. In the absence of FtsK (FtsK-independent pathway shown in the left panel), only XerC is active to promote strand exchange to form a Holliday junction intermediate. In this case (because XerD is not active), recombination is not completed, and the XerC reaction is frequently reversed. In the presence of FtsK (FtsK-dependent pathway shown in the right panel), XerD, now active, catalyzes formation of the Holliday junction intermediate, and XerC promotes second-strand exchange to complete the recombination event and generate chromosome monomers. (Adapted, with permission, from Aussel L. et al. 2002. *Cell* 108: 195–205, Fig. 6. © Elsevier.)

recombination sites in bacterial chromosomes, called *dif* sites, have a XerC recognition sequence on one side and an XerD recognition sequence on the other side of the crossover region (Box 12-2 Fig. 1). There is one *dif* site on the chromosome. It is located within the region where DNA replication terminates (see Chapter 9). When the chromosome is dimeric, this dimer will, of course, have two *dif* sites (see Fig. 12-15).

How do cells make sure that Xer-mediated recombination at *dif* sites will convert a chromosome dimer into monomers without ever promoting the reverse reaction? This directional regulation is achieved through the interaction between the Xer recombinase and a cell division protein called FtsK. This regulation is shown in Box 12-2 Figure 2 and occurs as follows. When FtsK is unavailable for interaction with the XerCD complex at the *dif* site, the recombinase complex adopts a conformation in which only the two XerC subunits are active. As a result, XerC will promote exchange of one pair of DNA strands to form the Holliday junction intermediate (see the discussion of the tyrosine recombinase mechanism above). Because XerD is never activated, recombination is never completed. Instead, reversal of the XerC cleavage reaction often occurs. This reversal simply regenerates the original DNA arrangement (see Box 12-2 Fig. 1).

In contrast, when the FtsK protein is available and interacts with the XerCD complex, it alters the conformation of the complex and activates XerD protein. In this case, XerD promotes recombination of the first pair of strands to generate the Holliday junction intermediate. Once this reaction is completed, XerC promotes the second pair of strand-exchange reactions, yielding the recombined DNA products (see Box 12-2 Fig. 1).

FtsK is an ATPase that tracks along DNA. It functions as a “DNA-pumping protein machine” similar to the RuvB protein that promotes DNA branch migration during homologous recombination (discussed in Chapter 11). FtsK is also a membrane-bound protein that is localized in the cell at the site where cell division occurs. It moves DNA away from the center of the cell before division so that the cell can divide at this site.

This localization of FtsK to the division site is key to how the cells ensure that XerD is activated specifically when a dimeric chromosome is present. In this case, the chromosome will be “stuck” in the middle of the dividing cell as one-half of the chromosome dimer is moved into each daughter cell. FtsK also interacts with specific polar DNA sequences (called KOPS) that are arranged asymmetrically around the *dif* site. As a result, FtsK translocates *dif* sites toward the septum and toward each other. This movement therefore facilitates their pairing, as well as activating XerD recombination. In this manner, site-specific recombination is regulated to occur at the right time and place within the cell division cycle.

as we described in Chapter 11. The gene rearrangements responsible for assembly of gene segments encoding critical proteins for the vertebrate immune system—known as V(D)J recombination—also occur at specific sites. This reaction is mechanistically similar to transposition and therefore is considered later in this chapter.

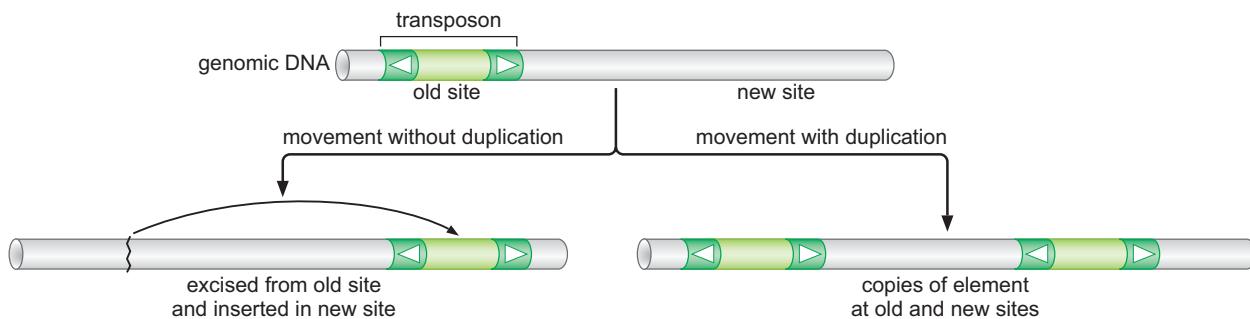
## TRANSPOSITION

### Some Genetic Elements Move to New Chromosomal Locations by Transposition

**Transposition** is a specific form of genetic recombination that moves certain genetic elements from one DNA site to another. These mobile genetic elements are called **transposable elements** or **transposons**. Movement occurs through recombination between the DNA sequences at the very ends of the transposable element and a sequence in the DNA of the host cell (Fig. 12-16); movement can occur with or without duplication of the element, as we shall see. In some cases, the recombination reaction involves a transient RNA intermediate.

When transposable elements move, they often show little sequence selectivity in their choice of insertion sites. As a result, transposons can insert within genes, often completely disrupting gene function. They can also insert within the regulatory sequences of a gene, where their presence may lead to changes in how that gene is expressed. It was these disruptions in gene function and expression that led to the discovery of transposable elements (see Box 12-3, Maize Elements and the Discovery of Transposons). Perhaps not surprisingly, therefore, transposable elements are the most common source of new mutations in many organisms. In fact, these elements are an important cause of mutations leading to genetic disease in humans. The ability of transposable elements to insert so promiscuously in DNA has also led to their modification and use as mutagens and DNA-delivery vectors in experimental biology.

Transposable elements are present in the genomes of all life-forms. The comparative analysis of genome sequences reveals two fascinating observations. First, transposon-related sequences can make up huge fractions of the genome of an organism. For example, more than 50% of both the human and maize genomes are composed of transposon-related sequences (including

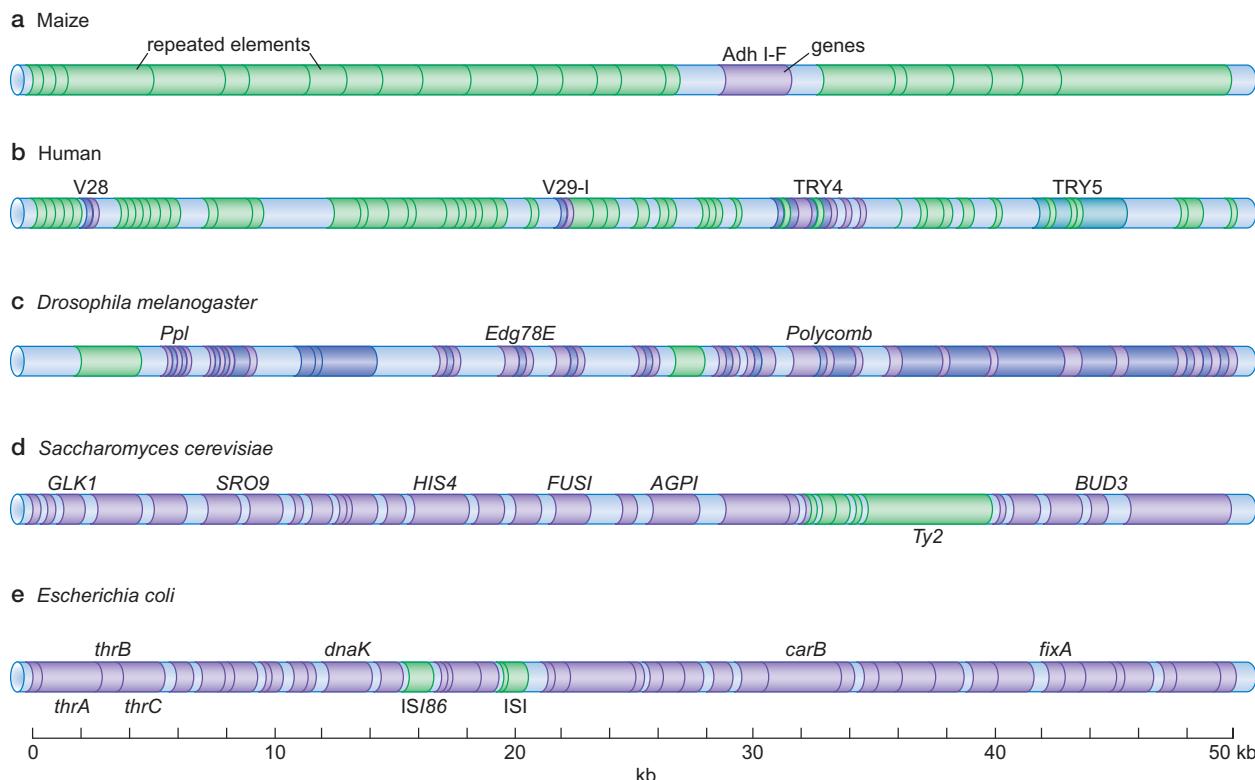


**FIGURE 12-16** Transposition of a mobile genetic element to a new site in the host DNA. Recombination, in some cases, involves excision of the transposon from the old DNA location (left). In other cases, one copy of the transposon stays at the old location, and another copy is inserted into the new DNA site (right).

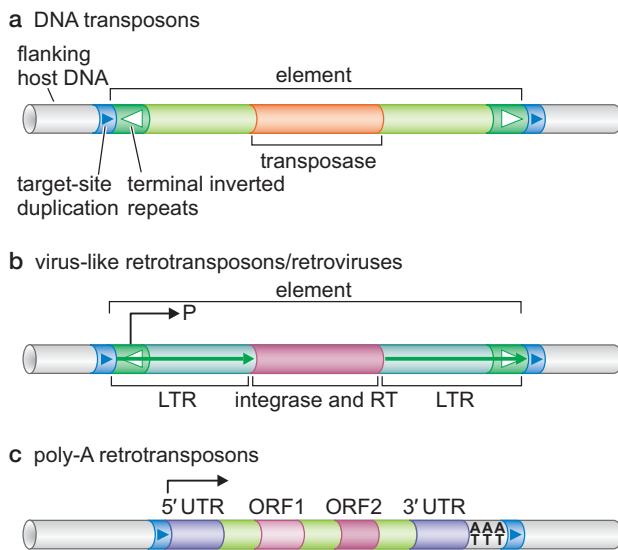
fragments of transposons or “dead” elements that have been inactivated by mutations). This contribution is in sharp contrast to the small percentage of the sequence that actually encodes cellular proteins (<2% in human). Second, the transposon content in different genomes is highly variable (Fig. 12-17). For example, compared with humans or maize, the fly and yeast genomes are very “gene-rich” and “transposon-poor.”

There are many different types of transposable elements. These elements can be divided into families that share common aspects of structure and recombination mechanism. In the following sections, we introduce three major families of transposable elements and the recombination mechanism associated with each family. Some of the best-studied individual elements are then described. In the description of individual elements, we focus on how transposition is regulated to balance the maintenance and propagation of these elements with their potential to disrupt or misregulate genes within the host organism.

The genetic recombination mechanisms responsible for transposition are also used for functions other than the movement of transposons. For example, many viruses use a recombination pathway nearly identical to transposition to integrate into the genome of the host cell during infection. These viral integration reactions will therefore be considered together with transposition. Likewise, some DNA rearrangements used by cells to alter gene expression occur using a mechanism very similar to DNA transposition. V(D)J recombination, a reaction required for development of a functional immune system in vertebrates, is a well-understood example. V(D)J recombination is discussed at the end of this chapter.



**FIGURE 12-17** Transposons in genomes: occurrence and distribution. (Green) Repeated elements, mostly composed of transposons or transposon-related sequences (such as truncated elements); (purple) cellular genes. (a) Maize; (b) human; (c) *Drosophila*; (d) budding yeast; (e) *Escherichia coli*. (With permission from Brown T.A. 2002. *Genomes*, 2nd ed., p. 34, Fig. 2.2, and references therein. © Taylor & Francis.)



**FIGURE 12-18** Genetic organization of the three classes of transposable elements. (a) DNA transposons. The element includes the terminal inverted-repeat sequences (green with white arrows), which are the recombination sites, and a gene encoding transposase. (b) Virus-like retrotransposons and retroviruses. The element includes two LTR sequences that flank a region encoding two enzymes: integrase and reverse transcriptase (RT). (c) Poly-A retrotransposons. The element terminates in the 5'- and 3'-untranslated region (UTR) sequences and encodes two enzymes: an RNA-binding enzyme (ORF1) and an enzyme having both reverse transcriptase and endonuclease activities (ORF2).

### There Are Three Principal Classes of Transposable Elements

Transposons can be divided into the following three families on the basis of their overall organization and mechanism of transposition:

1. *DNA transposons*.
2. *Virus-like retrotransposons*. This class includes the retroviruses. These elements are also called **long terminal repeat (LTR)** retrotransposons.
3. *Poly-A retrotransposons*. These elements are also called non-viral retrotransposons.

Figure 12-18 shows a schematic of the general genetic organization of each of these element families. DNA transposons remain as DNA throughout a cycle of recombination. They move using mechanisms that involve the cleavage and rejoining of DNA strands, and in this way, they are similar to elements that move by conservative site-specific recombination. Both types of retrotransposons move to a new DNA location using a transient RNA intermediate.

### DNA Transposons Carry a Transposase Gene, Flanked by Recombination Sites

DNA transposons carry both DNA sequences that function as recombination sites and genes encoding proteins that participate in recombination (Fig. 12-18a). The recombination sites are at the two ends of the element and are organized as inverted-repeat sequences. These terminal inverted repeats vary in length from ~25 bp to a few hundred base pairs, are not exact sequence repeats, and carry the recombinase recognition sequences. The recombinases responsible for transposition are usually called **transposases** (or, sometimes, **integrases**).

DNA transposons carry a gene encoding their own transposase. They may carry a few additional genes, sometimes encoding proteins that regulate transposition or provide a function useful to the element or its host cell. For example, many bacterial DNA transposons carry genes encoding proteins that promote resistance to one or more antibiotic(s). The presence of the transposon therefore causes the host cell to be resistant to that antibiotic.

The DNA sequences immediately flanking the transposon have a short (2–20 bp) segment of duplicated sequence. These segments are organized as direct repeats, are called **target-site duplications**, and are generated during the process of recombination, as we shall discuss later.

### Transposons Exist as Both Autonomous and Nonautonomous Elements

DNA transposons that carry a pair of terminal inverted repeats and a transposase gene have everything they need to promote their own transposition. These elements are called **autonomous transposons**. However, genomes also contain many even simpler mobile DNA segments known as **nonautonomous transposons**. These elements carry only the terminal inverted repeats, that is, the *cis*-acting sequences needed for transposition. In a cell that also carries an autonomous transposon, encoding a transposase that will recognize these terminal inverted repeats, the nonautonomous element will be able to transpose. However, in the absence of this “helper” transposon (to donate the transposase), nonautonomous elements remain frozen, unable to move.

### Virus-Like Retrotransposons and Retroviruses Carry Terminal Repeat Sequences and Two Genes Important for Recombination

Virus-like retrotransposons and retroviruses also carry inverted terminal repeat sequences that are the sites of recombinase binding and action (Fig. 12-18b). The terminal inverted repeats are embedded within longer repeated sequences; these sequences are organized on the two ends of the element as direct repeats and are called long terminal repeats (LTRs). Virus-like retrotransposons encode two proteins needed for their mobility: integrase (the transposase) and reverse transcriptase.

**Reverse transcriptase (RT)** is a special type of DNA polymerase that can use an RNA template to synthesize DNA. This enzyme is needed for transposition because an RNA intermediate is required for the transposition reaction. Because these elements convert RNA into DNA, the reverse of the normal pathway of biological information flow (DNA to RNA), they are known as “retro” elements. The distinction between virus-like retrotransposons and retroviruses is that the genome of a retrovirus is packaged into a viral particle, escapes its host cell, and infects a new cell. In contrast, the retrotransposons can move only to new DNA sites within a cell but can never leave that cell. Like the DNA transposons, these elements are flanked by short target-site duplications that are generated during recombination.

### Poly-A Retrotransposons Look Like Genes

The poly-A retrotransposons do not have the terminal inverted repeats present in the other transposon classes. Instead, the two ends of the element have distinct sequences (Fig. 12-18c). One end is called the 5'-UTR, whereas the other end has a region called the 3'-UTR followed by a stretch of A:T base pairs called the **poly-A sequence**. These elements are also flanked by short target-site duplications.

Retrotransposons carry two genes, known as *ORF1* and *ORF2*. *ORF1* encodes an RNA-binding protein. *ORF2* encodes a protein with both reverse transcriptase activity and an endonuclease activity. This protein, although distinct from the transposases and integrases encoded by the

other classes of elements, has essential roles during recombination. Like their DNA and virus-like transposon counterparts, poly-A retrotransposons exist commonly in both autonomous and nonautonomous forms. Furthermore, genome sequence analysis reveals that there are many truncated elements that do not have a complete 5'-UTR sequence and have lost their ability to transpose.

### DNA Transposition by a Cut-and-Paste Mechanism

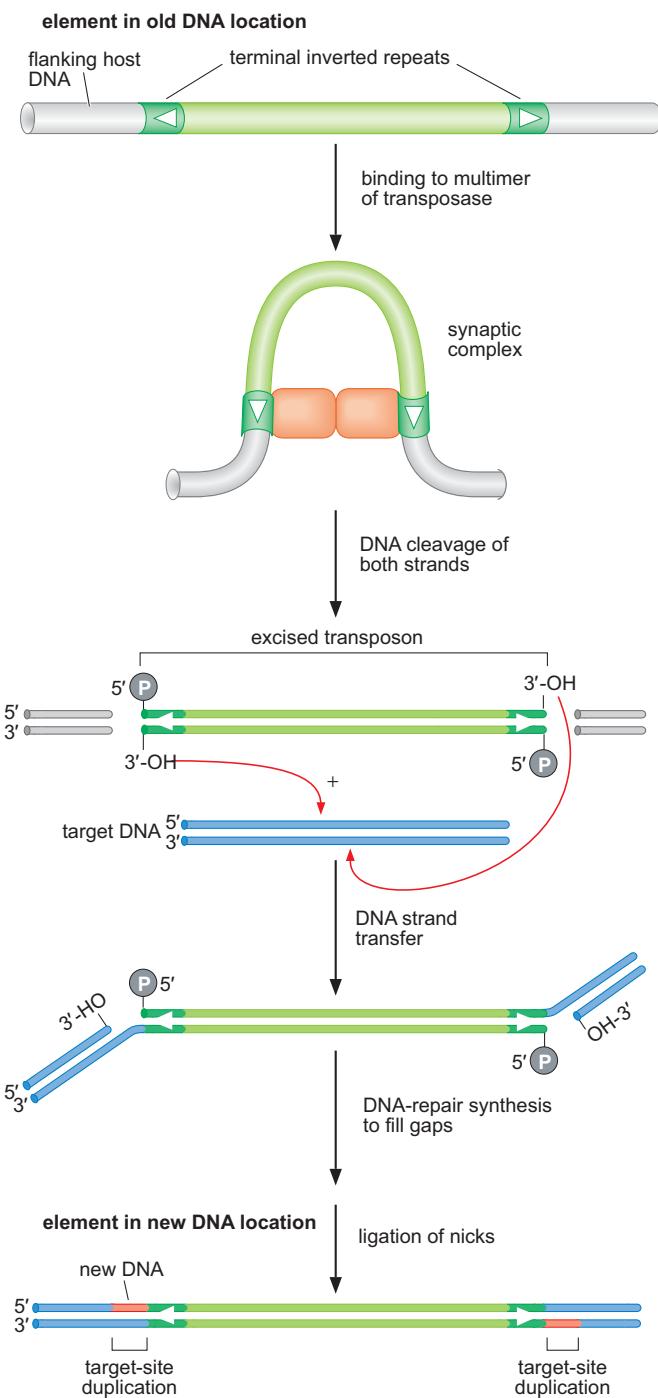
DNA transposons, virus-like retrotransposons, and retroviruses all use a similar mechanism of recombination to insert their DNA into a new site. First, let us consider the simplest transposition reaction: the movement of a DNA transposon by a nonreplicative mechanism. This recombination pathway involves the excision of the transposon from its initial location in the host DNA, followed by integration of this excised transposon into a new DNA site. This mechanism is therefore called **cut-and-paste transposition** (Fig. 12-19).

To initiate recombination, the transposase binds to the terminal inverted repeats at the end of the transposon. Once the transposase recognizes these sequences, it brings the two ends of the transposon DNA together to generate a stable protein–DNA complex. This complex is called the **synaptic complex** or **transpososome**. It contains a multimer of transposase—usually two or four subunits—and the two DNA ends (see later discussion). This complex functions to ensure that the DNA cleavage and joining reactions needed to move the transposon occur simultaneously on the two ends of the element’s DNA. It also protects the DNA ends from cellular enzymes during recombination. The next step is the excision of the transposon DNA from its original location in the genome. To achieve this, the transposase subunits within the transpososome first cleave one DNA strand at each end of the transposon, exactly at the junction between the transposon DNA and the host sequence in which it is inserted (a region called the **flanking host DNA**). The transposase cleaves the DNA such that the transposon sequence terminates with free 3'-OH groups at each end of the element’s DNA. To finish the excision reaction, the other DNA strand at each end of the element must also be cleaved. Different transposons use different mechanisms to cleave these “second” DNA strands (those strands that terminate with 5' ends at the transposon host DNA junction). These mechanisms are described in a following section.

After excision of the transposon, the 3'-OH ends of the transposon DNA—the ends first liberated by the transposase—attack the DNA phosphodiester bonds at the site of the new insertion. This DNA segment is called the **target DNA**. Recall that for most transposons, the target DNA can have essentially any sequence. As a result of this attack, the transposon DNA is covalently joined to the DNA at the target site. During each DNA-joining reaction, a nick is also introduced into the target DNA (Fig. 12-19). This DNA-joining reaction occurs by a one-step transesterification reaction that is called **DNA strand transfer** (Fig. 12-20). A similar mechanism for joining nucleic acid strands is used for RNA splicing (see Chapter 14).

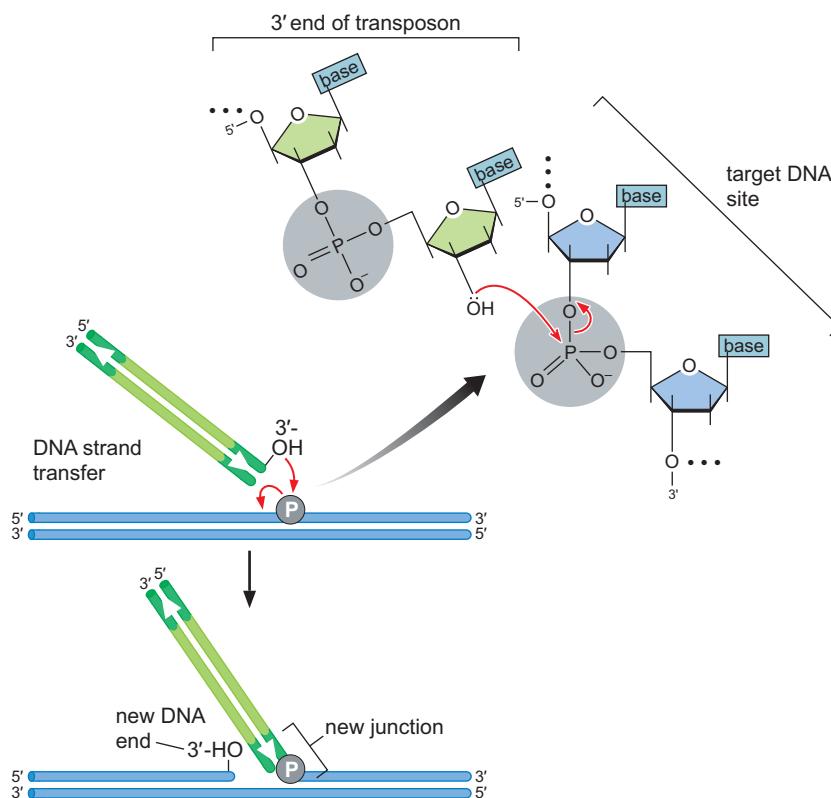
The transpososome ensures that the two ends of the transposon DNA attack the two DNA strands of the same target site together. The sites of attack on the two strands are usually separated by a few nucleotides (e.g., 2-, 5-, and 9-nucleotide spacings are common). This distance is fixed for each type of transposon and gives rise to the short target-site duplications that flank transposed copies of the element (as explained in the next section). Once DNA strand transfer is complete, the job of the transpososome is also complete. The remaining recombination steps are performed by cellular DNA-repair proteins.

**FIGURE 12-19** The cut-and-paste mechanism of transposition. Movement of a transposon from a target site in the (gray) host DNA to a new site in the (blue) DNA. Note the staggered cleavage sites on the target DNA during the DNA strand transfer reaction that give rise to short repeated sequences at the new target site (the target-site duplications). The DNA at the original insertion site (here in gray) will be left with a double-stranded DNA break as a result of transposon excision. This break can be repaired by nonhomologous end joining or homologous recombination (see Chapters 10 and 11). Note that the chemical steps are shown without the bound protein for clarity.



### The Intermediate in Cut-and-Paste Transposition is Finished by Gap Repair

The structure of the DNA intermediate generated after DNA strand transfer has the 3' ends of the transposon DNA attached to the target DNA. This structure also carries the two nicks in the target DNA that were generated during the process of DNA strand transfer. The fact that the two sites of DNA strand transfer on the two strands are separated by a few nucleotides results in short single-stranded DNA gaps flanking the joined transposon. These gaps are filled by a DNA-repair polymerase encoded by the host cell. Note that



**FIGURE 12-20** Close-up view of the chemical step of DNA strand transfer. In the inset, only one strand is shown for the transposon and for the target DNA for clarity.

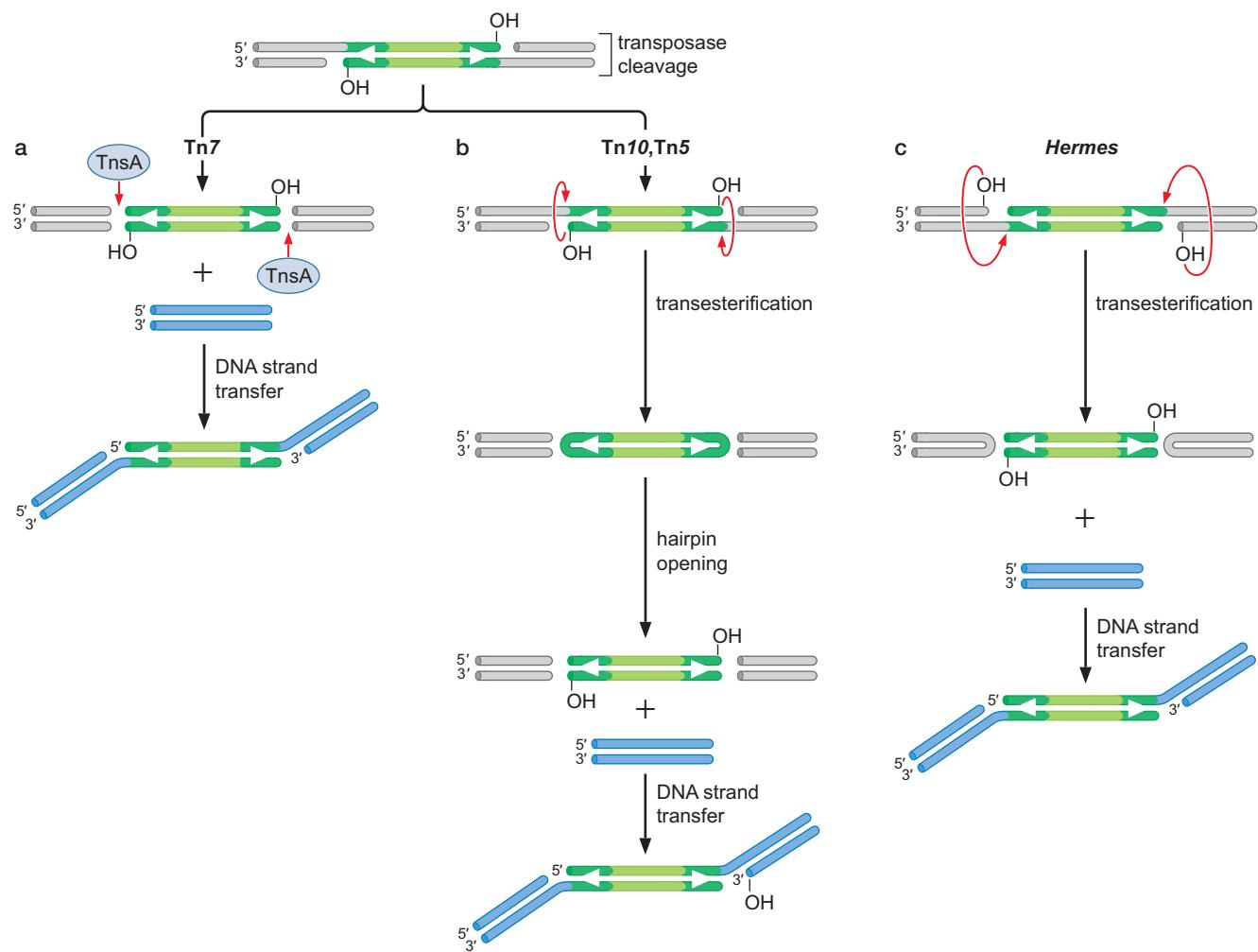
the target DNA is cleaved during the DNA strand transfer step to generate 3'-OH ends that can serve as the primers for this repair synthesis (see Fig. 12-18). Filling in the gaps gives rise to the target-site duplications that flank transposons (see above). Thus, the length of the target-site duplication reveals the distance between the sites attacked on the two strands of the target DNA during DNA strand transfer. After gap-repair synthesis, DNA ligase is needed to seal the DNA strands.

Cut-and-paste transposition also leaves a double-stranded break in the DNA at the site of the “old” insertion, which must be repaired to maintain the integrity of the host cell’s genome. Repair of double-stranded DNA breaks by homologous recombination is described in Chapter 11. These breaks are also sometimes more directly rejoined, as we shall see later in the discussion of the *Tc1/mariner* family of transposons.

### There Are Multiple Mechanisms for Cleaving the Nontransferred Strand during DNA Transposition

As described above, the transposase cleaves the 3' ends of the element DNA and promotes DNA strand transfer to catalyze cut-and-paste transposition. However, transposons that move by this mechanism also need to cleave the 5'-terminating strands at the junctions between the transposon and the flanking host DNA. These DNA strands are called the **nontransferred strands**, because their 5' ends are not directly linked to the target DNA during the DNA strand transfer reaction. Different transposons use different mechanisms to catalyze this second-strand cleavage reaction (Fig. 12-21). Two methods are described here.

An enzyme other than the transposase can be used to cleave the nontransferred strand (Fig. 12-21). For example, the bacterial transposon *Tn7* en-



**FIGURE 12-21** Three mechanisms for cleaving the nontransferred strand. (a) An enzyme other than transposase is used. (b) The transposase catalyzes the attack of one DNA strand on the opposite strand to form the DNA–hairpin intermediate. In this case, attack is of the transferred strand on the nontransferred strand. The two hairpin ends are subsequently hydrolyzed by the transposase. (c) The *Hermes* transposon uses a second mechanism of second-strand cleavage by hairpin formation. In this case, cleavage of the top strand (nontransferred strand) occurs first, and the hairpins are generated on the original insertion site DNA, rather than at the transposon ends.

codes a specific protein (called TnsA) that does this job (Fig. 12-21a). TnsA has a structure very similar to that of a restriction endonuclease. TnsA assembles with the Tn7-encoded transposase (the TnsB protein). By working together, the transposase and TnsA excise the transposon from its original target site.

The other way of cleaving the nontransferred strand is promoted by the transposase itself, using a DNA transesterification mechanism that is similar to DNA strand transfer. For example, the transposons Tn5 and Tn10 cleave the nontransferred strand by generating a structure known as a “DNA hairpin.” To form this hairpin, the transposase uses the initially cleaved 3'-OH end of the transposon DNA to attack a phosphodiester bond directly across the DNA duplex on the opposite strand (Fig. 12-21b). This reaction both cleaves the attacked DNA strand and covalently joins the 3' end of the transposon DNA to one side of the break. As a result, the two DNA strands are covalently joined by a looped end, reminiscent in shape to a hairpin.

This hairpin DNA end is then cleaved (i.e., “opened”) by the transposase to generate a standard double-strand break in the DNA. The opening reaction occurs on both ends of the transposon DNA. Once these steps are complete, the 3'-OH ends of the element DNA are ready to be joined to a new target DNA by the DNA strand transfer reaction.

The *Hermes* transposon, a member of the *hAT* family of elements, also uses DNA-hairpin intermediates to excise the transposon from the old DNA insertion site. However, in this case, the order of the strand cleavage and transesterification reactions is different, such that the DNA hairpins are formed in the host cell’s DNA, rather than on the end of the transposable element (Fig. 12-21c). As we shall see later in the chapter, this pathway of DNA cleavage and joining reactions is highly reminiscent of that observed during the early steps of V(D)J recombination. This mechanistic similarity strongly supports the hypothesis that V(D)J recombination arose from the capture and “taming” of a transposon by a host organism during vertebrate evolution.

Although not shown in Figure 12-21, DNA cleavage via a transesterification reaction can also occur *between* the two ends of the transposon. In this case, one cleaved 3'-OH end attacks the DNA strand at the opposite end of the element’s DNA and the resulting DNA intermediate is further processed to generate the excised transposon. IS3 family transposons use this mechanism.

Why might transposases use transesterification as a cleavage mechanism? It is probably an economic solution. Transposases have the intrinsic ability to promote (1) site-specific hydrolysis of the 3' ends of the transposon DNA and (2) transesterification of this end into a nonspecific DNA site. These same activities, with the transesterification reaction simply applied to a new DNA site (e.g., the strand opposite the initial cleavage site), can allow the transposase to promote transposon excision. This mechanism therefore avoids the need for the transposon to encode a second enzyme to cleave the nontransferred strand.

### DNA Transposition by a Replicative Mechanism

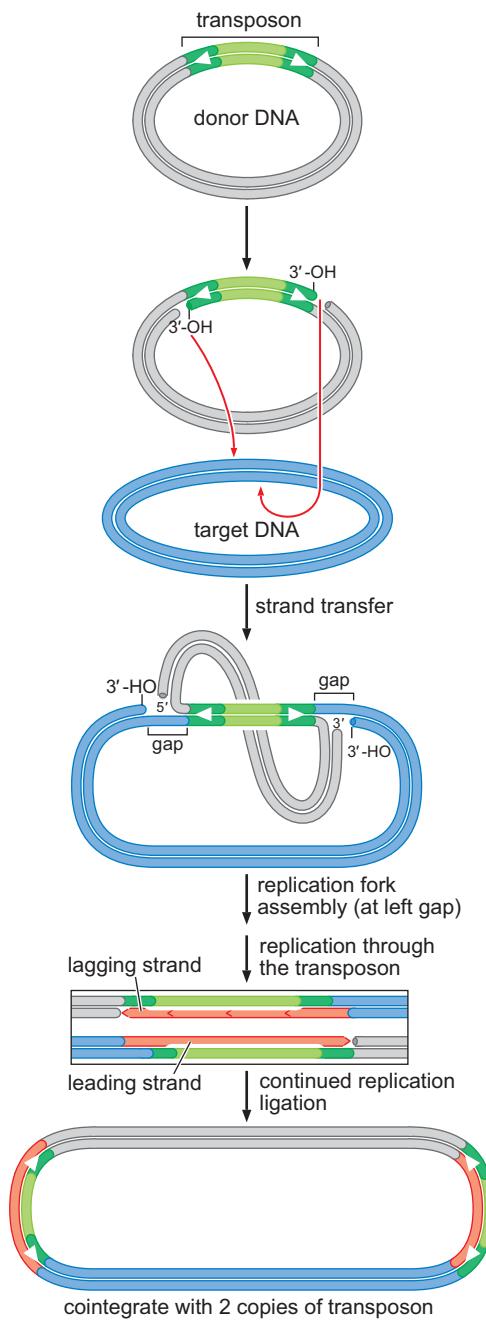
Some DNA transposons move using a mechanism called **replicative transposition**, in which the element DNA is duplicated during each round of transposition. Although the products of the transposition reaction are clearly different, as we shall now see, the mechanism of recombination is very similar to that used for cut-and-paste transposition (Fig. 12-22).

The first step of replicative transposition is the assembly of the transposase protein on the two ends of the transposon DNA to generate a transpososome. As we saw for cut-and-paste transposition, transpososome formation is essential to coordinate the DNA cleavage and joining reactions on the two ends of the transposon’s DNA.

The next step is DNA cleavage at the ends of the transposon DNA. This reaction is catalyzed by the transposase within the transpososome. The transposase introduces a nick into the DNA at each of the two junctions between the transposon sequence and the flanking host DNA (see Fig. 12-22). This cleavage liberates two 3'-OH DNA ends on the transposon sequence. In contrast to cut-and-paste transposition, the transposon DNA is not excised from the host sequences at this stage. This is the major difference between replicative and cut-and-paste transposition.

The 3'-OH ends of the transposon DNA are then joined to the target DNA site by the DNA strand transfer reaction. The mechanism is the same as discussed above for cut-and-paste transposition. However, the intermediate generated by DNA strand transfer is in this case a doubly branched DNA molecule (see Fig. 12-22). In this intermediate, the 3' ends of the transposon

**FIGURE 12-22 Mechanism for replicative transposition.** The transpososome introduces a single-strand nick at each of the ends of the transposon DNA. This cleavage generates a 3'-OH group at each end. These OH groups then attack the target DNA and become joined to the target by DNA strand transfer. Note that at each end of the transposon, only one strand is transferred into the target at this point, resulting in the formation of a doubly branched DNA structure. The replication apparatus assembles at one of these “forks” (the left one in the figure). Replication continues through the transposon sequence. The resulting product, called a cointegrate, has the two starting circular DNA molecules joined by two copies of the transposon. The single-stranded DNA gaps in the branched intermediate give rise to the target-site duplications. These duplications are not shown in the cointegrate for clarity.



are covalently joined to the new target site, whereas the 5' ends of the transposon sequence remain joined to the old flanking DNA.

The two DNA branches within this intermediate have the structure of a replication fork (see Chapter 9). After DNA strand transfer, the DNA replication proteins from the host cell can assemble at these forks. In the best-understood example of replicative transposition (phage Mu, which we shall discuss later), this assembly specifically occurs at only one of the two forked structures (see Fig. 12-22, bottom panels). The 3'-OH end in the cleaved target DNA serves as a primer for DNA synthesis. Replication proceeds through the transposon sequence and stops at the second fork. This replication reaction generates two copies of the transposon DNA. These copies are flanked by the short direct target-site duplications.

Replicative transposition frequently causes chromosomal inversions and deletions that can be highly detrimental to the host cell. This propensity to cause rearrangements may put replicative transposons at a selective disadvantage. Perhaps this is why so many elements have developed ways to excise completely from their original DNA location before joining to a new DNA site. By excision, transposons avoid generating these major disruptions to the host genome. As we describe later, transposition via an RNA intermediate also avoids generating these disruptive re-arrangements.

### Virus-Like Retrotransposons and Retroviruses Move Using an RNA Intermediate

Virus-like retrotransposons and retroviruses insert into new sites in the genome of the host cell, using the same steps of DNA cleavage and DNA strand transfer we have described for the DNA transposons. In contrast to the DNA transposons, however, recombination for these retroelements involves an RNA intermediate.

A cycle of transposition starts with transcription of the retrotransposon (or retroviral) DNA sequence into RNA by a cellular RNA polymerase. Transcription initiates at a promoter sequence within one of the LTRs (Fig. 12-23) and continues across the element to generate a nearly full-length RNA copy of the element's DNA. The RNA is then reverse-transcribed to generate a double-stranded DNA molecule. This DNA molecule is called the cDNA (for "copied DNA") and is free from any flanking host DNA sequences.

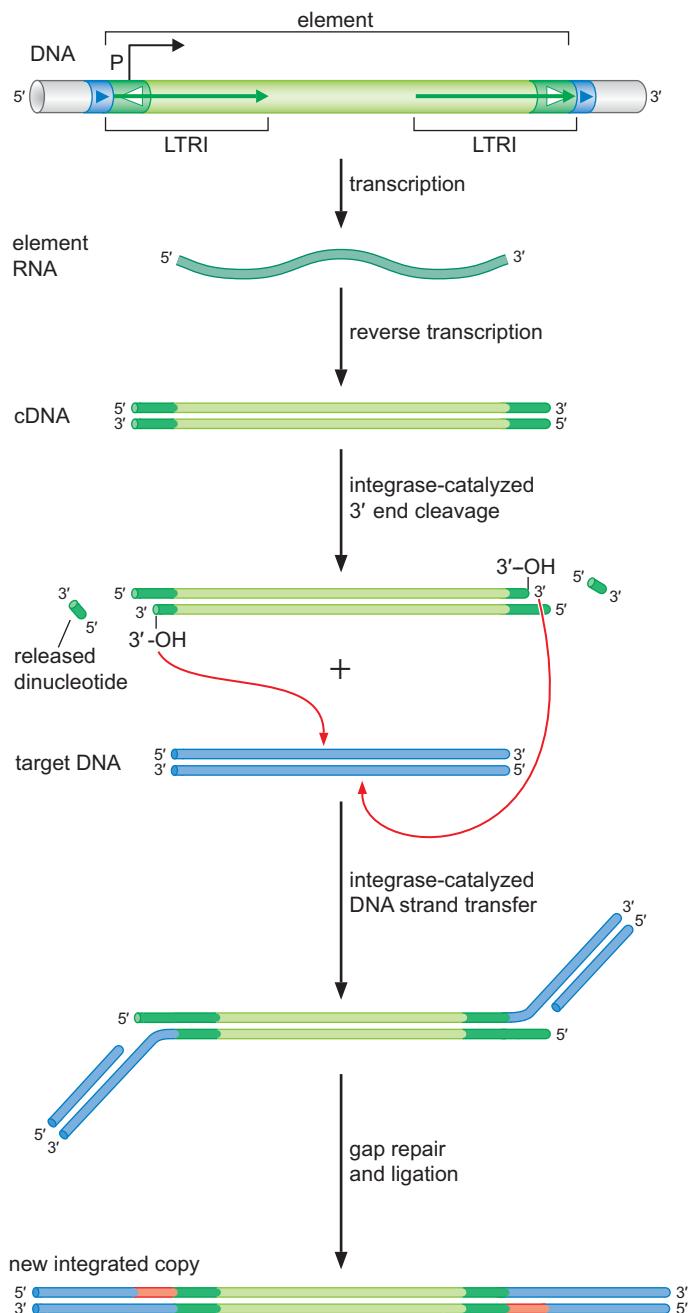
It is the cDNA that is recognized by the integrase protein (a protein highly related to the transposases of DNA elements, as we shall see later) for recombination with a new target DNA site. Integrase assembles on the ends of this cDNA and then cleaves a few nucleotides off the 3' end of each strand. This cleavage reaction is identical to the DNA cleavage step of DNA transposition. Because the direct precursor DNA for integration is generated from the RNA template by reverse transcription, it is already in the form of an excised transposon. Therefore, a mechanism to cleave the second strand is unnecessary for these elements. Integrase then catalyzes the insertion of these cleaved 3' ends into a DNA target site in the host-cell genome using the DNA strand transfer reaction. As we discussed above, this target site can have essentially any DNA sequence. Host-cell DNA-repair proteins fill the gaps at the target site generated during DNA strand transfer to complete recombination. This gap-repair reaction generates the target-site duplications.

Because transcription to generate the RNA intermediate initiates within one of the LTRs, this RNA does not carry the entire LTR sequence; the sequence between the transcription start site and the end of the element is missing. Therefore, a special mechanism is needed to regenerate the full-length element sequence during reverse transcription. The pathway of reverse transcription involves two internal priming events and two strand switches. These switching events result in the duplication of sequences at the ends of the cDNA. Thus, the cDNA has complete reconstructed LTR sequences to compensate for regions of sequence lost during transcription. This reconstruction of the LTRs is essential for recognition of the cDNA by integrase and for subsequent recombination.

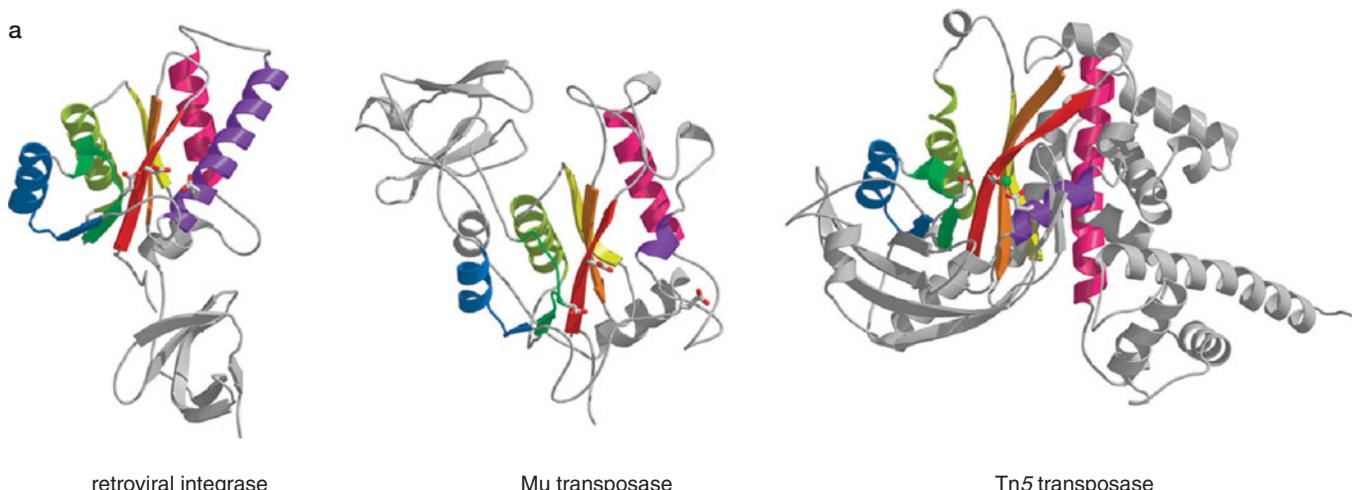
### DNA Transposases and Retroviral Integrases Are Members of a Protein Superfamily

As we have seen, DNA cleavage of the 3' ends of the transposon DNA (or cDNA) and DNA strand transfer are common steps used for DNA

**FIGURE 12-23** Mechanism of retroviral integration and transposition of virus-like retrotransposons. (Top panel) The integrated provirus. For a more detailed view of the LTR sequences, see the figures in Box 12-3. The promoter for transcription of the viral RNA is embedded in the left LTR as shown. cDNA synthesis from this viral RNA is explained in Box 12-3. The integrase-catalyzed DNA cleavage and DNA strand transfer steps are shown.



transposition and the movement of virus-like retrotransposons and retroviruses. This conserved recombination mechanism is reflected in the structure of the transposase/integrase proteins (Fig. 12-24). High-resolution structures reveal that many different transposases and integrases carry a catalytic domain that has a common three-dimensional (3D) shape. This catalytic domain contains three evolutionarily invariant acidic amino acids: two aspartates (D) and a glutamate (E). Therefore, recombinases of this class are referred to as DDE-motif transposase/integrase proteins. These acidic amino acids form part of the active site and coordinate divalent metal ions (such as  $Mg^{2+}$  or  $Mn^{2+}$ ) that are required for activity (as described for the DNA polymerases, see Chapter 9). An unusual feature of the transposase/integrase proteins is that they use this same active site to catalyze both the DNA



**FIGURE 12-24** Similarities of catalytic domains of transposases and integrases. (a) Structures of the conserved core domains (shown right to left) of Tn5 transposase (Davies D.R. et al. 2000. *Science* **289**: 77–85), of phage Mu transposase (Rice P. and Mizuuchi K. 1995. *Cell* **82**: 209–220), and of RSV integrase (Chook Y.M. et al. 1994. *J. Mol. Biol.* **240**: 476–500). Common secondary structure elements are shown in the same colors. The DDE-motif active-site residues are shown in stick form. Images prepared with MolScript, BobScript, and Raster3D. (b) Schematic of the domain organization of the three proteins shown in part a. The amino-terminal domains bind to the element DNA. The middle domains contain the catalytic regions shown in a. The carboxy-terminal domains are involved in protein–protein contacts needed to assemble the transpososome and/or to interact with other proteins that regulate transposition. (Derived from Rice P.A. and Baker T.A. 2001. *Nat. Struct. Biol.* **8**: 302–307.)

cleavage and DNA strand transfer, rather than having two active sites, each specialized for one chemical reaction.

In contrast to the highly conserved structure of the catalytic domains, the remaining regions of proteins in this family are not conserved. These regions encode site-specific DNA-binding domains and regions involved in protein–protein interactions needed to assemble the protein–DNA complex specific for each individual element. Thus, these unique domains ensure that transposases and integrases catalyze recombination specifically only on the element that encoded them or on a very highly related element.

Transposases and integrases are only active when assembled into a synaptic complex, also called a transpososome, on DNA (see above). The cocrystal structure of Tn5 transposase bound to a pair of transposon-end DNA fragments provides insight into why this is the case (Fig. 12-25). The transposase subunit that is bound to the recombinase recognition sequences on one of these DNA fragments (i.e., on one transposon end) donates the catalytic domain that promotes the DNA cleavage and DNA strand transfer reactions on the other end of the transposon. Because of this subunit organization, the transposase will be properly positioned for recombination only when two subunits and a pair of DNA ends are present together in the complex.

### Poly-A Retrotransposons Move by a “Reverse Splicing” Mechanism

The poly-A retrotransposons (e.g., human LINE elements) move using an RNA intermediate but use a mechanism different from that used by the virus-like elements. This mechanism is called **target-site-primed reverse transcription** (Fig. 12-26). The first step is transcription of the DNA of an integrated element by a cellular RNA polymerase (Fig. 12-26a). Although

**FIGURE 12-25 Cocrystal of Tn5 bound to substrate DNA.** The complex contains a dimer of transposase. The catalytic domains are colored as in Figure 12-24. (Green balls) Divalent metal ions bound in the protein's active site. Note that the subunit bound via its DNA-binding domain to one transposon end donates the catalytic domain for recombination on the other DNA end. (Light blue and pink) The DNA. (Davies D.R. et al. 2000. *Science* **289**: 77–85.) Image prepared with MolScript, BobScript, and Raster3D with additional modeling of the DNA by Leemor Joshua-Tor.



the promoter is embedded in the 5'-UTR, it can in this case direct RNA synthesis to begin at the first nucleotide of the element's sequence.

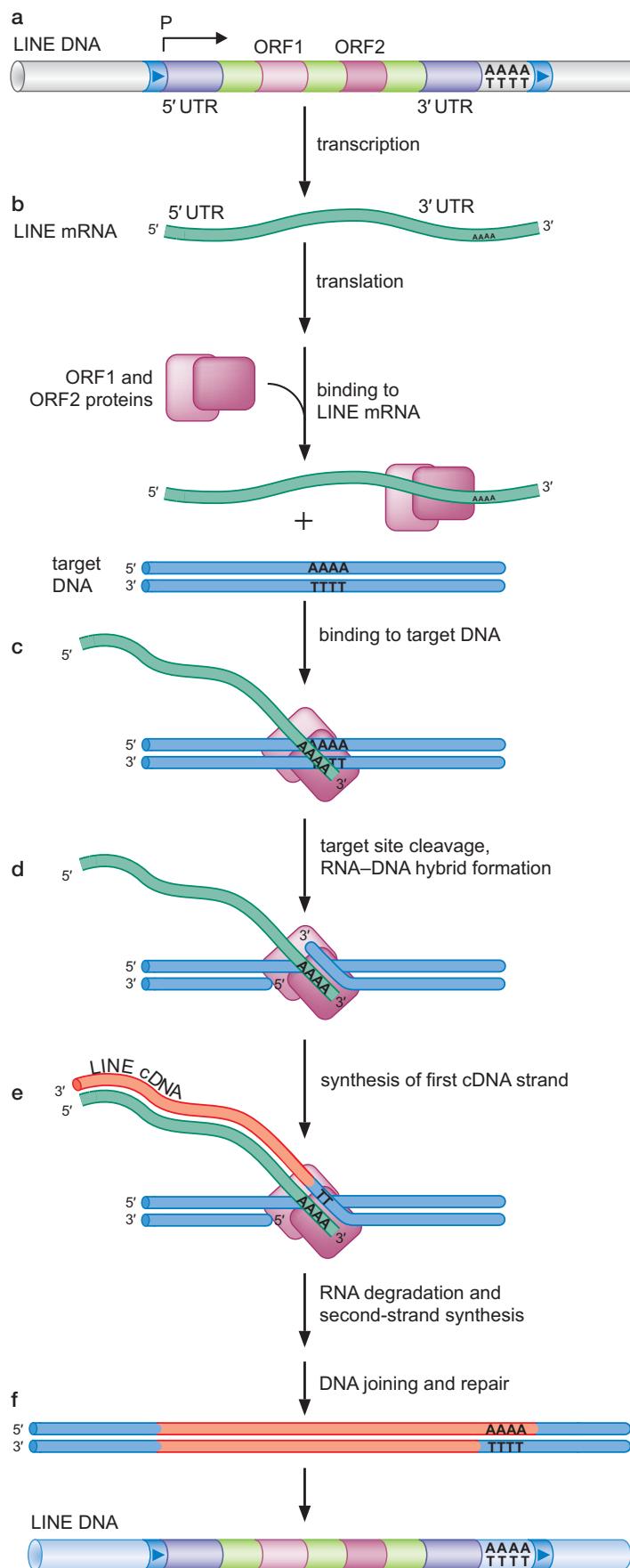
This newly synthesized RNA is exported to the cytoplasm and translated to generate the ORF1 and ORF2 proteins (see above). These proteins remain associated with the RNA that encoded them (Fig. 12-26b). In this way, an element promotes its own transposition and does not donate proteins to competing elements.

The protein–RNA complex then re-enters the nucleus and associates with the cellular DNA (Fig. 12-26c). Recall that the ORF2 protein has both a DNA endonuclease activity and a reverse transcriptase activity. The endonuclease initiates the integration reaction by introducing a nick in the chromosomal DNA (see Fig. 12-26d). T-rich sequences are preferred cleavage sites. The presence of these Ts at the cleavage site permits the DNA to base-pair with the poly-A tail sequence of the element RNA. The 3'-OH DNA end generated by the nicking reaction then serves as the primer for reverse transcription of the element RNA (Fig. 12-26e). The ORF2 protein also catalyzes this DNA synthesis. The remaining steps of transposition, although not yet well-understood, include synthesis of the second cDNA strand, repair of DNA gaps at the insertion site, and ligation to seal the DNA strands.

Many of the poly-A retrotransposons that have been detected by large-scale genomic sequencing are truncated elements. Most of these are missing regions from their 5' ends and do not have complete copies of element-encoded genes or an intact promoter. These truncated elements have therefore lost the ability to transpose.

## EXAMPLES OF TRANSPOSSABLE ELEMENTS AND THEIR REGULATION

Transposons have successfully invaded and colonized the genomes of all life-forms. Clearly, they are very robust biological entities. Some of this success can be attributed to the fact that transposition is regulated in ways that help to establish a harmonious coexistence with the host cell. This coexistence is essential for the survival of the element because transposons cannot exist without a host organism. On the other hand, as introduced above, transposons can wreak havoc in a cell, causing insertion mutations, altering gene expression, and promoting large-scale DNA rearrangements. These



**FIGURE 12-26** Transposition of a poly-A retrotransposon by target-site-primed reverse transcription. A model for the movement of a LINE element. (a) A cellular RNA polymerase initiates transcription of an integrated LINE sequence. (b) The resulting mRNA is translated to produce the products of the two encoded ORFs that then bind to the 3' end of their mRNA. (c) The protein–mRNA complex then binds to a T-rich site in the target DNA. (d) The proteins initiate cleavage in the target DNA, leaving a 3'-OH at the DNA end and forming an RNA:DNA hybrid. (e) The 3'-OH end of the target DNA serves as a primer for reverse transcription of the element RNA to produce cDNA (first-strand synthesis). (f) The final steps of the transposition reaction include second-strand synthesis and DNA joining and repair to create a newly inserted LINE element.

disruptions are particularly noticeable in plants, a feature that led to the discovery of transposons in maize (Box 12-3, Maize Elements and the Discovery of Transposons).

In the following sections, we briefly describe some of the best-understood individual transposons and transposon families. (A larger list of transposons and some of their important features is summarized in Table 12-2.) Each subsection provides a brief overview of a specific element and an example of regulation that is of particular importance to that element. As we shall see, two types of regulation appear as recurring themes.

- Transposons control the number of their copies present in a given cell. By **regulating copy number**, these elements limit their deleterious impact on the genome of the host cell.
- Transposons control target-site choice. Two general types of target-site regulation are observed. In the first, some elements preferentially insert into regions of the chromosome that tend not to be harmful to the host

#### ► KEY EXPERIMENTS

##### Box 12-3 Maize Elements and the Discovery of Transposons

Plant genomes are very rich in transposons. Furthermore, the ability of transposable elements to alter gene expression can often be readily observed as dramatic variation in the coloration of the plant (Box 12-3 Figs. 1 and 2). Thus, it is not surprising that transposable elements, and many of their salient features, were first discovered in plants.

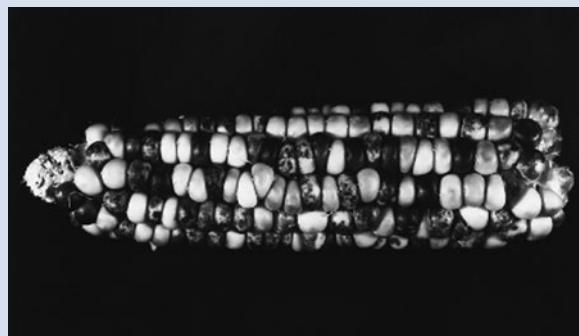
Barbara McClintock discovered “controlling elements” in maize in the late 1940s. It was actually the ability of transposable elements to break chromosomes that first came to McClintock’s attention. She found that some strains experienced broken chromosomes

very frequently, and she named the genetic element responsible for these chromosome breaks *Ds* (dissociator). Surprisingly, she observed that the sites of these “hot spots” for chromosome breaks were different in different strains and could even be in different chromosomal locations in the descendants of an individual plant. This observation provided the first insight that genetic elements could move (i.e., “transpose”) within chromosomes.

*Ds*, in fact, is a nonautonomous DNA transposon that moves by cut-and-paste transposition. *Ds* movement requires the *Ac* (activator) element (also discovered by McClintock) to be present in the same cell and provide the transposase protein. *Ac* is now recognized to be part of a large family of DNA transposons called the *hAT* family, named for the *hobo* elements from flies, the *Ac* elements from maize, and the *Tam* elements from snapdragon. The *Hermes* element from housefly is also a member of this family and has proved amenable to mechanistic analysis.



**BOX 12-3 FIGURE 1** Examples of color variegation in snapdragon flowers due to *Tam3* transposition. The size of white patches is related to the frequency of transposition. (Reprinted, with permission, from Chatterjee M. and Martin C. 1997. *Plant J.* **11**: 759–771, Fig. 2a. © Blackwell Publishing.)



**BOX 12-3 FIGURE 2** Example of corn (maize) cob showing color variegation due to transposition. (Photograph taken by Barbara McClintock; image courtesy of Cold Spring Harbor Laboratory Archives.)

**TABLE 12-2** Major Types of Transposable Elements

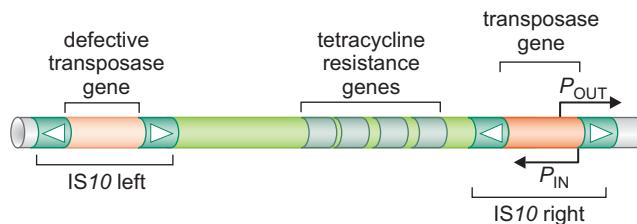
Type	Structural Features	Mechanism of Movement	Examples
<b>DNA-mediated transposition</b>			
Bacterial replicative transposons	Terminal inverted repeats that flank antibiotic-resistance and transposase genes	Copying of element DNA accompanying each round of insertion into a new target site	Tn3, $\gamma\delta$ , phage Mu
Bacterial cut-and-paste transposons	Terminal inverted repeats that flank antibiotic-resistance and transposase genes	Excision of DNA from old target site and insertion into new site	Tn5, Tn10, Tn7, IS911, Tn917
Eukaryotic transposons	Inverted repeats that flank coding region with introns	Excision of DNA from old target site and insertion into new site	P-elements ( <i>Drosophila</i> ), hAT family elements, Tc1/ <i>Mariner</i> elements
<b>RNA-mediated transposition</b>			
Virus-like retrotransposons	$\sim$ 250- to 600-bp direct terminal repeats (LTRs) flanking genes for reverse transcriptase, integrase, and retrovirus-like Gag protein	Transcription into RNA from promoter in left LTR by RNA polymerase II followed by reverse transcription and insertion at target site	Ty elements (yeast), Copia elements ( <i>Drosophila</i> )
Poly-A retrotransposons	3'-A-T-rich sequence and 5'-UTR flank genes encoding an RNA-binding protein and reverse transcriptase	Transcription into RNA from internal promoter; target-primed reverse transcription initiated by endonuclease cleavage	F and G elements ( <i>Drosophila</i> ), LINE and SINE elements (mammals), Alu sequences (humans)

cell. These regions are called **safe havens** for transposons. In the second type of regulation, some transposons specifically avoid transposing into their own DNA. This phenomenon is called **transposition target immunity**.

### IS4 Family Transposons Are Compact Elements with Multiple Mechanisms for Copy Number Control

The bacterial transposon Tn10 is a well-characterized representative of the IS4 family, which also includes Tn5. Tn10 is a compact element of 9 kb and encodes a gene for its own transposase and genes imparting resistance to the antibiotic tetracycline (Fig. 12-27).

Tn10 transposes via the cut-and-paste mechanism (described above), using the DNA hairpin strategy to cleave the nontransferred strands (Figs. 12-18 and 12-21). Tn10 is organized into three functional modules. This organization is relatively common, and elements that have it are called **composite transposons**. The two outermost modules, called IS10L (left) and



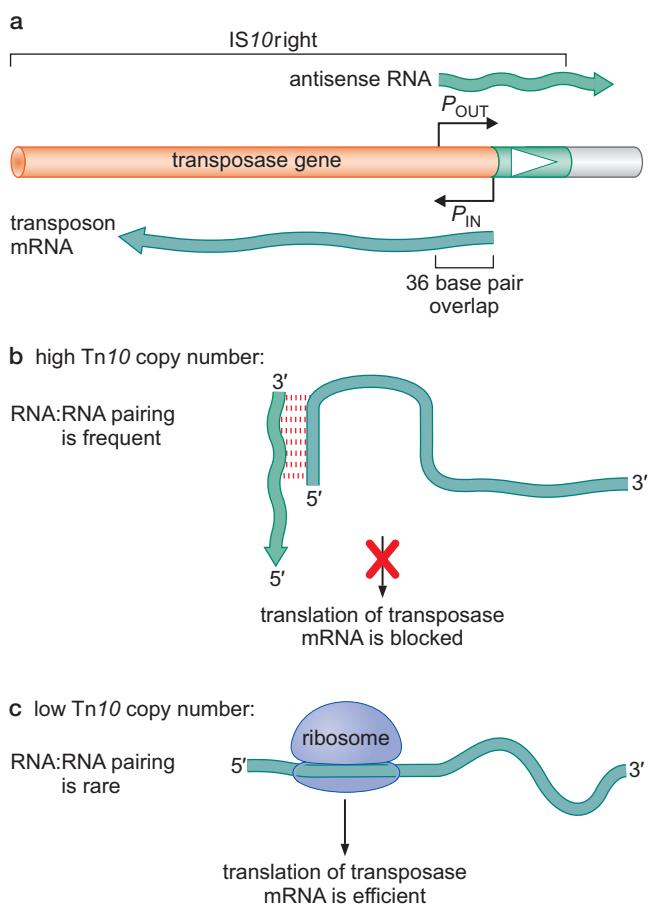
**FIGURE 12-27** Genetic organization of bacterial transposon Tn10. The map shows the functional elements in the bacterial transposon Tn10. Tn10, like many bacterial transposons, actually carries two “minitransposons” at its termini. For Tn10, these elements are called IS10L (left) and IS10R (right). Both types of IS10 elements can transpose and are found in DNA separately from Tn10. The white triangles show the inverted repeat sequences at the ends of the IS elements and Tn10. Although these four copies are not exactly the same in sequence, all are recognized by the Tn10 transposase and are used as recombination sites.

*IS10R* (right), are actually minitransposons. (“IS” stands for **insertion sequence**.) *IS10R* encodes the gene for the transposase that recognizes the terminal inverted repeat sequences of *IS10R*, *IS10L*, and *Tn10*. *IS10L*, although very similar in sequence to *IS10R*, does not encode a functional transposase. Thus, both *IS10R* and *Tn10* are autonomous, whereas *IS10L* is a nonautonomous transposon. Both types of *IS10* elements are found, as expected considering their own mobility, unassociated with *Tn10* in genomes.

*Tn10* limits its copy number in any given cell by strategies that restrict its transposition frequency. For example, one mechanism is the use of an **antisense RNA** to control the expression of the transposase gene (see Fig. 12-29) (for a discussion of antisense RNA regulation, see Chapters 19 and 20). Near the end of *IS10R* are two promoters that direct the synthesis of RNA by the host cell’s RNA polymerase. The promoter that directs RNA synthesis inward (called  $P_{IN}$ ) is responsible for the expression of the transposase gene. The promoter that directs transcription outward ( $P_{OUT}$ ), in contrast, serves to regulate transposase expression by making an antisense RNA.

By this mechanism, cells that carry more copies of *Tn10* will transcribe more of the antisense RNA, which, in turn, will limit expression of the transposase gene (Fig. 12-28; see legend for more details). The transposition frequency will therefore be very low in such a strain. In contrast, if there is only one copy of *Tn10* in the cell, the level of antisense RNA will be low, synthesis of the transposable protein will be efficient, and transposition will occur at a higher frequency.

**FIGURE 12-28** Antisense regulation of *Tn10* expression. (a) A map of the overlapping promoter regions is shown. The leftward promoter ( $P_{IN}$ ) promotes expression of the transposase gene; the rightward promoter ( $P_{OUT}$ ), which lies 36 bases to the left of  $P_{IN}$ , promotes expression of an antisense RNA. The first 36 bases of each transcript are complementary to one another. Note that in cells, the antisense transcript initiated at  $P_{OUT}$  is longer-lived than is the mRNA initiated at  $P_{IN}$ . (b) In cells having a high copy number of *Tn10*, the RNA:RNA pairing occurs frequently and blocks translation of the transposase mRNA (thereby eventually reducing the copy number of the element). (c) In cells having a low copy number of the transposon, RNA:RNA pairing is rare; the translation of transposase mRNA is efficient, and the copy number in the cell is increased.



### Phage Mu Is an Extremely Robust Transposon

Phage Mu, like bacteriophage  $\lambda$ , is a lysogenic bacteriophage (see Appendix 1). Mu is also a large DNA transposon. This phage uses transposition to insert its DNA into the genome of the host cell during infection and in this way is similar to the retroviruses (discussed above). Mu also uses multiple rounds of replicative transposition to amplify its DNA during lytic growth. During the lytic cycle, Mu completes approximately 100 rounds of transposition per hour, making it the most efficient transposon known. Furthermore, even when present as a quiescent lysogen, the Mu genome transposes quite frequently, compared with traditional transposons such as Tn10. The name *Mu* is short for mutator and stems from this ability to transpose promiscuously: cells carrying an inserted copy of the Mu DNA frequently accumulate new mutations due to insertion of the phage DNA into cellular genes.

The Mu genome is  $\sim 40$  kb and carries more than 35 genes, but only two encode proteins with dedicated roles in transposition. These are the *A* and *B* genes, which encode the proteins MuA and MuB. MuA is the transposase and is a member of the DDE protein superfamily discussed above. MuB is an ATPase that stimulates MuA activity and controls the choice of the DNA target site (Fig. 12-29). This process is explained in the next section.

### Mu Uses Target Immunity to Avoid Transposing into Its Own DNA

Mu, like many transposons, shows very little sequence preference at its target sites. As a result, “good” target sites occur very frequently in DNA including the DNA of the Mu genome itself. Given this nearly random sequence preference, how does Mu avoid transposing into its own DNA, a situation that would likely result in serious disruptions of the phage’s genes?

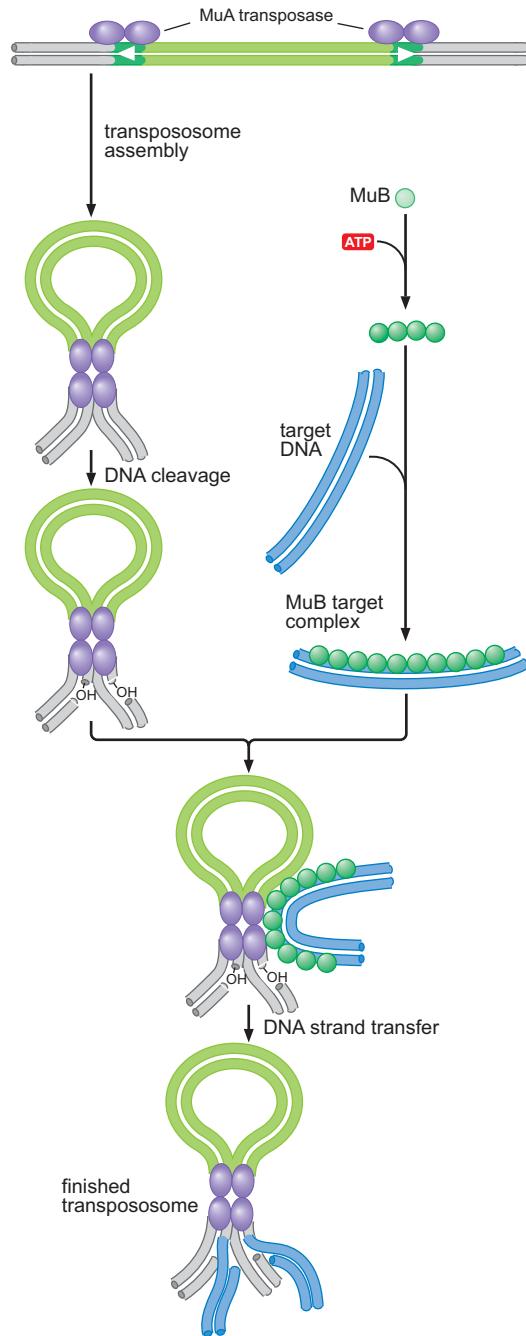
This problem is solved because Mu transposition is regulated by a process called **transposition target immunity** (see Box 12-4, Mechanism of Transposition Target Immunity). DNA sites surrounding a copy of the Mu element, including the element’s own DNA, are rendered very poor targets for a new transposition event.

Transposition target immunity is observed for several different transposable elements and can work over very long distances. For Mu, sequences within  $\sim 15$  kb of an existing Mu insertion are immune to new insertions. For some elements—for example, Tn3 and Tn7—target immunity occurs over distances  $>100$  kb. Target immunity protects an element from transposing into itself or from having another new copy of the same type of element insert into its genome. Furthermore, this type of regulation of target DNA selection also provides a driving force for elements to move to new locations “far” from where they are initially inserted, a feature that may also be advantageous for their overall propagation and survival.

### Tc1/mariner Elements Are Highly Successful DNA Elements in Eukaryotes

Recognizable members of the Tc1/mariner family of elements are widespread in both invertebrate and vertebrate organisms. Elements in this family are the most common DNA transposons present in eukaryotes. Although these elements are clearly related, members isolated from different organisms have distinguishing features and are named differently. For example, elements from the worm *Caenorhabditis elegans* are called Tc elements, whereas the original element named *Mariner* was isolated from a *Drosophila* species.

**FIGURE 12-29** Overview of the early steps of Mu transposition. Four subunits of the MuA transposase assemble on the ends of Mu DNA. MuB binds ATP and then binds to DNA of any sequence. A protein–protein interaction between MuA and MuB brings the MuA DNA–transpososome complex to a new DNA target site. MuB is not shown in the final panel because, after DNA strand transfer, it is no longer needed and probably leaves the complex.



*Tc1/mariner* elements are among the simplest autonomous transposons known. Typically, they are 1.5–2.5 kb long and carry only a pair of terminal inverted repeat sequences (the site of transposase binding) and a gene encoding a transposase protein of the DDE transposase superfamily (see above). In contrast to many transposons, no accessory proteins are required for transposition, although the final steps of recombination do require cellular DNA-repair proteins. This simplicity in structure and mechanism may be responsible for the huge success of these elements in such a wide range of host organisms.

*Tc1/mariner* elements move by a cut-and-paste transposition mechanism (Fig. 12-19). The transposon DNA is cleaved out of the old flanking host DNA using pairs of cleavages that are staggered by two base pairs. These

## ► ADVANCED CONCEPTS

**Box 12-4 Mechanism of Transposition Target Immunity**

Interplay between the MuA transposase and the MuB ATPase is at the center of the mechanism of transposition target immunity. MuA–MuB interactions prevent MuB from binding to the DNA near where MuA is bound. The interactions responsible for this interplay are listed below.

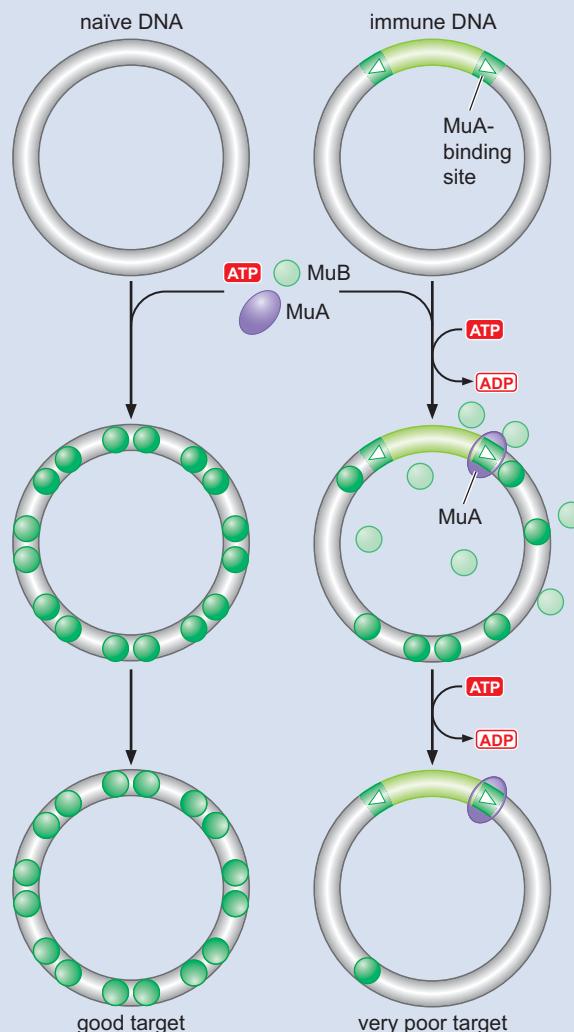
- MuA inhibits MuB from binding to nearby DNA sites. This inhibition requires ATP hydrolysis.
- MuB helps MuA find a target site for transposition.

To see how individual protein–protein and protein–DNA interactions function together to generate target immunity, consider transposition into two candidate DNA segments: one is any representative segment of DNA, whereas the second has a copy of Mu already inserted (see Box 12-4 Fig. 1). We call the first DNA segment the naive region and the second DNA segment the immune region.

What happens at each of these DNA regions as Mu prepares to transpose? First, we consider events at the naive region. MuB, in complex with ATP (MuB–ATP), will bind the DNA, using its non-specific DNA-binding activity. At the same time, MuA transposase will assemble a transpososome on the Mu DNA. This MuA in the transpososome can then make protein–protein contacts with the MuB–DNA complex at the naive region. As a result of this interaction, MuB delivers this DNA to MuA for use as a target site.

In contrast, both MuA and MuB bind to DNA in the immune region. MuA interacts with its specific binding sites on the Mu genome that is already present; MuB–ATP again binds using its affinity for any DNA sequence. However, when both MuA and MuB are bound to this region, they will interact. As a

result, MuA stimulates ATP hydrolysis by MuB and the disassociation of MuB from this DNA. MuB therefore does not accumulate on this immune DNA segment. By this means, the Mu transposition proteins use the energy stored in ATP to protect the Mu genome from becoming the target of transposition. As expected from this mechanism, even a single MuA-binding site within a DNA molecule is sufficient to impart target immunity.



**BOX 12-4 FIGURE 1** The interplay between MuA and MuB on DNA leads to the development of an immune target DNA. The MuA-binding sites are in the terminal inverted repeats on the ends of the transposon (dark green). MuA is shown bound to only one of the two repeat regions for clarity. Every time MuB hydrolyzes ATP, it dissociates from the DNA (MuB bound to ATP is darker green); MuA–MuB contact stimulates this hydrolysis reaction. Although shown contacting only two molecules of MuB, MuA will preferentially contact all of the MuB bound within close proximity to its DNA-binding site. DNA lengths of 5–15 kb can be rendered “immune” by a single MuA-bound terminal inverted repeat sequence.

elements strongly prefer to insert into DNA sites with the (obviously, very common) sequence 5'TA.

What happens to the “empty” site in the host chromosome when a transposon excises? In the case of *Tc1/mariner* elements, DNA sequence analysis of some sites that once carried a transposon reveals that sometimes the broken DNA ends are filled in (by repair DNA synthesis) and then directly joined (see the discussion on nonhomologous end joining in Chapter 10). These repair reactions result in the incorporation of a few extra base pairs

of DNA at the old insertion site. These small DNA insertions are known as “footprints,” because they are the traces left by a transposon that has “traveled through” a site in the genome.

In contrast to many transposons, the transposition of *Tc1/mariner* elements is not well-regulated. Perhaps as a result of this lack of control, many elements found by genome sequencing are “dead” (i.e., unable to transpose). For example, many elements carry mutations in the transposase gene that inactivate it. Using a large number of sequences from both inactive and active elements, researchers constructed an artificial hyperactive *Tc1/mariner* element. This element, named *Sleeping Beauty*, transposes at very high frequencies compared with naturally isolated elements. *Sleeping Beauty* is promising as a tool for mutagenesis and DNA insertion in many eukaryotic organisms. Furthermore, this reconstruction experiment reveals that the frequency of transposition by *Tc1/mariner* elements is naturally kept at bay because of the suboptimal activity of their transposase proteins.

### Yeast Ty Elements Transpose into Safe Havens in the Genome

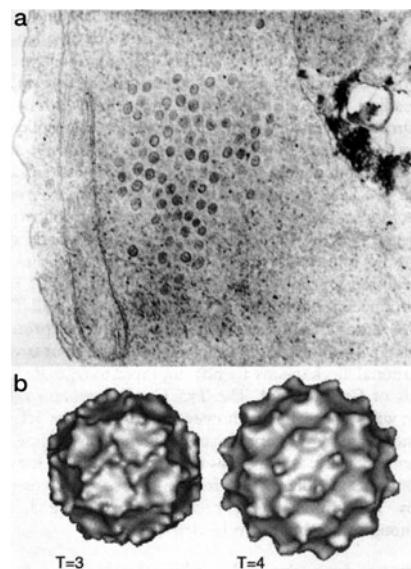
The Ty elements (transposons in yeast), prominent transposons in yeast, are virus-like retrotransposons. In fact, their similarity to retroviruses extends beyond their mechanism of transposition: Ty RNA is found in cells packaged into virus-like particles (Fig. 12-30). Thus, these elements seem to be viruses that cannot escape one cell and infect new cells. There are many types of well-studied Ty elements; for example, *S. cerevisiae* carries members of the Ty1, Ty3, Ty4, and Ty5 classes (although the Ty5 elements in this yeast species all appear to be inactive). Each of these classes of Ty elements promotes its own mobility but does not mobilize elements of another class.

Ty elements preferentially integrate into specific chromosomal regions (Fig. 12-31). For example, Ty1 elements nearly always transpose into DNA within  $\sim 200$  bp upstream of a start site for transcription by the host RNA polymerase III (Pol III) enzyme (see Chapter 13). RNA Pol III specifically transcribes tRNA genes, and most Ty1 insertions are near these genes. Ty3 integration is also tightly linked to Pol III promoters. In this case, integration is precisely targeted to the start site of transcription ( $\pm 2$  bp). In contrast, Ty5 preferentially integrates into regions of the genome that are in a silenced, transcriptionally quiescent state. Silenced regions targeted by Ty5 include the telomeres and the silent copies of the mating-type loci (see Chapter 11). In all of these cases, the mechanism of regional target-site selection involves the formation of specific protein–protein complexes between the element’s integrase—bound in a complex to the cDNA—and host-specific proteins bound to these chromosomal sites. For example, Ty5 integrase forms a specific complex with the DNA-silencing protein Sir4 (see Chapter 19).

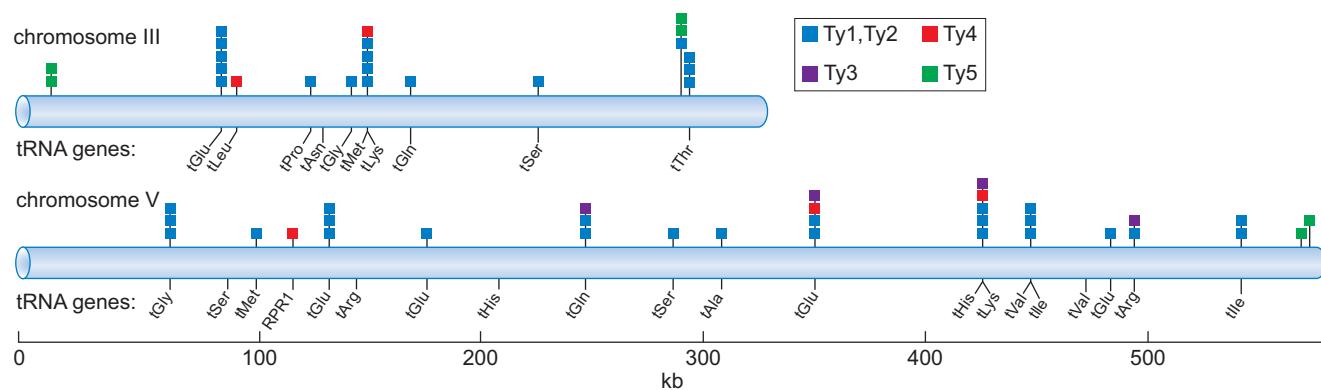
Why do Ty elements exhibit this regional target-site preference? It is proposed that this target specificity enables the transposons to persist in a host organism by focusing most of their insertions away from important regions of the genome that are involved directly in coding for proteins. The use of this type of targeted transposition may be especially important in organisms with small gene-rich genomes, such as yeast.

### LINEs Promote Their Own Transposition and Even Transpose Cellular RNAs

The autonomous poly-A retrotransposons known as LINEs are abundant in the genomes of vertebrate organisms. In fact,  $\sim 20\%$  of the human genome is



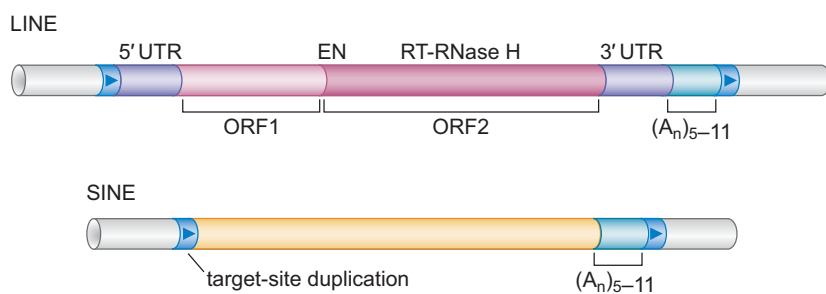
**FIGURE 12-30** Yeast Ty elements packaged into viral particles. (a) An electron micrograph of *Saccharomyces cerevisiae* cells overexpressing Ty1 virus-like particles. The particles are seen as oval electron-dense structures. (b) Cryoelectron microscopy showing the 3D reconstructions of Ty1 virions. These Ty1 elements carry a truncated Gag protein that forms the spiky shells with trimeric units of the particles. (Reprinted, with permission, from Craig N. et al. 2002. *Mobile DNA II*, © ASM Press. b, Also courtesy of H. Saibil.)



**FIGURE 12-31** Clustered integration sites observed for Ty elements. Each colored box represents a known site for transposon insertion. Note that Ty1, Ty2, Ty3, and Ty4 insertions are near tRNA genes, which are transcribed by the cellular RNA polymerase III. Insertion occurs upstream of the actual gene and therefore does not disrupt expression. Ty1 and Ty2 are closely related elements and therefore are grouped together. Ty5 is found near the ends of chromosomes and near the mating-type loci (see Chapter 11) that are “silenced” (i.e., not highly transcribed). (Courtesy of Dan Voytas.)

composed of LINE sequences. These elements were first recognized as a family of repeat sequences. Their name is derived from this initial identification: **LINE** is the acronym for “long interspersed nuclear element.” L1 is one of the best understood LINEs in humans. In addition to promoting their own mobility, LINEs also donate the proteins needed to reverse-transcribe and integrate another related class of repeat sequences, the nonautonomous poly-A retrotransposons, known as short interspersed nuclear elements (**SINEs**). Genome sequences reveal, once again, the presence of huge numbers of these elements, which are typically only between 100 and 400 bp in length. The *Alu* sequence is an example of a widespread SINE in the human genome. A comparison of the structures of typical LINE and SINE elements is shown in Figure 12-32. The sequences of LINEs and SINEs look like simple genes. In fact, the *cis*-acting sequences important for transposition simply include a promoter, to direct transcription of the element into RNA, and a poly-A sequence. Recall that these A residues pair with the DNA at the target site to help generate the primer terminus for reverse transcription (see Fig. 12-23).

These simple sequence requirements for transposition pose a problem for LINEs: how do they avoid transposing cellular mRNA molecules? All genes have a promoter, and most are transcribed into an mRNA that will carry a poly-A sequence at the 3' end of the molecule (Chapter 13). Thus, any mRNA should be an attractive “substrate” for transposition. In fact, genome sequences provide clear evidence for transposition of cellular RNA via the target-primed reverse transcription mechanism.



**FIGURE 12-32** Genetic organization of a typical LINE and SINE. Note the variable-length poly-A sequence at the right end of the elements. This is a defining feature of the poly-A retrotransposons. These elements are also flanked by target-site duplications that are variable in length (blue arrows). Sequence elements are not shown to scale. Both types of elements also carry promoter sequences. See Figures 12-20 and 12-28. (Adapted, with permission, from Bushman F. 2002. *Lateral DNA transfer*, p. 251, Fig. 8.4. © Cold Spring Harbor Laboratory Press.)

For many cellular genes, there are additional copies of a highly related sequence in the genome. These copies appear to have lost their promoter and their introns (regions of sequence present within a gene but removed from the mRNA by RNA splicing) (see Chapter 14) and often carry truncations near their 5' ends. These sequences are known as **processed pseudogenes** and usually are not expressed by the cell. These pseudogenes are often flanked by short repeats in the target DNA. This structure is exactly that expected of LINE-promoted transposition of a cellular mRNA.

Although transposition of cellular RNAs can occur, it is a rare event. The principal mechanism used to avoid this process is that the LINE-encoded proteins bind immediately to their own RNA during translation (see Fig. 12-23). Thus, they show a strong bias to catalyzing reverse transcription and integration of the RNA that encoded them.

## V(D)J RECOMBINATION

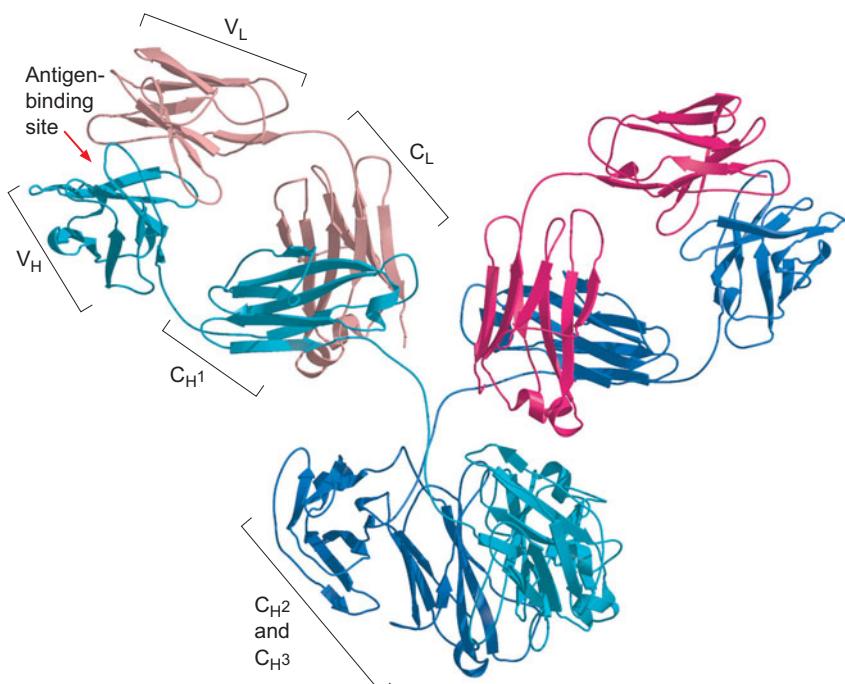
---

We have seen that transposition is involved in the movement of many different genetic elements. Cells, however, have also harnessed this recombination mechanism for functions that directly help the organism. The best example is V(D)J recombination, which occurs in the cells of the vertebrate immune system.

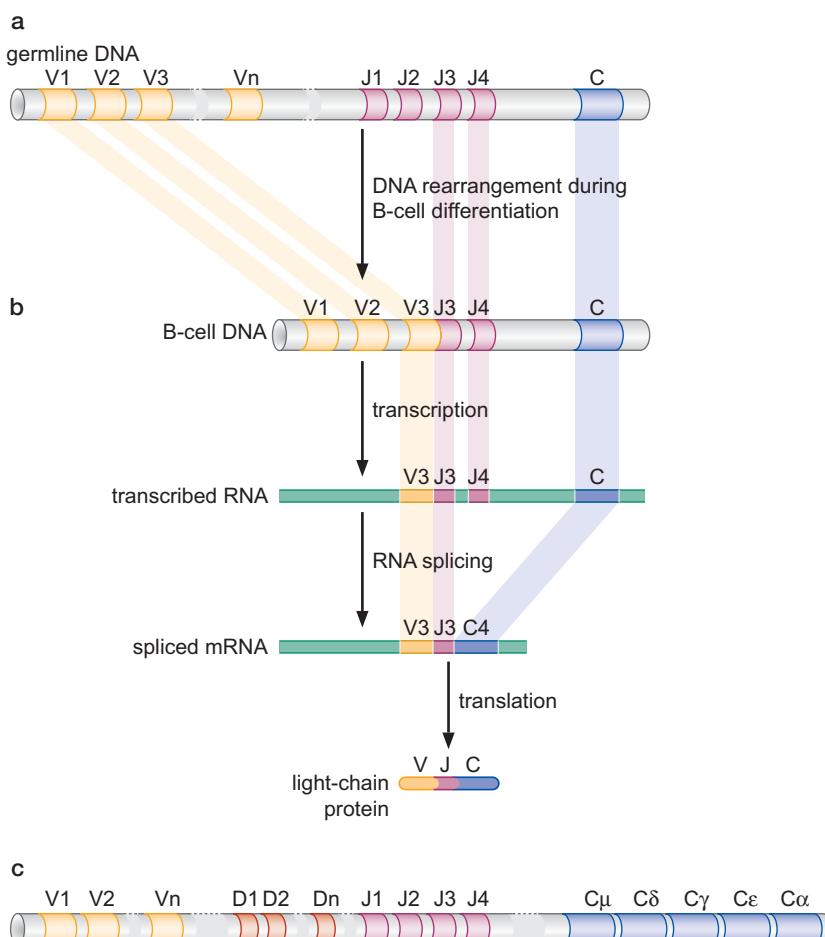
The immune system of vertebrates has the job of recognizing and fending off invading organisms, including viruses, bacteria, and pathogenic eukaryotes. Vertebrates have two specialized cell types dedicated to recognizing these invaders: B-cells and T-cells. B-cells produce **antibodies** that circulate in the bloodstream, whereas T-cells produce cell surface-bound receptor proteins (called **T-cell receptors**). Recognition of a “foreign” molecule by either of these classes of proteins starts a cascade of events focused on destruction of the invader. To fulfill their functions successfully, antibodies and T-cell receptors must be able to recognize an enormously diverse group of molecules. The principal mechanism cells use to generate antibodies and T-cell receptors with such diversity relies on a specialized set of DNA rearrangement reactions known as **V(D)J recombination**.

Antibody and T-cell receptor genes are composed of gene segments that are assembled by a series of sequence-specific DNA rearrangements. To understand how this recombination generates the needed diversity, we need to look at the structure of an antibody molecule (Fig. 12-33); T-cell receptors have a similar modular structure. A genomic region encoding an antibody molecule is shown in Figure 12-34. Antibodies are constructed of two copies each of a light chain and a heavy chain. The part of the protein that interacts with foreign molecules is called the **antigen-binding site**. This binding region is constructed from the  $V_L$  and  $V_H$  domains of the antibody molecule, shown in Figure 12-33. The “V” signifies that the protein sequence in this region is highly variable. The remaining domains of the antibody are called “C,” or constant, regions and do not differ among different antibody molecules.

Figure 12-34a shows the genomic region encoding an antibody light chain (from a mouse), called the  $\kappa$  locus. This region carries about 300 gene segments coding for different versions of the light-chain  $V_L$  protein region. There are also four gene segments encoding a short region of protein sequence called the J region, followed by a single coding region for the  $C_L$  domain. By the mechanism we describe later, V(D)J recombination can fuse the DNA between any pair of V and J segments. Thus, as a result of



**FIGURE 12-33** Structure of an antibody molecule. (Pink) The two light chains; (blue) the heavy chains. The variable and constant regions are labeled on the left side of the molecule only. Note that the antigen-binding region is formed at the interface between the  $V_L$  and  $V_H$  domains. (Harris L.J. et al. 1998. *J. Mol. Biol.* **275**: 861–872.) Image prepared with MolScript, BobScript, and Raster3D.



**FIGURE 12-34** Overview of the process of V(D)J recombination. (Top panels) The steps involved in producing the light chain of an antibody protein. (a) The genetic organization of part of the light-chain DNA in cells that have not experienced V(D)J recombination (germline DNA). (b) Recombination between two specific gene segments (V3 and J3) as occurs during B-cell development. This is only one of the many types of recombination events that can occur in different pre-B-cells. The recombined locus is then transcribed and the RNA spliced (Chapter 14) to juxtapose a constant-region gene segment. This mRNA is then translated to generate the light-chain protein. (c) Schematic of the even more complex heavy-chain genetic region, with its additional "D" gene segments and multiple types of constant-region segments ( $C_\mu$ ,  $C_\gamma$ , etc.). (Adapted, with permission, from Bushman F. 2002. *Lateral DNA transfer*, p. 345, Fig. 11.3. © Cold Spring Harbor Laboratory Press.)

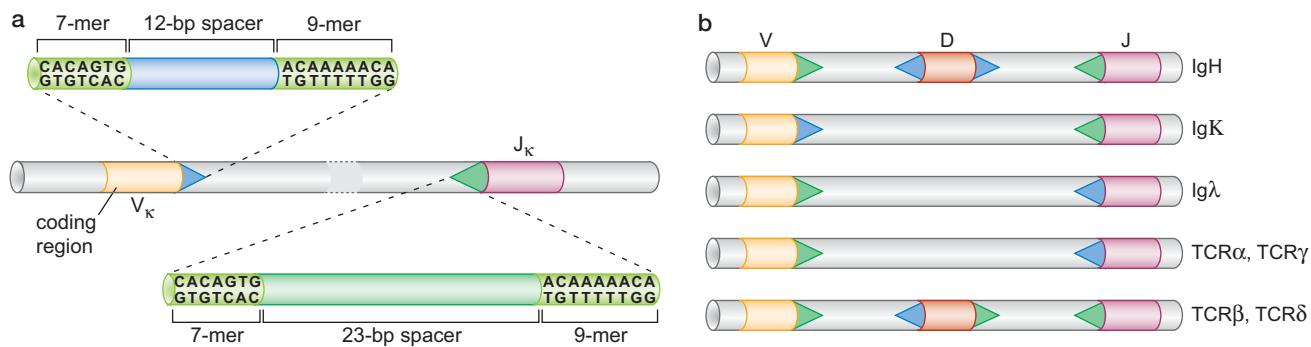
recombination, 1200 variants of the antibody light chain can be produced from this single genomic region. These segments are then brought together with the C<sub>L</sub>-coding region by RNA splicing (Chapter 14).

The situation for assembly of the gene segments encoding the antibody heavy chain is similar. In this case, however, there is an additional type of gene segment, called D (for “diversity”) (Fig. 12-34c). Heavy-chain genes can be very complex. For example, a specific heavy-chain locus in a mouse has more than 100 V regions, 12 D regions, and four J regions. V(D)J recombination can assemble this gene to generate more than 4800 different protein sequences. Because functional antibodies can be constructed from any pair of light and heavy chains, the diversity generated by recombination at the light and heavy loci has a multiplicative impact on protein structure.

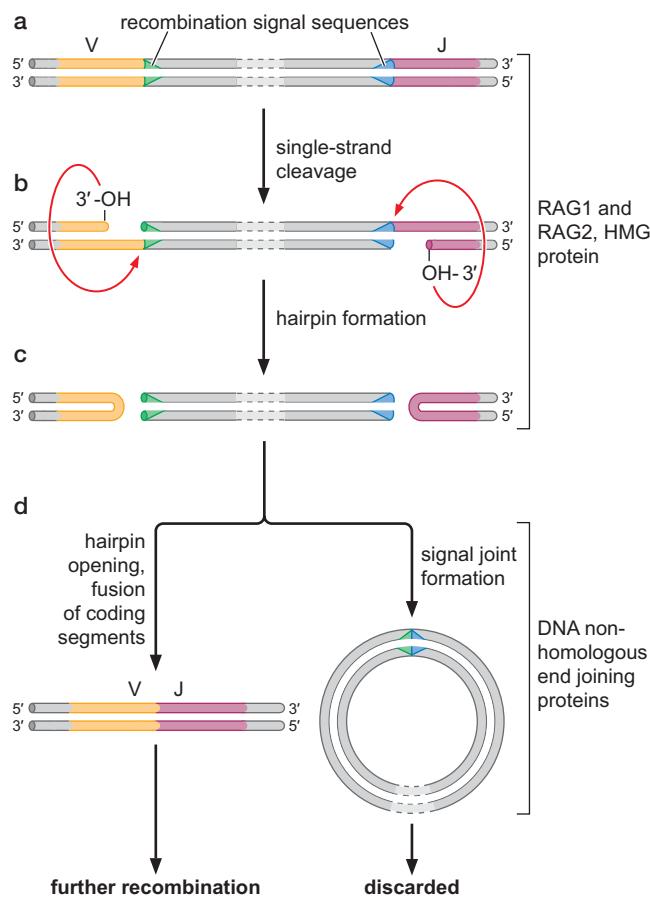
### The Early Events in V(D)J Recombination Occur by a Mechanism Similar to Transposon Excision

Recombination sequences, called **recombination signal sequences**, flank the gene segments that are assembled by V(D)J recombination. These signals all have two highly conserved sequence motifs, one 7 bp (the 7-mer) and the second 9 bp (the 9-mer) in length (Fig. 12-35). These motifs are bound by the recombinase (see later discussion). The recombination signal sequences come in two classes. One class has the 7-mer and 9-mer motifs spaced by 12 bp of sequence, whereas the second class has these motifs spaced by 23 bp (Fig. 12-35a). Recombination always occurs between a pair of recombination signal sequences in which one partner has the 12-bp “spacer” and the other partner has the 23-bp “spacer.” These pairs of recombination signal sequences are organized as inverted repeats flanking the DNA segments that are destined to be joined (Fig. 12-35b).

The recombinase responsible for recognizing and cleaving the recombination signal sequences is composed of two protein subunits called **RAG1** and **RAG2** (RAG for “recombination-activating gene”). These proteins function in a manner very similar to that of a transposase (Fig. 12-36). They recognize the recombination signal sequences and pair the two sites to form a protein–DNA synaptic complex.



**FIGURE 12-35** Recombination signal sequences recognized in V(D)J recombination. (a) Close-up of the two types of recombination signal sequences (RSSs). (Blue) The 12-bp spacer; (green) the 23-bp spacer; (light green) conserved 7-mer and 9-mer sequence elements, shared by both types of sequences. The nucleotide sequence in the spacer region is not important. The length, however, is critical. (b) Examples of RSS arrangements in the genetic regions encoding antibodies (Ig genes) and T-cell receptor proteins (TCR genes). (a, Adapted, with permission, from Bushman F. 2002. *Lateral DNA transfer*, p. 346, Fig. 11.5. © Cold Spring Harbor Laboratory Press.)



**FIGURE 12-36** The V(D)J recombination pathway: cleavages occur by a mechanism similar to transposon excision. The recombinases catalyze single-strand cleavage at the ends of the signal sequences, leaving a free 3'-OH. Each 3'-OH then initiates attack on the opposite strands to form a hairpin intermediate (see Fig. 12-23b). The hairpin structures are subsequently hydrolyzed and then joined together to form a coding joint between the V and J regions. The two ends carrying the recombination signal sequences are also joined to form a signal joint. The former structure undergoes further recombination, whereas the latter is discarded. (Adapted, with permission, from Bushman F. 2002. *Lateral DNA transfer*, p. 348, Fig. 11.6. © Cold Spring Harbor Laboratory Press.)

The RAG1 proteins within this complex then introduce single-strand breaks in the DNA at each of the junctions between the recombination signal sequence and the gene segment that will be rearranged (Fig. 12-36a). The site of cleavage is such that the protein-coding segment now has a free 3'-OH DNA end (Fig. 12-36b). Then, as we have seen previously for some transposon excision reactions (especially in the *Hermes* pathway) (see Fig. 12-21), this 3'-OH DNA end attacks the opposite strand of the DNA double helix. This attack results in the coupled DNA cleavage and joining reaction that generates a hairpin DNA end. It is the protein-coding sequence segments that have the DNA hairpin ends, whereas the recombination signal sequences now have normal double-strand breaks at their ends (Fig. 12-36c). This same mechanism generates a DNA hairpin at each of the two recombining DNA segments.

Once the two DNA sequences in the synaptic complex have been nicked and “hairpinned” by the RAG recombinase, cellular DNA-repair proteins take over to finish the recombination reaction (Fig. 12-36d). The DNA hairpin ends on the two protein-coding segments must be opened, and these ends must then be joined together. Cellular nonhomologous end-joining proteins (see Chapter 10) participate in this reaction. Interestingly, DNA joining is often accompanied by the addition (or deletion) of a few nucleotides. These additions are analogous to the “footprints” left in the old target DNA when transposons excise, as we described for the *Tc1/mariner* transposons. The added nucleotides contribute an extra component to the sequence diversity of the resulting protein molecule. The pair of cleaved recombination signal sequences is also joined together during recombination. This event generates a circular DNA molecule that is usually discarded by the cell.

The similarities between the mechanism of DNA cleavage to initiate V(D)J recombination and transposon excision are remarkable. In fact, the recombination signal sequences also look similar to the terminal inverted repeats found at the ends of a transposon, and the RAG1 protein has some sequence similarity to the DDE transposase protein family. In fact, genomic analysis has recently uncovered a transposon family called *Transib* that is the likely source of both RAG1 and the recombination signal sequences. These observations, together with many others, provide overwhelming evidence for the proposal that V(D)J recombination, now a critical feature of the immune system of higher animals, evolved from a DNA transposon. This conclusion speaks to the critical importance of transposable elements in the evolution of cellular genomes.

## SUMMARY

---

Although DNA is normally thought of as a very static molecule that archives the genetic material, it is also subject to numerous types of rearrangements. Two classes of genetic recombinations—conservative site-specific recombination and transposition—are responsible for many of these events.

Conservative site-specific recombination occurs at defined sequence elements in the DNA. Recombinase proteins recognize these sequence elements and act to cleave and join DNA strands to rearrange DNA segments containing the recombination sites. Three types of rearrangements are common: DNA insertion, DNA deletion, and DNA inversion. These rearrangements have many functions, including insertion of a viral genome into that of the host cell during infection, resolving DNA multimers, and altering gene expression.

The organization of the recombination sites on the DNA and the participation of DNA architectural proteins dictate the outcome of a specific recombination reaction. The architectural proteins function to bend DNA segments and can have a large influence on the reactions occurring on a specific region of DNA.

There are two families of conservative site-specific recombinases. Both families cleave DNA using a protein–DNA covalent intermediate. For the serine recombinases, this linkage is via an active-site serine residue; for the tyrosine recombinases, it is via a tyrosine. Structures of the tyrosine recombinases yield many insights into the details of the recombination mechanism.

Transposition is a class of recombination that moves mobile genetic elements, called transposons, to new genomic sites. There are three major classes of transposons: DNA transposons, virus-like retrotransposons, and poly-A retrotransposons. The DNA transposons exist as DNA throughout a cycle of transposition. They move either by a cut-and-paste recombination mechanism, which involves an excised transposon intermediate, or by a replicative mechanism. The two classes of retrotransposons move using an RNA intermediate. These “retro” elements require the RNA-dependent DNA polymerase, called reverse transcriptase, as well as a recombinase protein for mobility.

Transposons are present in the genomes of all organisms, where they can constitute a huge fraction of the total DNA sequence. They are a major cause of mutations and genome rearrangements. Transposition is often regulated to help ensure that transposons do not cause too much of a disruption to the genome of the host cell. Control of transposon copy number and regulation of the choice of new insertion sites are commonly observed.

Finally, a transposition-like mechanism can be used for other types of DNA rearrangement reactions. The prime example of this is the V(D)J recombination reaction, responsible for assembly of gene fragments during development of the vertebrate immune system.

## BIBLIOGRAPHY

---

### Books

- Bushman F. 2002. *Lateral DNA transfer: Mechanisms and consequences*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
- Craig N.L., Craigie R., Gellert M., and Lambowitz A.M. eds. 2002. *Mobile DNA II*. American Society for Microbiology, Washington, D.C.

### Site-Specific Recombination

- Chen Y. and Rice P.A. 2003. New insight into site-specific recombination from FLP recombinase–DNA structures. *Annu. Rev. Biophys. Biomol. Struct.* **32**: 135–159.
- Grindley N.D.F., Whiteson K.L., and Rice P.A. 2006. Mechanisms of site-specific recombination. *Annu. Rev. Biochem.* **75**: 567–605.

- Hallet B. and Sherratt D.J. 1997. Transposition and site-specific recombination: Adapting DNA cut-and-paste mechanisms to a variety of genetic rearrangements. *FEMS Microbiol. Rev.* **21**: 157–178.
- Matthews A.G. and Oettinger M.A. 2009. RAG: A recombinase diversified. *Nat. Immunol.* **10**: 817–821.
- Smith M.C. and Thorpe H.M. 2002. Diversity in the serine recombinases. *Mol. Microbiol.* **44**: 299–307.
- Yang W. 2010. Topoisomerases and site-specific recombinases: Similarities in structure and mechanism. *Crit. Rev. Biochem. Mol. Biol.* **45**: 520–534.
- Prak E.T.L. and Kazazian H.H. Jr. 2000. Mobile elements in the human genome. *Nat. Rev. Genet.* **1**: 134–144.
- Rebollo R., Romanish M.T., and Mager D.L. 2012. Transposable elements: An abundant and natural source of regulatory sequences for host genes. *Annu. Rev. Genet.* (in press). doi: 10.1146/annurev-genet-110711-155621.
- Rice P.A. and Baker T.A. 2001. Comparative architecture of transposase and integrative complexes. *Nat. Struct. Biol.* **8**: 302–307.
- Smit A.F.A. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9**: 657–663.

## Transposition

- Gueguen E., Rousseau P., Duval-Valentin G., and Chandler M. 2005. The transpososome: Control of transposition at the level of catalysis. *Trends Microbiol.* **13**: 543–549.
- Haren L., Ton-Hoang B., and Chandler M. 1999. Integrating DNA: Transposases and viral integrases. *Annu. Rev. Microbiol.* **53**: 245–281.
- Plasterck R. 1995. The *Tc1/mariner* transposon family. *Curr. Top. Microbiol. Immunol.* **204**: 125–143.

## V(D)J Recombination

- Fugmann S.D., Lee A.I., Schockett P.E., Villey I.J., and Schatz D.G. 2000. The RAG proteins and V(D)J recombination: Complexes, ends, and transposition. *Annu. Rev. Immunol.* **18**: 495–527.
- Gellert M. 2002. V(D)J recombination: RAG proteins, repair factors, and regulation. *Annu. Rev. Biochem.* **71**: 101–132.
- Oettinger M.A. 2004. Hairpins at split ends in DNA. *Nature* **432**: 960–961.

## QUESTIONS

## MasteringBiology®

For instructor-assigned tutorials and problems, go to MasteringBiology.

For answers to even-numbered questions, see Appendix 2: Answers.

**Question 1.** Given a linear piece of double-stranded DNA that includes two separate crossover regions surrounded by recombinase recognition sites, describe the alignment of the recombination sites that determines if the recombination outcome is a deletion or inversion. Explain why this arrangement dictates the outcome of the reaction.

**Question 2.** Explain why serine and tyrosine recombinases do not require an external source of energy such as ATP hydrolysis for catalysis.

**Question 3.** Explain a major difference between site-specific recombination and transposition.

**Question 4.** List the similarities and differences between the mechanisms of tyrosine and serine recombinases during conservative site-specific recombination.

**Question 5.** Describe advantages of using Cre recombinase for genetic engineering in eukaryotic cells.

**Question 6.** Infection of *E. coli* with bacteriophage  $\lambda$  involves integrative recombination for the phage to enter the lysogenic state and excisive recombination for it to enter lytic growth. Does  $\lambda$ Int have a role in integrative recombination, excisive recombination, or both? Explain your reasoning.

**Question 7.** Describe the biological relevance to the Hin recombinase catalyzing DNA inversion in the *Salmonella typhimurium* genome.

**Question 8.** Explain the major feature in the cycle of recombination that distinguishes DNA transposons from retrotransposons.

**Question 9.** Provide an explanation for how the human genome can contain greater than 50% transposon-related sequence but does not experience major genetic instability as a consequence of transposon movement.

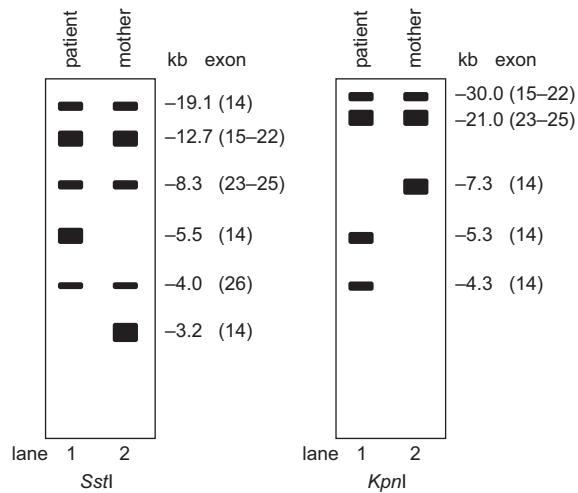
**Question 10.** Explain why scientists use transposons like Tn5 as a genetic engineering tool to screen a population of bacterial or yeast cells for mutants for a given phenotype. Why does the presence of the transposon in the mutant provide an experimental advantage in a genetic screen, compared to mutations generated by chemical mutagenesis?

**Question 11.** Researchers have found that treating human tumor cell lines with an inhibitor of reverse transcriptase reduces the rate of tumor-cell proliferation. Predict why reverse transcriptase is expressed in human cells. Hypothesize a general reason why reverse transcriptase activity could be associated with tumor cells.

**Question 12.** Compare and contrast the cut-and-paste mechanism of transposition with the replicative mechanism of transposition.

**Question 13.** Describe the role of V(D)J recombination in antibody diversification. Explain why nonhomologous end joining is an advantageous mechanism to repair the double-stranded DNA breaks (fusion of the coding segments after the hairpins are hydrolyzed) in the V(D)J recombination pathway.

**Question 14.** A deficiency in factor VIII causes hemophilia A, a blood disorder. Researchers studying hemophilia A evaluated the DNA from an affected patient and the patient's unaffected mother. They analyzed the 186-kb-long factor VIII gene that includes 26 exons (see Chapter 14). After digesting the genomic DNA with KpnI or SstI and separating the products by gel electrophoresis, the researchers probed the DNA with a radiolabeled cDNA probe that binds the factor VIII gene in a region that includes exons 14–26. The size of the fragments and the corresponding exon(s) are shown on the right of autoradiograms. The researchers conclude a transposon inserted into one of the exons of the factor VIII gene.

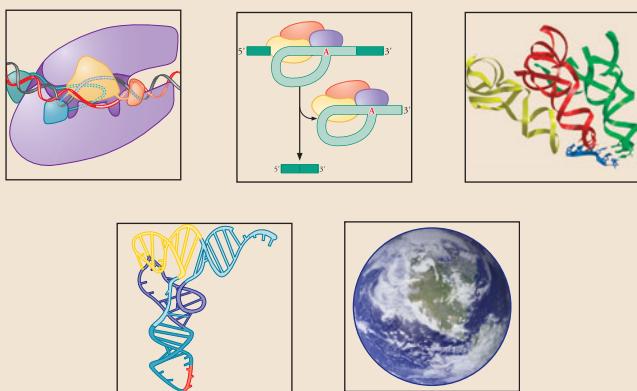


- Describe the differences between the patient and mother *SstI* digest results. Be specific.
- Describe the differences between the patient and mother *KpnI* digest. Be specific.
- Propose an hypothesis explaining the observed differences including what exon of the factor VIII gene contains the transposon.

Data adapted from Kazazian Jr. et al. (1988. *Nature* **332**: 164–166).

P A R T      4

# EXPRESSION OF THE GENOME



## O U T L I N E

- CHAPTER 13  
Mechanisms of Transcription, 429
- 
- CHAPTER 14  
RNA Splicing, 467
- 
- CHAPTER 15  
Translation, 509
- 
- CHAPTER 16  
The Genetic Code, 573
- 
- CHAPTER 17  
The Origin and Early Evolution  
of Life, 593

PART 4 IS PRIMARILY CONCERNED WITH HOW information in the form of the linear sequence of nucleotides in a polynucleotide chain (DNA) is converted into the linear sequence of amino acids in a polypeptide chain (protein). We also consider how these processes came about and evolved from simpler beginnings.

Chapters 13–16 trace the flow of information from the copying of the gene into an RNA replica known as the messenger RNA to the decoding of the messenger RNA into a polypeptide chain. The process by which nucleotide sequence information is transferred from DNA to RNA is known as transcription, and this is the subject of Chapter 13.

The enzyme RNA polymerase unwinds a short stretch of DNA locally and uses one of the two transiently separated DNA strands as a template upon which it progressively builds a complementary RNA copy by base pairing in a chemical reaction very similar to DNA synthesis. Although the basic enzyme that makes the RNA is very similar in all cells, the rest of the machinery involved in transcription in eukaryotes is more complex than its prokaryotic counterparts.

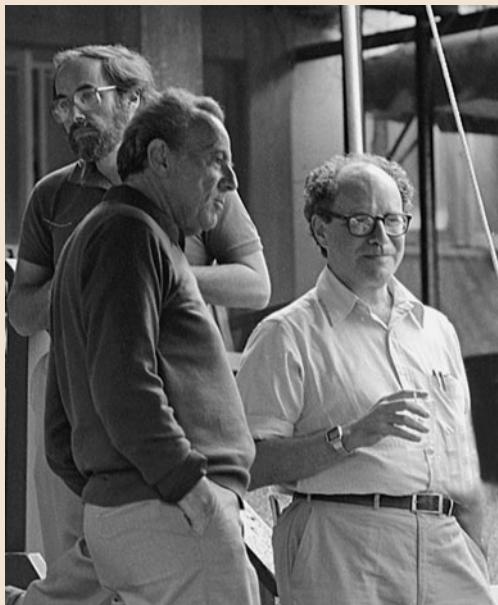
In prokaryotes, once the messenger RNA is synthesized, it is ready for the next stage of information flow in which RNA is used as a template for protein synthesis. But not in eukaryotes: there the RNA product of transcription must undergo a series of maturation events before it is competent to serve as a messenger RNA. The most dramatic processing event is called mRNA splicing; it is described in Chapter 14. Genes in eukaryotic cells are frequently interrupted by non-protein-coding segments known as introns. The number of introns found within the coding sequence varies for each gene and can range from one to several. When the gene is transcribed into an RNA copy, these introns must be removed so that the protein-coding segments, known as exons, can be joined to each other to create a contiguous protein-coding sequence. Chapter 14 describes the elaborate molecular machine responsible for removing introns with great precision.

The details of the intricate process known as translation are discussed in Chapters 15 and 16. This is the process whereby genetic information, in the form of the sequence of nucleotides in messenger RNA, is used to direct the ordered incorporation of amino acids into the polypeptide chain of a protein. Chapter 15 describes the principal participants in translation: the coding sequence in messenger RNA; adaptor molecules known as tRNAs; enzymes that load amino acids onto the tRNA adaptors; and the protein-synthesizing factory itself, the ribosome, which is composed of RNA and protein.

Chapter 16 describes the classic experiments that led to the elucidation of the genetic code and lays out the rules by which the code is translated. The nucleotide sequence information is based on a three-letter code, whereas the protein sequence information is based on 20 different amino acids. The code is degenerate with two or more codons (in most cases) specifying the same amino acid. There are also specific codons that indicate where translation should start and where it should stop.

Finally, in Chapter 17, we consider how life arose in the first place, and how the crude mechanisms for coding, replicating, and expressing information evolved into the elaborate systems we see today, as described in Parts 3 and 4.

## PHOTOS FROM THE COLD SPRING HARBOR LABORATORY ARCHIVES



**David Baltimore, François Jacob, and Walter Gilbert, 1985 Symposium on the Molecular Biology of Development.** Baltimore codiscovered, with Howard Temin, the enzyme reverse transcriptase, which makes DNA using RNA as a template (Chapter 12). Jacob, with Jacques Monod, proposed the basic model for how gene expression is regulated (Chapter 18) and also proposed a model for how DNA replication is regulated (Chapter 9). Gilbert provided biochemical validation for aspects of the Jacob and Monod model of gene regulation; he also invented a chemical method for sequencing DNA (Chapter 7). They all separately shared in Nobel Prizes, in 1975 (in Physiology or Medicine), 1965 (in Physiology or Medicine), and 1980 (in Chemistry), respectively.



**Ada Yonath, 2001 Symposium on The Ribosome.** Inspired by the fact that ribosomes form 2D crystals in the cells of hibernating bears, Yonath produced crystals of the ribosome in an attempt to solve its structure, the object of her research from long before most people believed its structure could be solved. For her contributions to this achievement, she shared, with Venki Ramakrishnan and Tom Steitz, the 2009 Nobel Prize in Chemistry.



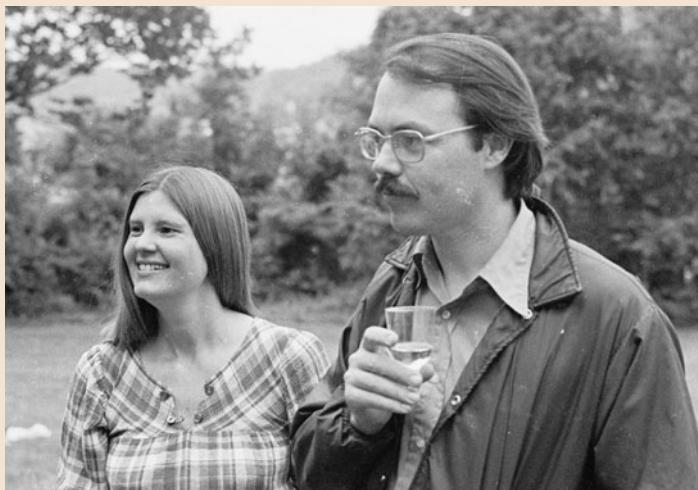
**David Allis and Emily Bernstein, 2004 Symposium on Epigenetics.** Allis was the first to identify an enzyme that modifies histones—a histone acetyltransferase from *Tetrahymena* (Chapter 8). Since that discovery, a whole field has grown up examining the range of histone modifications that exist and their effects on gene expression. Allis is here shown with Bernstein, a postdoc in his lab at the time the photo was taken and formerly a graduate student in Greg Hannon's lab, where she identified the Dicer enzyme involved in RNAi (Chapter 20).



**Robert Roeder, 1998 Symposium on Mechanisms of Transcription.** Roeder discovered the three eukaryotic RNA polymerases—Pol I, II, and III—purifying all three enzymes and other factors they each need to initiate transcription from their respective promoters (Chapter 13). On the left, looking on skeptically, is Camilo Parada, at the time a postdoc in Roeder's lab.



**Roger D. Kornberg, 1977 Symposium on Chromatin.** Having earlier worked on the structure of the nucleosome (Chapter 8), Kornberg won the Nobel Prize in Chemistry in 2006 for his structural studies of RNA polymerase II (Chapter 13). His father is Arthur Kornberg, whose picture is on page 196.



**Phillip Sharp, 1974 Symposium on Tumor Viruses.** Sharp and Richard Roberts shared the 1993 Nobel Prize in Physiology or Medicine for discovering that many eukaryotic genes are "split"—that is, their coding regions are interrupted by stretches of noncoding DNA. The noncoding regions are removed from the RNA copy by "splicing" (Chapter 14). Sharp is shown here with his wife Ann.



**Paul Zamecnik, 1969 Symposium on The Mechanism of Protein Synthesis.** Zamecnik developed *in vitro* systems of protein synthesis that proved critical to understanding how the genetic code works and how cells manufacture proteins (Chapters 2 and 15). Together with Mahlon Hoagland, he also discovered tRNAs, a key component in that process (Chapter 15).



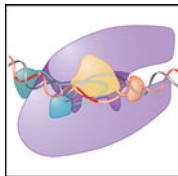
**Richard Roberts, 1977 Symposium on Chromatin.** Much of Roberts' research has focused on the function and diversity of restriction enzymes (Chapter 7), but he was also a codiscoverer of "split genes," for which he shared the Nobel Prize in Physiology or Medicine with Phillip Sharp in 1993. Shown here with him are, left to right, Yasha Gluzman, the tumor virologist; Ahmad Bukhari, who worked on phage Mu transposition (Chapter 12); and James Darnell, whose work focuses on signal transduction in gene regulation (Chapter 19).



**Venki Ramakrishnan and Jack Szostak, 2009 Symposium on Evolution.** Ramakrishnan (left) shared, with Ada Yonath and Tom Steitz, the 2009 Nobel Prize for Chemistry for his work on the crystal structure of the ribosome, while Szostak (center) shared the Physiology or Medicine Prize that same year (with Elizabeth Blackburn and Carol Greider) for his work on telomeres. They are pictured here at the symposium picnic with Alex Gann, one of this book's authors.

*This page intentionally left blank*

CHAPTER 13



# Mechanisms of Transcription

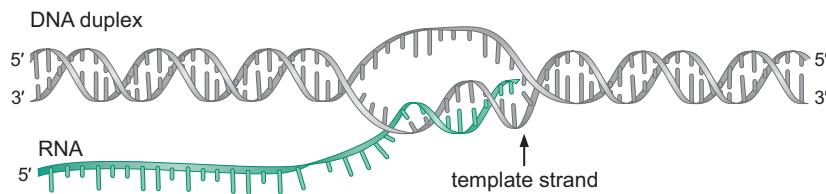
UP TO THIS POINT, WE HAVE BEEN CONSIDERING maintenance of the genome—that is, how the genetic material is organized, protected, and replicated. We now turn to the question of how that genetic material is *expressed*—that is, how the series of bases in the DNA directs the production of the RNAs and proteins that perform cellular functions and define cellular identity. In the next few chapters, we describe the basic processes responsible for gene expression: transcription, RNA processing, and translation.

Transcription is, chemically and enzymatically, very similar to DNA replication (Chapter 9). Both involve enzymes that synthesize a new strand of nucleic acid complementary to a DNA template strand. There are some important differences, of course; most notably, in the case of transcription, the new strand is made from *ribonucleotides* rather than *deoxyribonucleotides* (see Chapter 5). Other mechanistic features of transcription that differ from that of replication include the following.

- **RNA polymerase** (the enzyme that catalyzes RNA synthesis) does not need a primer; rather, it can initiate transcription *de novo* (although *in vivo*, initiation is permitted only at certain sequences, as we shall see).
- The RNA product does not remain base-paired to the template DNA strand: the enzyme displaces the growing chain only a few nucleotides behind where each ribonucleotide is added (Fig. 13-1). This displacement is critical for the RNA to perform its functions (e.g., as is most often the case, to be translated to produce its protein product). Furthermore, because this release follows so closely behind the site of polymerization, multiple RNA polymerase molecules can transcribe the same gene at the same time, each following closely behind another. Thus, a cell can synthesize large numbers of transcripts from a single gene (or other DNA sequence) in a short time. It is important to note that as the RNA product dissociates from the DNA template just behind each advancing RNA polymerase, the two DNA strands reanneal (Fig. 13-1).
- Transcription, although very accurate, is less accurate than replication (one mistake occurs in 10,000 nucleotides added, compared with one in 10 million for replication). This difference reflects the lack of extensive proofreading mechanisms for transcription, although two forms of proofreading for RNA synthesis do exist.

## O U T L I N E

RNA Polymerases and the Transcription Cycle, 430	•
The Transcription Cycle in Bacteria, 434	•
Transcription in Eukaryotes, 448	•
Transcription by RNA Polymerases I and III, 462	•
Visit Web Content for Structural Tutorials and Interactive Animations	



**FIGURE 13-1** Transcription of DNA into RNA. The figure shows, in the absence of the enzymes involved, how the DNA double helix is unwound and an RNA strand is built on the template strand. It also shows how the RNA transcript dissociates from the DNA template a few nucleotides behind the point of synthesis, and how the DNA strands reanneal. In the figure, transcription proceeds from left to right.

It makes sense for the cell to worry more about the accuracy of replication than of transcription. DNA is the molecule in which the genetic material is stored, and DNA replication is the process by which that genetic material is passed on. Any mistake that arises during replication can therefore easily be catastrophic: it becomes permanent in the genome of that individual and gets passed on to subsequent generations. Transcription, in contrast, produces only transient copies and normally several from each transcribed region. Thus, a mistake during transcription will rarely do more harm than render one out of many transient transcripts defective.

Beyond these mechanistic differences between DNA replication and transcription, one profound difference reflects the different purposes served by these processes. Transcription selectively copies only certain parts of the genome and makes anywhere from one to several hundred, or even thousand, copies of any given section. In contrast, replication must copy the entire genome and do so once (and only once) every cell division (as we saw in Chapter 9). The choice of which regions to transcribe is not random: there are specific DNA sequences that direct the initiation of transcription at the start of each region and others at the end that terminate transcription.

Not only are different parts of the genome transcribed to different extents, but the choice of which part to transcribe, and how extensively, can also be regulated. Thus, in different cells, or in the same cell at different times, different sets of genes might be transcribed. Therefore, for example, two genetically identical cells in a human will, in many cases, transcribe different sets of genes, leading to differences in the character and function of those two cells (e.g., one might be a muscle cell and the other a neuron). Or a given bacterial cell will transcribe a different set of genes, depending on the medium in which it is growing. These questions of transcriptional regulation are dealt with in Part 4.

## RNA POLYMERASES AND THE TRANSCRIPTION CYCLE

### RNA Polymerases Come in Different Forms but Share Many Features

RNA polymerase performs essentially the same reaction in all cells, from bacteria to humans. It is thus not surprising that the enzymes from these organisms share many features, especially in those parts of the enzyme directly involved with catalyzing the synthesis of RNA. From bacteria to mammals, the cellular RNA polymerases are made up of multiple subunits (although some phage and organelles do encode single-subunit enzymes that perform the same task, as we shall see in Box 13-2). Table 13-1 shows

**TABLE 13-1** The Subunits of RNA Polymerases

Prokaryotic		Eukaryotic		
Bacterial Core	Archaeal Core	RNAP I (Pol I)	RNAP II (Pol II)	RNAP III (Pol III)
$\beta'$	A'/A''	RPA1	RPB1	RPC1
$\beta$	B	RPA2	RPB2	RPC2
$\alpha^I$	D	RPC5	RPB3	RPC5
$\alpha^{II}$	L	RPC9	RPB11	RPC9
$\omega$	K	RPB6	RPB6	RPB6
	[+six others]	[+nine others]	[+seven others]	[+11 others]

The subunits in each column are listed in order of decreasing molecular weight. (Adapted, with permission, from Ebright R.H. 2000. *J. Mol. Biol.* **304**: 687–698, Fig. 1, p. 688. © Elsevier.)

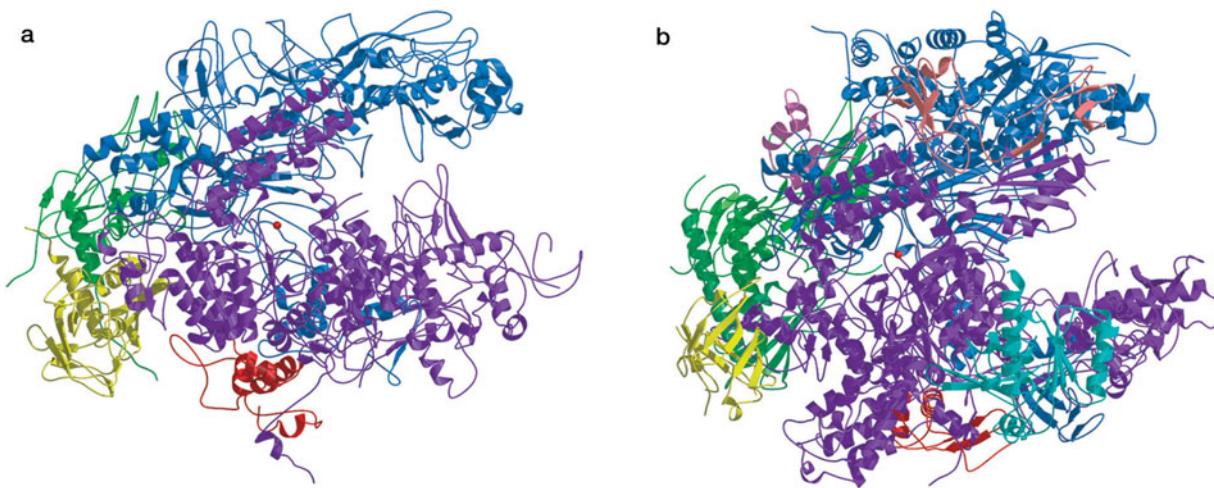
the numbers and sizes of subunits found in each case and also shows which subunits are conserved at the sequence level between different enzymes.

As can be seen from the table, bacteria have only a single RNA polymerase, whereas eukaryotic cells have three: RNA polymerases I, II, and III (RNA Pol I, II, and III). **Pol II** is the enzyme we focus on when dealing with eukaryotic transcription in the second half of this chapter, because it is the most studied of these enzymes. It is also the polymerase responsible for transcribing most genes—indeed, essentially all protein-encoding genes. **Pol I** and **Pol III** are each involved in transcribing specialized, RNA-encoding genes. Specifically, Pol I transcribes the large rRNA precursor gene, whereas Pol III transcribes tRNA genes, some small nuclear RNA genes, and the 5S rRNA gene. We return to these enzymes at the end of the chapter. Finally, two more DNA-dependent RNA polymerases have been identified in recent years, and have been called **Pol IV** and **Pol V**. These are found only in plants, where they transcribe **small interfering RNAs** involved in transcriptional silencing (see Chapter 20). They are both closely related to Pol II and clearly evolved from that enzyme relatively recently: some of their subunits are identical to those from Pol II, encoded by the same genes, and the others are from recently duplicated copies.

The bacterial RNA polymerase **core enzyme** alone is capable of synthesizing RNA and comprises two copies of the  $\alpha$  subunit and one each of the  $\beta$ ,  $\beta'$ , and  $\omega$  subunits. This enzyme is closely related to the eukaryotic polymerases (see Table 13-1). Specifically, the two large subunits,  $\beta$  and  $\beta'$ , are homologous to the two large subunits found in RNA Pol II (RPB1 and RPB2). The  $\alpha$  subunits are homologous to RPB3 and RPB11, and  $\omega$  is homologous to RPB6. The structure of a bacterial RNA polymerase core enzyme is similar to that of the yeast Pol II enzyme. These are shown side by side in Figure 13-2. Later, we describe some of the structural details that shed light on how these enzymes work. For now, we just highlight some of the general features.

The bacterial and yeast enzymes share an overall shape and organization—indeed, they are more alike than the comparison of the subunit sequences would predict. This is particularly true of the internal parts, near the active site, and less so on the peripheries. This distribution of similarities and differences makes sense: the internal parts of the enzyme are involved in synthesis of RNA on a DNA substrate—the same in all organisms; many of the peripheral regions of the enzyme, however, are involved in interactions with other proteins, and these differ in eukaryotic cells compared with prokaryotic cells, as we shall see.

Overall, the shape of each enzyme resembles a crab claw. This is reminiscent of the “hand” structure of DNA polymerases described in Chapter 9 (Fig. 9-5). The two pincers of the crab claw are made up predominantly of the two largest subunits of each enzyme ( $\beta'$  and  $\beta$  for the bacterial case



**FIGURE 13-2** Comparison of the crystal structures of prokaryotic and eukaryotic RNA polymerases. (a) Structure of RNA polymerase core enzyme from *Thermus aquaticus*. The subunits are colored as follows: (blue)  $\beta$ ; (purple)  $\beta'$ ; (yellow and green) the two  $\alpha$  subunits; (red)  $\omega$ . The  $Mg^{2+}$  ion (red ball) marks the active site here and in part b (Seth Darst, The Rockefeller University, pers. comm.). (b) Structure of RNA Pol II from yeast *Saccharomyces cerevisiae*. The subunits are colored to show their relatedness to those in the bacterial enzyme (see Table 13-1). Thus, RPB1 (purple); RPB2 (blue); RPB3 (green) RPB11 (yellow); RPB6 (red). (From Cramer P. et al. 2001. *Science* **292**: 1863.) Images prepared with MolScript, BobScript, and Raster3D.

and RPB1 and RPB2 for the eukaryotic enzyme). The active site, which is made up of regions from both these subunits, is found at the base of the pincers within a region called the “active center cleft” (see Fig. 13-2). The active site works according to the two-metal ion catalytic mechanism for nucleotide addition proposed for all types of polymerase (see Chapter 9). In this case, however, the active site contains only one tightly bound  $Mg^{2+}$  ion, and the second  $Mg^{2+}$  is brought in with each new nucleotide in the addition cycle and released with the pyrophosphate.

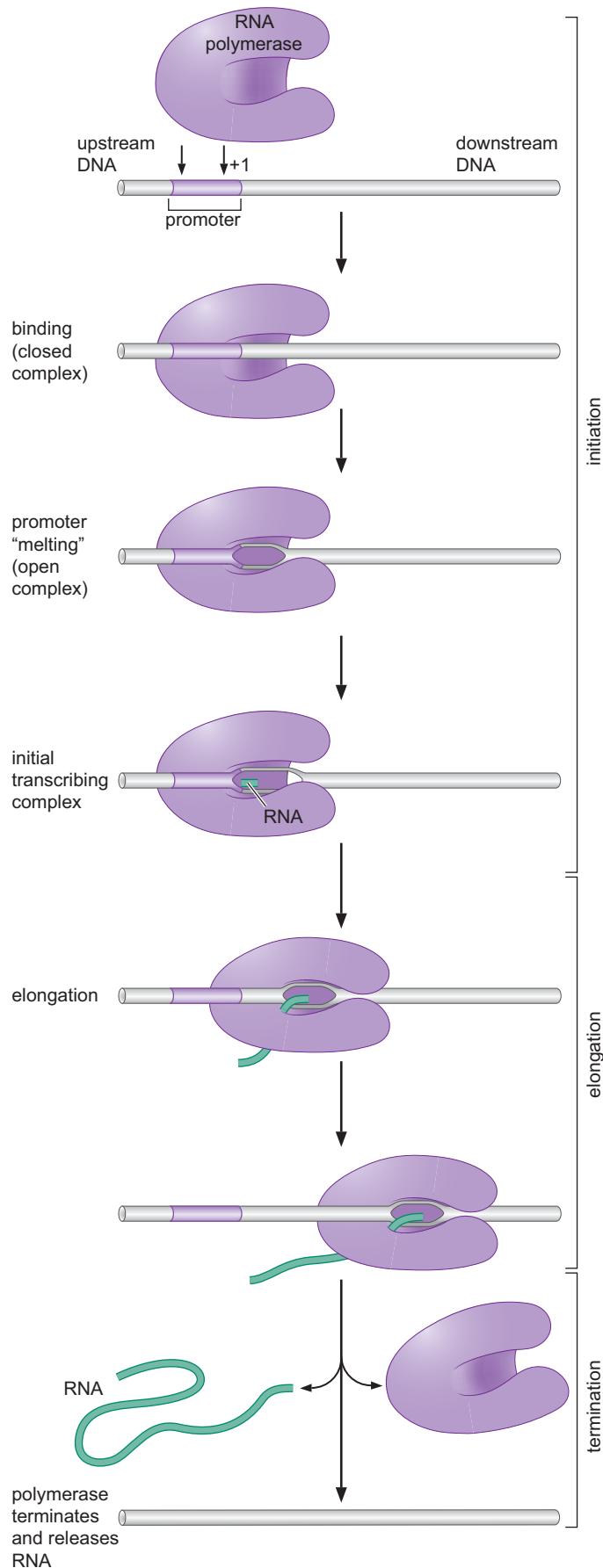
There are various channels that allow DNA, RNA, and ribonucleotides into and out of the enzyme’s active center cleft. We discuss these later when considering the mechanisms of transcription.

### Transcription by RNA Polymerase Proceeds in a Series of Steps

To transcribe a gene, RNA polymerase proceeds through a series of well-defined steps grouped into three phases: **initiation**, **elongation**, and **termination**. Here, and in Figure 13-3, we summarize the basic features of each phase.

**Initiation** A **promoter** is the DNA sequence that initially binds the RNA polymerase (together with any initiation factors required). Once formed, the promoter–polymerase complex undergoes structural changes required for initiation to proceed. As in replication initiation, the DNA around the point where transcription will start unwinds. The base pairs are disrupted, producing a “transcription bubble” of single-stranded DNA. Again, like DNA replication, transcription always occurs in a 5'-to-3' direction: the new ribonucleotide is added to the 3' end of the growing chain. Unlike replication, however, only one of the DNA strands acts as a template on which the RNA strand is built. Because RNA polymerase binds promoters in a defined orientation, the same strand is always transcribed from a given promoter.

The choice of promoter determines which stretch of DNA is transcribed and is the most common step at which regulation is imposed.



**FIGURE 13-3** The phases of the transcription cycle: Initiation, elongation, and termination. The figure shows the general scheme for the transcription cycle. The features shown hold for both bacterial and eukaryotic cases. Other factors required for initiation, elongation, and termination are not shown here but are described in the text. The DNA nucleotide encoding the beginning of the RNA chain is called the transcription start site and is designated the "+1" position. Sequences in the direction in which transcription proceeds are referred to as downstream from the start site. Likewise, sequences preceding the start site are referred to as upstream sequences. When referring to a specific position in the upstream sequence, this is given a negative value. Downstream sequences are allotted positive values.

**Elongation** Once the RNA polymerase has synthesized a short stretch of RNA (~10 bases), it shifts into the elongation phase. During elongation, the enzyme performs an impressive range of tasks in addition to the catalysis of RNA synthesis. It unwinds the DNA in front and reanneals it behind; it dissociates the growing RNA chain from the template as it moves along; and it performs proofreading functions. Recall that during replication, in contrast, several different enzymes are required to catalyze a similar range of functions.

**Termination** Once the polymerase has transcribed the length of the gene (or genes), it must stop and release the RNA product (as well as dissociating from the DNA itself). This step is called *termination*. In some cells, specific, well-characterized sequences trigger termination. In others, it is less clear what instructs the enzyme to cease transcribing and dissociate from the template.

### Transcription Initiation Involves Three Defined Steps

The first phase in the transcription cycle—initiation—can itself be broken down into a series of defined steps (as indicated in Fig. 13-3). The first step is the initial binding of polymerase to a promoter to form what is called a **closed complex**. In this form, the DNA remains double-stranded, and the enzyme is bound to one face of the helix. In the second step of initiation, the closed complex undergoes a transition to the **open complex** in which the DNA strands separate over a distance of ~13 bp around the start site to form the transcription bubble. In the next stage of initiation, polymerase enters the phase of initial transcription followed by promoter escape, as we now describe.

The opening up of the DNA frees the template strand. The first two ribonucleotides are brought into the active site, aligned on the template strand, and joined together. In the same way, subsequent ribonucleotides are incorporated into the growing RNA chain. Incorporation of the first 10 or so ribonucleotides is a rather inefficient process, and at that stage, the enzyme often releases short transcripts (each of less than 10 or so nucleotides) and then begins synthesis again. In this phase, the polymerase–promoter complex is called the **initial transcribing complex**. Once an enzyme makes a transcript longer than 10 nucleotides, it is said to have **escaped** the promoter. At this point, it has formed a stable ternary complex, containing enzyme, DNA, and RNA. This is the transition to the elongation phase.

In the remainder of this chapter, we describe the transcription cycle in more detail—first for the bacterial case and then for eukaryotic systems.

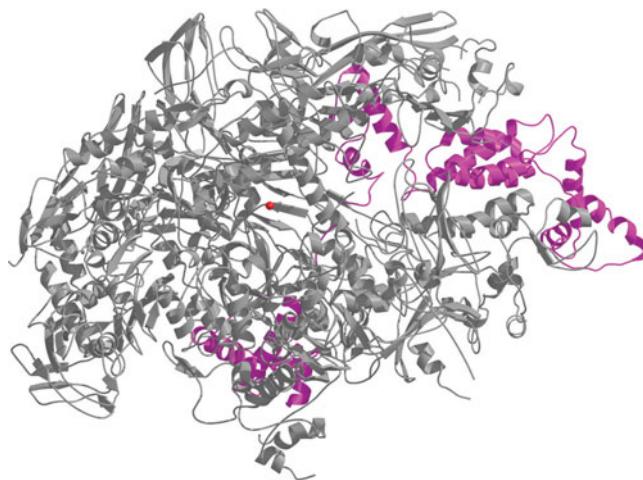
## THE TRANSCRIPTION CYCLE IN BACTERIA

---

### Bacterial Promoters Vary in Strength and Sequence but Have Certain Defining Features

The bacterial core RNA polymerase can, in principle, initiate transcription at any point on a DNA molecule, and this can be shown *in vitro* using purified core enzyme. In cells, however, polymerase initiates transcription only at promoters. It is the addition of an initiation factor called  $\sigma$  that converts core enzyme ( $\alpha_2 \beta \beta' \omega$ ) into the form that initiates only at promoters. This form of the enzyme is called the RNA polymerase **holoenzyme** (Fig. 13-4).

In the case of *Escherichia coli*, the predominant  $\sigma$  factor is called  $\sigma^{70}$  (we consider other alternative  $\sigma$  factors and their roles in transcriptional regulation in Chapters 18 and 22). Promoters recognized by polymerase containing  $\sigma^{70}$  share the following characteristic structure: two conserved sequences,

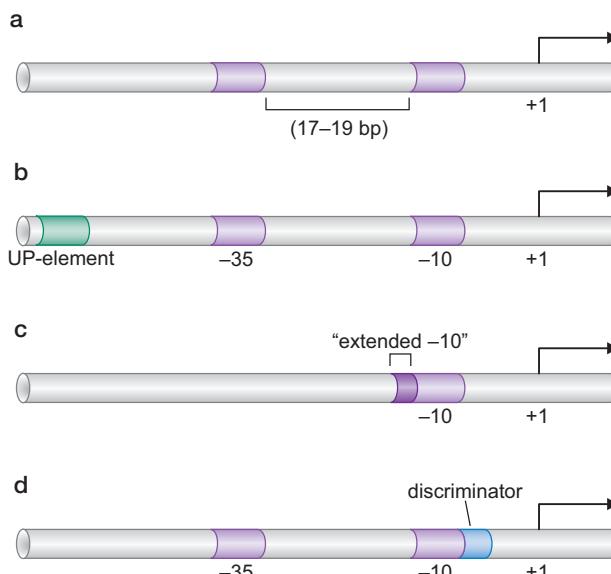


**FIGURE 13-4** RNA polymerase holoenzyme from *Thermus aquaticus*. Shown in gray is the core enzyme (the same enzyme shown in Fig. 13-2a). The  $\sigma^{70}$  subunit is shown in purple (regions 2, 3, and 4; see Fig. 13-6). On the right is region 2; at the top, region 3; and at the bottom, region 4. As described later in the text, it is  $\sigma$  regions 2 and 4 that recognize the  $-10$  and  $-35$  regions of the promoter, respectively. (From Murakami K.S. et al. 2002. *Science* **296**: 1280.) Image prepared with MolScript, BobScript, and Raster3D.

each of 6 nucleotides, separated by a nonspecific stretch of 17–19 nucleotides (Fig. 13-5a). The two defined sequences are centered, respectively, at ~10 bp and at ~35 bp upstream of the site where RNA synthesis starts. The sequences are thus called the  **$-35$**  (minus 35) and  **$-10$**  (minus 10) **regions, or elements**, according to the numbering scheme described in Figure 13-3, in which the DNA nucleotide encoding the beginning of the RNA chain is designated +1.

Although the vast majority of  $\sigma^{70}$  promoters contain recognizable  $-35$  and  $-10$  regions, the sequences are not identical. By comparing many different promoters, a **consensus sequence** can be derived (for a discussion of how these are derived, see Box 13-1, Consensus Sequences). The consensus sequence reflects preferred  $-10$  and  $-35$  regions, separated by the optimum spacing (17 bp). Very few promoters have this exact sequence, but most differ from it only by a few nucleotides.

Promoters with sequences closer to the consensus are generally “stronger” than those that match less well. By the strength of a promoter, we mean how many transcripts it initiates in a given time. That measure is influenced by how well the promoter binds polymerase initially, how efficiently it supports



**FIGURE 13-5** Features of bacterial promoters. Various combinations of bacterial promoter elements are shown. Details of how each element contributes to polymerase binding and function are described in the text.

## TECHNIQUES

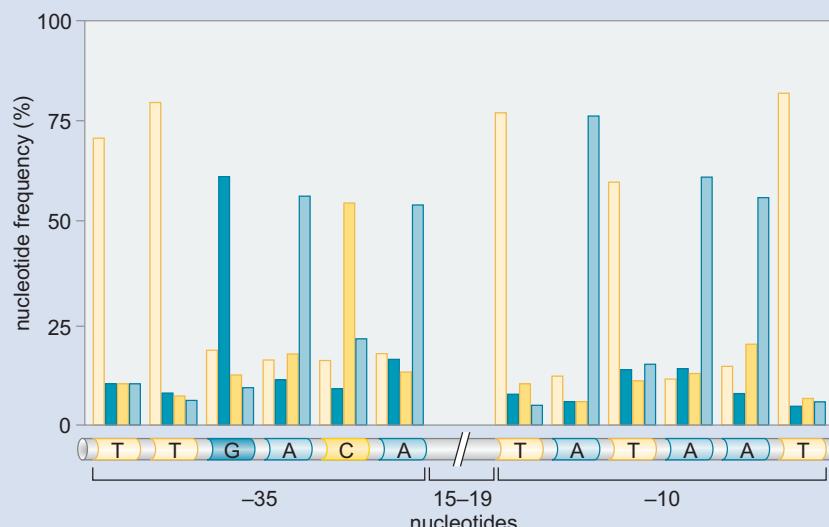
### Box 13-1 Consensus Sequences

The DNA sequences of binding sites recognized by a given protein may not always be exactly the same. Likewise, a stretch of amino acids that bestows upon a protein a particular function may be slightly different in different proteins. A consensus sequence is, in each case, a version of the sequence having at each position the nucleotide (or amino acid) most commonly found there in different examples. Thus, the consensus sequence for promoters in *E. coli* recognized by RNA polymerase containing  $\sigma^{70}$  is shown in the figure (Box 13-1 Fig. 1). This consensus sequence was derived by aligning 300 sequences known to function as  $\sigma^{70}$  promoters and ascertaining the most common base found at each position in the  $-35$  and  $-10$  hexamers. That nucleotide is then chosen as the nucleotide of choice at that position in the consensus; its relative frequency and the frequencies with which the other three nucleotides occur at each position are portrayed in the graph. Note that there is no significant consensus among the 17–19 nucleotides that lie in the region between  $-35$  and  $-10$ .

In that example, each individual promoter sequence had previously been identified, thus aligning the sequences is tri-

vial. But consider a rather different example. In this case, no binding site has been identified for the DNA-binding protein in question. However, several regions of a chromosome are known to contain binding sites somewhere within their lengths. A computer algorithm is used that scans each of the sequences of these chromosomal regions, searching for a potential binding site common to them all.

A second approach to deriving the consensus sequence for a DNA-binding protein when the binding site is not already known takes advantage of chemical methods for synthesizing vast sets of short DNA fragments of random sequence. The protein of interest is mixed with the population of DNA molecules, and those DNAs to which it binds are retrieved and sequenced. A comparison of the sequences bound reveals the consensus readily, because each of the fragments is very short. This last method (often called SELEX) is widely used to define binding sites for previously uncharacterized DNA-binding proteins. SELEX is described in more detail in Chapter 7.



**BOX 13-1 FIGURE 1** Promoter consensus sequence and spacing consensus. (Redrawn, with permission, from Alberts B. et al. 2002. *Molecular biology of the cell*, 4th ed., p. 308, Fig. 6.12. © Garland Science/Taylor & Francis Books LLC.)

isomerization, and how readily the polymerase can then escape. The correlation between promoter strength and sequence explains why promoters are so heterogeneous: some genes need to be expressed more highly than others, and the former are likely to have sequences closer to the consensus.

An additional DNA element that binds RNA polymerase is found in some strong promoters, for example, those directing expression of the ribosomal RNA (rRNA) genes. This is called an **UP-element** (see Fig. 13-5b) and increases polymerase binding by providing an additional specific interaction between the enzyme and the DNA.

Another class of  $\sigma^{70}$  promoters lacks a  $-35$  region and instead has a so-called “extended  $-10$ ” element (see Fig. 13-5c). This comprises a standard  $-10$  region with an additional short sequence element at its upstream

end. Extra contacts made between polymerase and this additional sequence element compensate for the absence of a  $-35$  region. The *E. coli gal* genes (whose products direct metabolism of the sugar galactose; see Chapter 18) use such a promoter.

A final DNA element that binds RNA polymerase is sometimes found just downstream from the  $-10$  element. This element is called the **discriminator** and is shown in Figure 13-5d. The strength of the interaction between the discriminator and polymerase influences the stability of the complex between the enzyme and the promoter.

### The $\sigma$ Factor Mediates Binding of Polymerase to the Promoter

The  $\sigma^{70}$  factor can be divided into four regions called  $\sigma$  region 1 through  $\sigma$  region 4 (see Fig. 13-6). The regions that recognize the  $-10$  and  $-35$  elements of the promoter are regions 2 and 4, respectively.

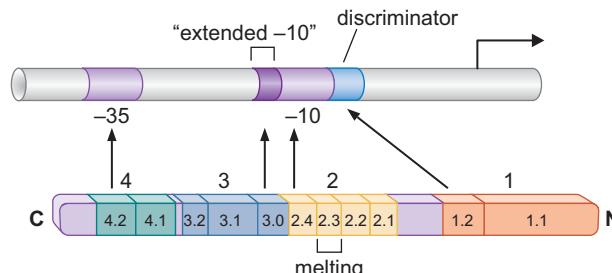
Two helices within region 4 form a common DNA-binding motif called a **helix-turn-helix**. One of these helices inserts into the major groove and interacts with bases in the  $-35$  region; the other lies across the top of the groove, making contacts with the DNA backbone. This structural motif is found in many DNA-binding proteins—for example, almost all transcriptional activators and repressors found in bacterial cells (described in Chapter 18)—and was discussed in Chapter 6 (Fig. 6-13).

The  $-10$  region is also recognized by an  $\alpha$  helix. But in this case, the interaction is more complicated: whereas the  $-35$  region simply provides binding energy to secure polymerase to the promoter, the  $-10$  region has a more elaborate role in transcription initiation, because it is within that element that DNA melting is initiated in the transition from the closed to the open complex. Thus, the region of  $\sigma$  that interacts with the  $-10$  region is doing more than simply binding DNA. In keeping with this expectation, the  $\alpha$  helix involved in recognition of the  $-10$  region contains several essential aromatic amino acids that can interact with bases on the non-template strand in a manner that stabilizes the melted DNA. In Chapter 9, we described a similar role for the single-strand binding protein (SSB) during DNA replication.

Recent structural studies of  $\sigma$  region 2 bound to a single-stranded  $-10$  element, and also of the entire intact open complex, reveal exactly how melting is driven through favorable binding interactions between  $\sigma$  and the single-stranded DNA. Two bases in the non-template strand are flipped out and inserted into pockets within the  $\sigma$  protein where they make favorable contacts that stabilize the unwound state of the promoter region.

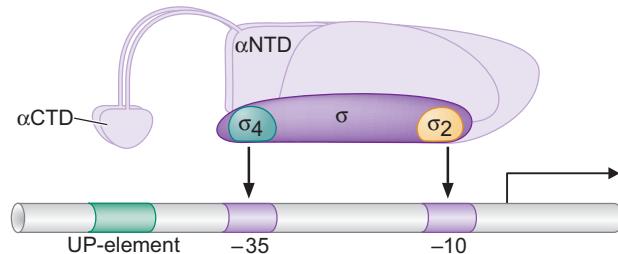
The extended  $-10$  element, where present, is recognized by an  $\alpha$  helix in  $\sigma$  region 3. This helix makes contact with the two specific base pairs that constitute that element. The discriminator is recognized by  $\sigma$  region 1.2.

Unlike the other elements within the promoter, the UP-element is not recognized by  $\sigma$  but is recognized by a carboxy-terminal domain of the  $\alpha$



**FIGURE 13-6** Regions of  $\sigma$ . Those regions of  $\sigma$  factor that recognize specific regions of the promoter are indicated by arrows. Region 2.3 is responsible for melting the DNA. For a schematic view of  $\sigma$  recruiting RNA polymerase core enzyme to a standard promoter, see Figure 13-7. (Adapted, with permission, from Young B.A. et al. 2002. *Cell* 109: 417–420, Fig. 1. © Elsevier.)

**FIGURE 13-7**  $\sigma$  and  $\alpha$  subunits recruit RNA polymerase core enzyme to the promoter. The carboxy-terminal domain of the  $\alpha$  subunit ( $\alpha$ CTD) recognizes the UP-element (where present), whereas  $\sigma$  regions 2 and 4 recognize the  $-10$  and  $-35$  regions, respectively (see Fig. 13-6). In this figure, RNA polymerase is shown in a schematic representation rather different from that presented in earlier figures. This representation is particularly useful for indicating surfaces that touch DNA and regulatory proteins, and we use it again in some figures in Chapter 18 when we consider regulation of transcription in bacteria.



subunit, called the  **$\alpha$ CTD** (Fig. 13-7). The  $\alpha$ CTD is connected to the  $\alpha$ NTD by a flexible linker. Thus, although the  $\alpha$ NTD is embedded in the body of the enzyme, the  $\alpha$ CTD can reach the upstream element and can do so even when that element is not located immediately adjacent to the  $-35$  region, but further upstream.

The  $\sigma$  subunit is positioned within the holoenzyme structure in such a way as to make feasible the recognition of various promoter elements. Thus, the DNA-binding regions point away from the body of the enzyme, rather than being embedded. Moreover, the spacing between those regions is consistent with the distance between the DNA elements they recognize:  $\sigma$  regions 2 and 4 are separated by  $\sim 75$  Å when  $\sigma$  is bound in the holoenzyme, and this is about the same distance as that between the centers of the  $-10$  and  $-35$  elements of a typical  $\sigma^{70}$  promoter (see Fig. 13-7). This rather large spacing of the protein domains is accommodated by the region between  $\sigma$  regions 2 and 4, that is, by region 3—especially region 3.2, also called the  $\sigma_{3/4}$  linker (see Figs. 13-4 and 13-6).

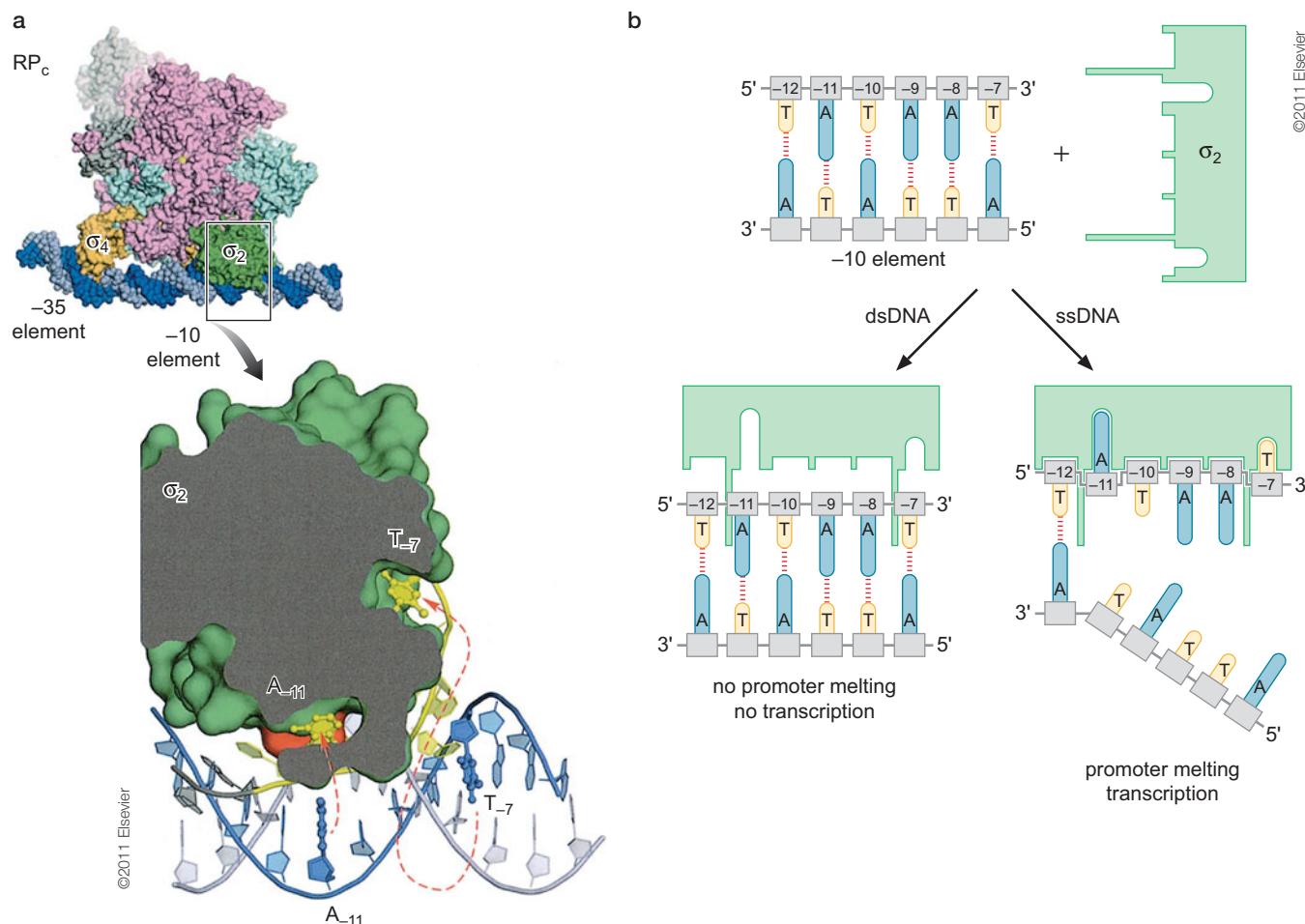
### Transition to the Open Complex Involves Structural Changes in RNA Polymerase and in the Promoter DNA

The initial binding of RNA polymerase to the promoter DNA in the closed complex leaves the DNA in double-stranded form. The next stage in initiation requires the enzyme to become more intimately engaged with the promoter, in the open complex. The transition from the closed to the open complex involves structural changes in the enzyme and the opening of the DNA double helix to reveal the template and nontemplate strands. This “melting” occurs between positions  $-11$  and  $+2$ , with respect to the transcription start site.

In the case of the bacterial enzyme bearing  $\sigma^{70}$ , this transition, often called **isomerization**, does not require energy derived from ATP hydrolysis and is instead the result of a spontaneous conformational change in the DNA–enzyme complex to a more energetically favorable form. As we noted above, two bases in the non-template strand of the  $-10$  element ( $A_{11}$  and  $T_7$ ) flip out from their base-stacking interactions and instead insert into pockets within the  $\sigma$  protein where they make more favorable interactions. By stabilizing the single-stranded form of the  $-10$  element, these interactions drive melting of the promoter region (see Fig. 13-8).

Isomerization is essentially irreversible and, once complete, typically guarantees that transcription will subsequently initiate (although regulation can still be imposed after this point in some cases). Formation of the closed complex, in contrast, is readily reversible: polymerase can as easily dissociate from the promoter as make the transition to the open complex.

To picture the global structural changes within the polymerase that accompany isomerization, we need to examine the structure of the holoenzyme in more detail. A channel runs between the pincers of the claw-shaped



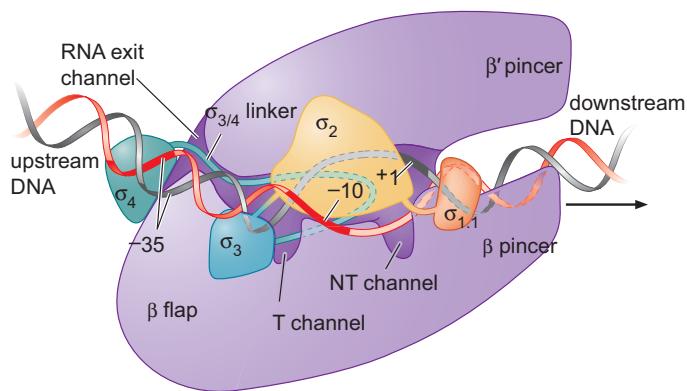
**FIGURE 13-8** Recognition and melting of the  $-10$  elements by  $\sigma$  region 2.  $\sigma$  region 2 has two pockets that each bind one flipped-out base of the non-template strand of the  $-10$  element. These energetically preferred binding reactions drive the melting of the promoter and thus the transition from the closed to the open complex, without the need for ATP hydrolysis. (a) The magnified view of  $\sigma_2$  of Taq RNA polymerase, with the dsDNA of the closed complex (blue) and the portion of ssDNA of the open complex (yellow). The cut-away reveals the two flipped-out bases, A and T (yellow), in the binding pockets. Red arrows show how the flipped-out bases relate to the same nucleotides in the closed complex. (Image reproduced, with permission, from Feklistov A. and Darst S.A. 2011. *Cell* 147: 1257, Fig. 6C, p. 1265. © Elsevier.) (b) The image displayed in panel a is rendered here schematically to show more clearly the interaction between  $\sigma$  region 2 and the non-template strand of the  $-10$  region, specifically the flipped out bases that drive DNA melting in this region. (Adapted, with permission, from Liu X., Bushnell D.A., and Kornberg R.D. 2011. *Cell* 147: 1218, Fig. 1, p. 1219. © Elsevier.)

enzyme, as we described above (see Fig. 13-2). The active site of the enzyme, which is made up of regions from both the  $\beta$  and  $\beta'$  subunits, is found at the base of the pincers within the active center cleft.

There are five channels into the enzyme, as shown in the illustration of the open complex in Figure 13-9. The NTP-uptake channel (not shown in the figure; see figure caption) allows ribonucleotides to enter the active center. The RNA-exit channel allows the growing RNA chain to leave the enzyme as it is synthesized during elongation. The remaining three channels allow DNA entry and exit from the enzyme, as follows.

The downstream DNA (i.e., DNA ahead of the enzyme, yet to be transcribed) enters the active center cleft in double-stranded form through the downstream DNA channel (between the pincers). Within the active center

**FIGURE 13-9** Channels into and out of the open complex. This figure shows the relative positions of the DNA strands (template strand in gray, nontemplate strand in orange), the four regions of  $\sigma$ , the  $-10$  and  $-35$  regions of the promoter, and the start site of transcription ( $+1$ ). The channels through which DNA and RNA enter or leave the RNA polymerase enzyme are also shown. The only channel not shown here is the nucleotide entry (NTP-uptake) channel, through which nucleotides enter the active site cleft for incorporation into the RNA chain as it is made. As drawn, that channel would enter the active site from the back of the page at about the position shown as " $+1$ " on the DNA. Where a DNA strand passes underneath a protein, it is drawn as a dotted ribbon.  $\sigma$  region 3/4 linker—also called  $\sigma_{3.2}$ —is the linker region between  $\sigma_{3.1}$  and  $\sigma_4$ . (Original figure design by Richard Ebright.)



cleft, the DNA strands separate from position +3. The non-template strand exits the active center cleft through the non-template-strand (NT) channel and travels across the surface of the enzyme. The template strand, in contrast, follows a path through the active center cleft and exits through the template-strand (T) channel. The double helix re-forms at  $-11$  in the upstream DNA behind the enzyme.

Two striking structural changes are seen in the enzyme upon isomerization from the closed to the open complex. First, the pincers at the front of the enzyme clamp down tightly on the downstream DNA. Second, there is a major shift in the position of the amino-terminal region of  $\sigma$  (region 1.1). When not bound to DNA,  $\sigma$  region 1.1 lies within the active center cleft of the holoenzyme, blocking the path that, in the open complex, is followed by the template DNA strand. In the open complex, region 1.1 shifts some 50 Å and is now found on the outside of the enzyme, allowing the DNA access to the cleft (see Fig. 13-9). Region 1.1 of  $\sigma$  is highly negatively charged (just like DNA). Thus, in the holoenzyme, region 1.1 acts as a **molecular mimic** of DNA. The space in the active center cleft, which may be occupied either by region 1.1 or by DNA, is highly positively charged.

### Transcription Is Initiated by RNA Polymerase without the Need for a Primer

Recall from Chapter 9 that DNA polymerase does not synthesize new DNA strands de novo—that is, it can only extend an existing polynucleotide chain. For this reason, replication always requires a primer strand. The primer is typically a short piece of RNA that binds to the DNA template strand to form a short hybrid double-strand region. DNA polymerase then adds nucleotides to the 3' end of the primer.

RNA polymerase can initiate a new RNA chain on a DNA template and thus does not need a primer. This impressive feat requires that the DNA template be brought into the polymerase active site and held stably in a helical conformation and that the initiating ribonucleotide be brought into the active site and held stably on the template while the next NTP is presented with correct geometry for the chemistry of polymerization to occur. This is particularly difficult because RNA polymerase starts most transcripts with an A, and that ribonucleotide binds the template nucleotide (T) with only two hydrogen bonds (as opposed to the three between C and G).

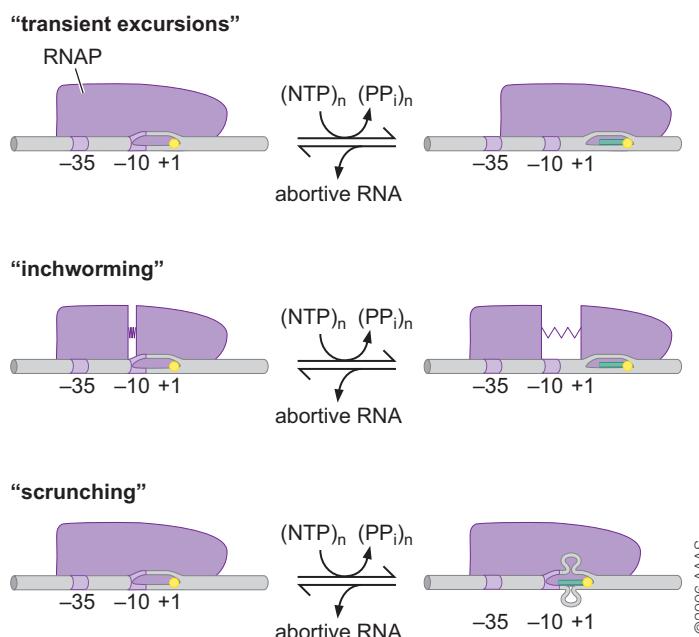
Thus, the enzyme has to make specific interactions with one or all of the DNA template strand, the initiating ribonucleotide, and the second ribonucleotide—holding one (or all) rigidly in the correct orientation to allow

chemical attack on the incoming NTP. The requirement for such specific interactions between the enzyme and the initiating nucleotide probably explains why most transcripts start with the same nucleotide. The structure of the open complex shows that the  $\sigma$  region 3/4 linker interacts with the template strand, organizing it in the correct conformation and location to allow initiation. Consistent with this, in experiments using an RNA polymerase containing a  $\sigma^{70}$  derivative lacking this part of  $\sigma$ , initiation requires much higher than normal concentrations of one or both of the first two ribonucleotides.

### During Initial Transcription, RNA Polymerase Remains Stationary and Pulls Downstream DNA into Itself

As we have outlined, during initial transcription, RNA polymerase produces and releases short RNA transcripts of  $<10$  nucleotides (abortive synthesis) before escaping the promoter, entering the elongation phase, and synthesizing the proper transcript. It has long been unclear how the enzyme's active site translocates along the DNA template during initial abortive cycles of transcription. Three general models were proposed (as shown in Fig. 13-10 and described below).

1. The “transient excursion” model proposes transient cycles of forward and reverse translocation of RNA polymerase. Thus, polymerase is thought to leave the promoter and translocate a short way along the DNA template, synthesizing a short transcript before aborting transcription, releasing the transcript, and returning to its original location on the promoter.
2. “Inchworming” invokes a flexible element within the polymerase that allows a module at the front of the enzyme, containing the active site, to move downstream, synthesizing a short transcript before aborting and retracting to the body of the enzyme still at the promoter.
3. “Scrunching” proposes that DNA downstream from the stationary, promoter-bound, polymerase is unwound and pulled into the enzyme.



**FIGURE 13-10 Mechanism of initial transcription.** During initial transcription, the active center of RNA polymerase is translocated forward relative to the DNA template and synthesizes short transcripts before aborting, then repeats this cycle until it escapes the promoter. Three models have been proposed to account for this and are shown in the figure. According to the first of these—transient excursions (shown at the top)—polymerase moves along the DNA. In the second—inchworming (shown in the middle)—the front part of the enzyme moves along the DNA, but because of a flexible region within the enzyme, the back part of the enzyme can remain stationary at the promoter. In the third model—scrunching (shown at the bottom)—the enzyme remains stationary and pulls the DNA into itself. The differences between these models are explained in the text, as is the evidence supporting scrunching as the true picture of what goes on. (Modified, with permission, from Kapanidis A.N. et al. 2006. *Science* 314: 1144–1147, Fig. 1a. © AAAS.)

The DNA thus accumulated within the enzyme is accommodated as single-stranded bulges.

It is now believed that the third model—**scrunching**—reflects what actually happens. This conclusion is based on a number of findings, including experiments using single-molecule analyses that allow the positions of different parts of polymerase to be measured relative to each other and to the template DNA during initial transcription. These experiments show that during initial transcription, the polymerase remains stationary on the promoter, unwinds downstream DNA, and pulls that DNA into itself. Only the scrunching model is consistent with these results.

### Promoter Escape Involves Breaking Polymerase–Promoter Interactions and Polymerase Core– $\sigma$ Interactions

As we have seen, during initial transcription, the process of abortive initiation takes place, and short—9 nucleotides or shorter—transcripts are generated and released. Polymerase manages to escape from the promoter and enter the elongation phase only once it has managed to synthesize a transcript of a threshold length of 10 or more nucleotides. Once this length, the transcript cannot be accommodated within the region where it hybridizes to the DNA and must start threading into the RNA exit channel (Fig. 13-9). Promoter escape is associated with the breaking of all interactions between polymerase and promoter elements and between polymerase and any regulatory proteins operating at the given promoter (Chapter 18).

It is not clear why RNA polymerase must undergo this period of abortive initiation before achieving escape, but once again a region of the  $\sigma$  factor appears to be involved, acting as a molecular mimic. In this case, it is the region 3/4 linker, and it mimics RNA. This region of  $\sigma$  lies in the middle of the RNA exit channel in the open complex (see Fig. 13-9), and for an RNA chain to be made longer than  $\sim$ 10 nucleotides, this region of  $\sigma$  must be ejected from that location, a process that can take the enzyme several attempts.

The ejection of the  $\sigma$  region 3/4 linker probably accounts for  $\sigma$  being more weakly associated with the elongating enzyme than it is with the open complex; indeed, it is often lost altogether from the elongating complex.

Scrunching is reversed upon escape: the DNA unwound during scrunching is rewound, with concomitant collapse of the transcription bubble from a size of 22–24 nucleotides back down to 12–14 nucleotides (Fig. 13-3). It is believed that this process provides the energy required by polymerase to break the polymerase–promoter and core– $\sigma$  interactions associated with escape. Thus, scrunching is a way to store and mobilize energy during transcription initiation, and its release upon escape is what enables polymerase to break free of the promoter and dislodge  $\sigma$  factor from the core.

In Box 13-2, The Single-Subunit RNA Polymerases, we see how these simple RNA polymerases, despite lacking a  $\sigma$  subunit, undergo a structurally comparable shift in transition from the initiating to the elongating complex.

### The Elongating Polymerase Is a Processive Machine That Synthesizes and Proofreads RNA

DNA passes through the elongating enzyme in a manner very similar to its passage through the open complex (Fig. 13-9). Thus, double-stranded DNA enters the front of the enzyme between the pincers. At the opening of the catalytic cleft, the strands separate to follow different paths through the enzyme before exiting via their respective channels and re-forming a

► ADVANCED CONCEPTS

**Box 13-2** The Single-Subunit RNA Polymerases

In the text, we discuss the multi-subunit RNA polymerases found in bacteria and eukaryotic cells. But there are several examples of single-subunit RNA polymerases that are capable of performing the same basic reaction as their more complex multicellular counterparts. Thus, many bacteriophage, for example, the *E. coli* phage T7, encode polymerases of this type with which, upon infection, they transcribe most of their genes. Similarly, the majority of mitochondrial and chloroplast genes are transcribed by polymerases closely related to the single-subunit phage enzymes. It is remarkable that evolution has produced these relatively simple enzymes capable of performing transcription, a task that we, in the text, emphasize as an impressive achievement even for the much larger and more complicated multi-subunit enzymes.

The T7 polymerase is the most widely studied of the single-subunit enzymes. It has a molecular mass of 100 kDa—compared with 400 kDa for the bacterial core enzyme (without  $\sigma$  factor)—and a structure shown in Box 13-2 Figure 1. Overall, it looks like the Pol I family of DNA polymerases that we considered in Chapter 9. Thus, the T7 RNA polymerase resembles a right hand, with the fingers, thumb, and palm representing domains arranged around a central cleft, within which lies the active site.

Although it is not structurally related to the cellular RNA polymerases (and instead is structurally related to the DNA polymerases), the T7 enzyme does have features functionally analogous to the cellular RNA polymerases as well, features that have become more apparent since the structure of the T7 and bacterial enzymes have been compared in complex with their templates. As we saw in the text, the bacterial enzyme has various channels into and out of the active center cleft (see Fig. 13-8). One of these, for example, allows the NTPs access to the active site and template, where they are polymerized, under the influence of the template, into the growing RNA chain. Another channel provides the growing RNA chain an exit from the enzyme. Analogous channels are seen in the structure of the phage polymerase as well.

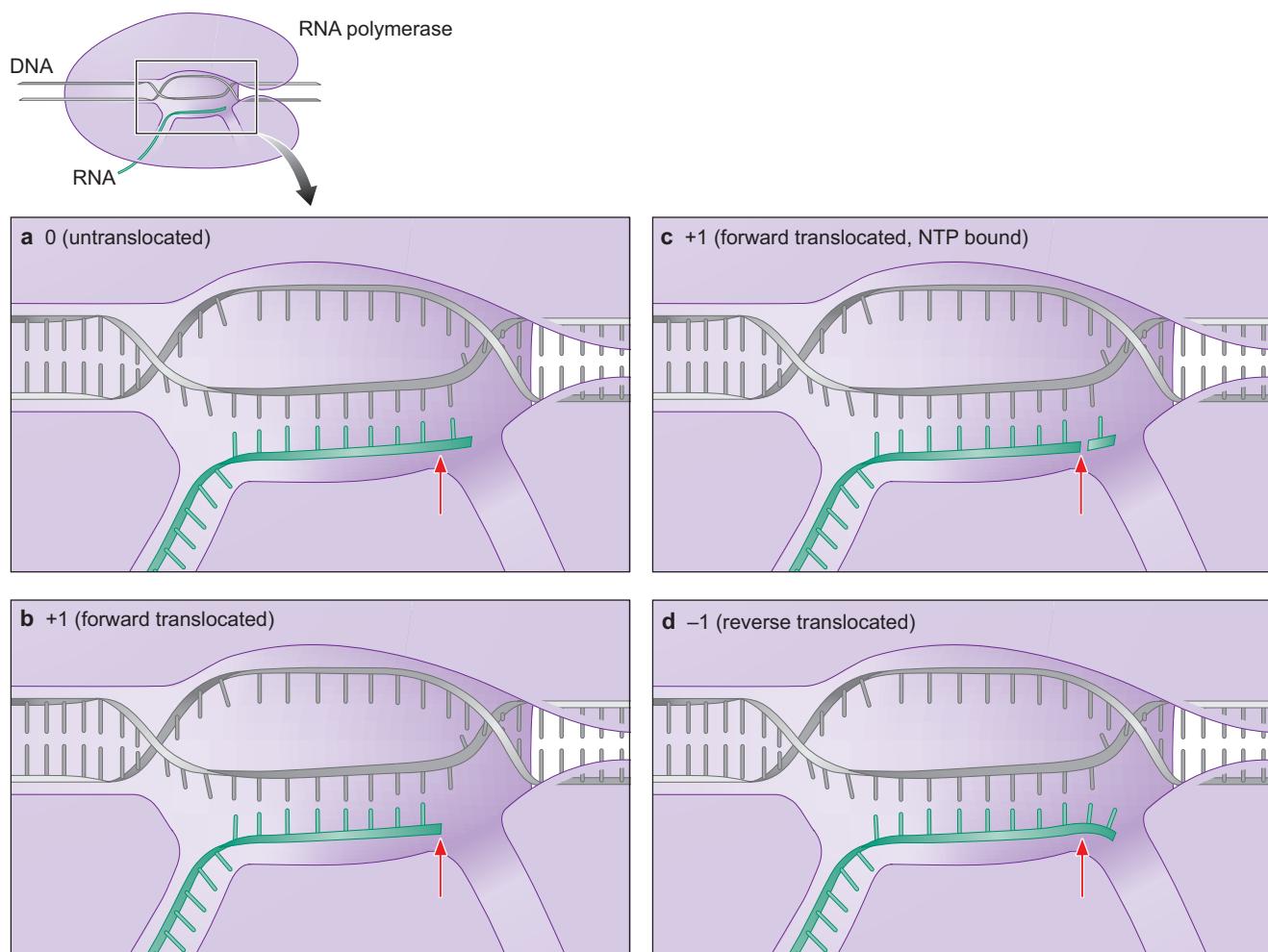
The initiation and elongation complexes of the bacterial and T7 polymerases have been compared. These comparisons highlight one striking example of how an analogous functional transition can be achieved through different kinds of structural change in the two cases. We note in the text that in the bacterial case the transition from initiation to elongation involves a significant shift in the location of a domain of the  $\sigma$  factor. This movement opens up the RNA-exit channel, thereby allowing production of transcripts larger than 10 nucleotides in length. The T7 enzyme has no  $\sigma$  factor, but a comparable structural change in the body of that single-subunit enzyme mediates the transition from the initiating to the elongating complex, and this structural change is required to form the RNA-exit channel.



**BOX 13-2 FIGURE 1** Bacteriophage T7 RNA polymerase. (From Jeruzalmi D. and Steitz T.A. 1998. *EMBO J.* 17: 4101.) Image prepared with MolScript, BobScript, and Raster3D.

double helix behind the elongating polymerase. Ribonucleotides enter the active site through their defined channel and are added to the growing RNA chain under the guidance of the template DNA strand. Only 8 or 9 nucleotides of the growing RNA chain remain base-paired to the DNA template at any given time; the remainder of the RNA chain is peeled off and directed out of the enzyme through the RNA exit channel. See Figure 13-11 for a schematic diagram of the elongating complex.

During elongation, the enzyme adds one nucleotide at a time to the growing RNA transcript. In contrast to initial transcription, where polymerase uses scrunching and this pulls DNA into the enzyme (Fig. 13-10), during elongation polymerase uses a step mechanism: using single-molecule techniques, it was shown that the enzyme steps forward as a molecular motor, advancing in a single step a distance equivalent to a base pair for every nucleotide it adds to the growing RNA chain. In addition, the size of the bubble, that is, the length of DNA that is not double-helical, remains constant



**FIGURE 13-11** Template and transcript within the RNA polymerase elongating complex.

The figure shows schematic diagrams of the relative positions of RNA and the DNA template within RNA polymerase at various states of the transcription process. (a) Untranslocated polymerase (0) shows the RNA chain paired with the template DNA strand for a 9-base stretch. (b) Forward translocated polymerase (+1) shows the situation when the enzyme has translocated one base forward. (c) Forward translocated polymerase with NTP bound shows the DNA and RNA in the same position as in b with the incoming NTP bound. (d) Reverse translocated polymerase (-1) shows the situation when the enzyme is translocated backward one base as it does during hydrolytic editing. (Red arrow) A set position within the polymerase, the same in all parts of the figure. See text for more details. For clarity, the polymerase shown here is in a different orientation from that in Figure 13-9, with the RNA exit channel downward. (Figures based on images courtesy of Richard Ebright.)

throughout elongation: as 1 bp is separated ahead of the processing enzyme, 1 bp is formed behind it.

As well as synthesizing the transcript, RNA polymerase performs two proofreading functions on that growing transcript. The first of these is called **pyrophosphorolytic editing**. In this, the enzyme uses its active site, in a simple back-reaction, to catalyze the removal of an incorrectly inserted ribonucleotide, by reincorporation of PP<sub>i</sub>. The enzyme can then incorporate another ribonucleotide in its place in the growing RNA chain. Note that the enzyme can remove either correct or incorrect bases in this manner, but spends longer hovering over mismatches than matches, and thus removes the former more frequently. In the second proofreading mechanism, called **hydrolytic editing**, the polymerase backtracks by one or more nucleotides (see Fig. 13-11d) and cleaves the RNA product, removing the error-containing sequence.

Hydrolytic editing is stimulated by Gre factors, which both enhance hydrolytic editing function and serve as elongation stimulating factors; that is, they ensure that polymerase elongates efficiently and help overcome “arrest” at sequences that are difficult to transcribe. This combination of functions is comparable to those imposed on the eukaryotic RNA polymerase II by the transcription factor TFIIS (see later in this chapter and Fig. 13-22). Another group of proteins—the Nus proteins—joins polymerase in the elongation phase and promotes, in still rather undefined ways, the processes of elongation and termination (for examples of regulation during elongation, see Chapter 18). One of the bacterial Nus proteins—NusG—is highly conserved in Archaea and Eukaryotes as well (where it is called Spt5; see later discussion).

### RNA Polymerase Can Become Arrested and Need Removing

Under certain circumstances, an elongating RNA polymerase can become arrested and cease transcribing. One common cause of arrest is a damaged DNA strand. The consequences of arrest can be catastrophic if the gene being transcribed is essential as no product will be made by the arrested polymerase, and that same enzyme will cause a roadblock to other polymerases attempting to transcribe the same gene.

To deal with this situation, the cell has machinery that removes the arrested polymerase and at the same time recruits repair enzymes (in particular, the endonuclease Uvr(A)BC); the repair that follows is called transcription-coupled repair, which we discussed in Chapter 10. Both polymerase removal and repair enzyme recruitment are performed by a single protein called TRCF.

TRCF has an ATPase activity. It binds double-stranded DNA upstream of the polymerase and uses the ATPase motor to translocate along the DNA until it encounters the stalled RNA polymerase. The collision pushes polymerase forward, either allowing it to restart elongation or, more often, causing dissociation of the ternary complex of RNA polymerase, template DNA, and RNA transcript. This terminates transcription by that enzyme, but it makes way for repair enzymes and for another RNA polymerase.

### Transcription Is Terminated by Signals within the RNA Sequence

We have already seen one way in which transcription can be terminated. When RNA polymerase arrests during elongation, it can be knocked off DNA by the action of the translocator TRCF (discussed above). This termination is triggered by damaged DNA or by other unanticipated hindrances. But termination is a normal and important function at the ends of genes. There, sequences called **terminators** trigger the elongating polymerase to dissociate from the DNA and release the RNA chain it has made. In bacteria, terminators come in two types: **Rho-dependent** and **Rho-independent**. The first, as its name suggests, requires a protein called Rho to induce termination. The second causes termination without the involvement of other factors. We deal with each kind of terminator in turn.

Rho-dependent terminators have rather ill-defined RNA elements called **rut** sites (discussed later), and for them to work requires the action of the Rho factor. Rho, which is a ring-shaped protein with six identical subunits, binds to single-stranded RNA as it exits the polymerase (Fig. 13-12). The protein also has an ATPase activity, and once attached to the transcript, Rho uses the energy derived from ATP hydrolysis to induce termination. The precise mechanism of termination remains to be determined, and models include the following: Rho pushes polymerase forward relative to the DNA and RNA, resulting in termination in a manner analogous to termination by

**FIGURE 13-12** The Rho transcription termination factor. The crystal structure of the Rho termination factor is shown in a top-down view. It consists of a hexamer of Rho protein, each monomer shown in a different color. The six monomers form an open ring. The ring is not flat; the sixth subunit is further down in the plane of the page than the first. The gap between the two subunits is 12 Å, and the helical pitch between them is 45 Å. The RNA transcript on which Rho acts (not shown) is believed to bind along the bottom of each subunit and then thread through the middle of the ring. (From Skordalakes E. and Berger J.M. 2003. *Cell* 114: 135.) Image prepared with MolScript, BobScript, and Raster3D.



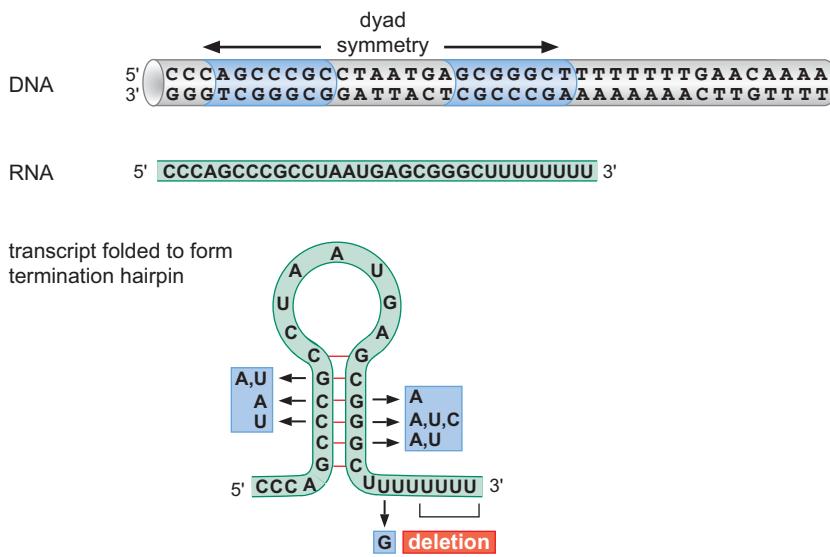
TRCF (described above); Rho pulls RNA out of the polymerase, resulting in termination; or Rho induces a conformational change in polymerase, causing the enzyme to terminate. Most recent experiments suggest that the last of these is at least an important part of the story and that the conformational change causes the elongating complex to stall, with dissociation following more slowly.

Recent studies also have suggested that Rho binds to RNA polymerase throughout the transcription cycle. Thus, Rho doesn't reach polymerase by translocating along a nascent, rut-containing transcript, but, rather, it binds polymerase early in transcription and then at some point also binds the RNA transcript being exuded from that elongating enzyme. The role of translocation by Rho is thus perhaps to tighten the resulting RNA loop, and when sufficiently tight, polymerase elongation ceases.

How is Rho directed to work on particular RNA transcripts? First, there is some specificity in the sites it binds (the rut sites for Rho utilization, mentioned above). Optimally, these sites consist of stretches of ~40 nucleotides that do not fold into a secondary structure (i.e., they remain largely single-stranded); they are also rich in C residues.

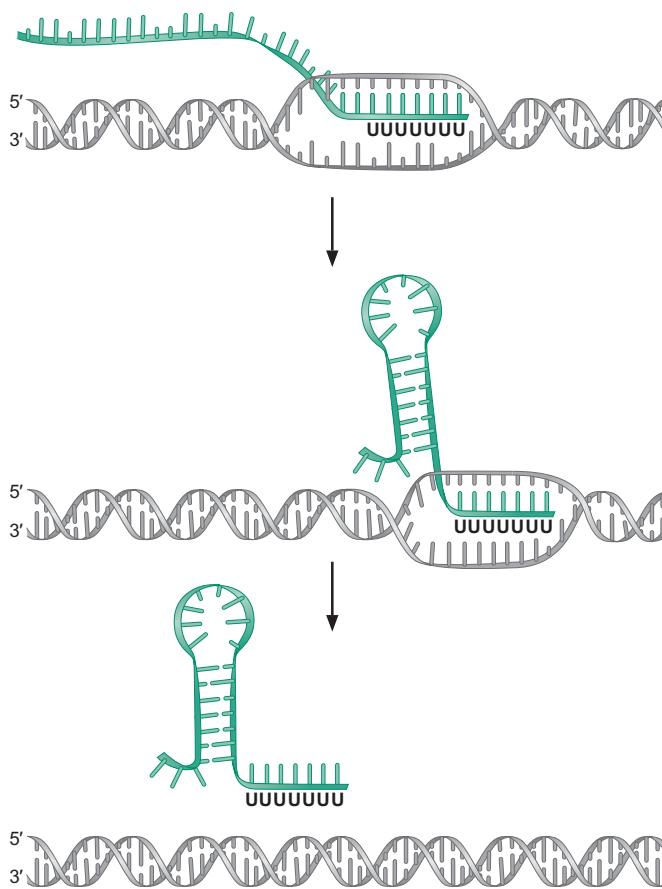
The second level of specificity is that Rho fails to bind any transcript that is being translated (i.e., a transcript bound by ribosomes). In bacteria, transcription and translation are tightly coupled—translation initiates on growing RNA transcripts as soon as they start exiting polymerase, while they are still being synthesized. Thus, Rho typically terminates only those transcripts still being transcribed beyond the end of a gene or operon.

Rho-independent terminators, also called **intrinsic terminators** because they need no other factors to work, consist of two sequence elements: a short inverted repeat (of ~20 nucleotides) followed by a stretch of about eight A:T base pairs (Fig. 13-13). These elements do not affect the polymerase until they have been transcribed—that is, they function in the RNA rather than in the DNA. When polymerase transcribes an inverted repeat sequence, the resulting RNA can form a stem-loop structure (often called a “hairpin”) by base-pairing with itself (see Chapter 5). Formation of the hairpin causes termination by disrupting the elongation complex. As with Rho-dependent termination, the mechanism remains to be determined, and current models are much the same as those proposed for Rho. That is, the hairpin induces termination by either pushing polymerase forward relative to the DNA and RNA, wresting the transcript from polymerase, or inducing a conformational change in polymerase.



**FIGURE 13-13 Sequence of a Rho-independent terminator.** At the top is the sequence, in the DNA, of the terminator. Below is shown the sequence of the RNA, and the bottom image shows the structure of the terminator hairpin. The terminator in question is from the *trp* attenuator, discussed in Chapter 18. The boxes show mutations isolated in the sequence that disrupt the terminator. (Adapted from Yanofsky C. 1981. *Nature* **289**: 751–758.)

The hairpin works as an efficient terminator only when it is followed by a stretch of A:U base pairs, as we have described. This is because, under those circumstances, at the time the hairpin forms, the growing RNA chain will be held on the template at the active site by only A:U base pairs. Because A:U base pairs are the weakest of all base pairs (weaker even than A:T base pairs), they are more easily disrupted by the effects of the stem-loop on the transcribing polymerase, and thus the RNA will more readily dissociate (Fig. 13-14).



**FIGURE 13-14 Transcription termination.** A model for how the Rho-independent terminator might work. (Top) The hairpin forms in the RNA (Fig. 13-11) as soon as that region has been transcribed by polymerase (the enzyme is not shown here). (Middle) That RNA structure disrupts polymerase just as the enzyme is transcribing the AT-rich stretch of DNA downstream. (Bottom) Exactly how the hairpin disrupts the transcribing polymerase is not clear (see text for alternative models), but the weak interactions between the transcript and the template DNA (Us in the transcript and As in the template) appear to make release of that transcript easier. (Adapted from Platt T. 1981. *Cell* **24**: 10–23.)

## TRANSCRIPTION IN EUKARYOTES

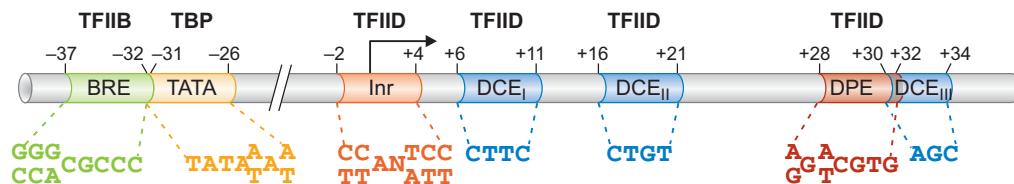
As we have already discussed, transcription in eukaryotes is undertaken by polymerases closely related to the RNA polymerases found in prokaryotes. This is hardly surprising as the process of transcription itself is identical in the two cases. There are, however, differences in the machinery used in each case—one of which we have already seen: whereas bacteria have only one RNA polymerase, all eukaryotes have at least three different ones (Pol I, II, and III; and plants also have a Pol IV and a Pol V). In addition, whereas bacteria require only one additional initiation factor ( $\sigma$ ), several initiation factors are required for efficient and promoter-specific initiation in eukaryotes. These are called the **general transcription factors (GTFs)**.

In vitro, the general transcription factors are all that are required, together with Pol II, to initiate transcription on a DNA template (without histones). In vivo, however, the DNA template in eukaryotic cells is incorporated into nucleosomes, as we have seen in Chapter 8. Under these circumstances, the general transcription factors are not alone sufficient to bind promoter sequences and elicit significant expression. Rather, additional factors are required, including DNA-binding regulatory proteins, the so-called Mediator complex, and often chromatin-modifying enzymes.

We first consider the basic mechanism by which Pol II and the general transcription factors assemble at a promoter to initiate transcription in vitro. We then consider the roles of the additional components required to promote transcription in vivo.

### RNA Polymerase II Core Promoters Are Made Up of Combinations of Different Classes of Sequence Element

The eukaryotic **core promoter** refers to the minimal set of sequence elements required for accurate transcription initiation by the Pol II machinery, as measured in vitro. A core promoter is typically  $\sim 40$ – $60$  nucleotides long, extending either upstream or downstream from the transcription start site. Figure 13-15 shows the location, relative to the transcription start site, of elements found in Pol II core promoters. These are the TFIIB recognition element (BRE), the TATA element (or box), the initiator (Inr), and the downstream promoter elements (known as DPE, DCE, and MTE). Typically, a promoter includes some subset of these elements. Thus, for example, promoters typically have either a TATA element or a DPE element, not both. Often, a TATA-containing promoter also contains a DCE. The Inr is the



**FIGURE 13-15** Pol II core promoter. The figure shows the positions of various DNA elements relative to the transcription start site (indicated by the arrow above the DNA). These elements, described in the text, are as follows: (BRE) TFIIB recognition element; (TATA) TATA box; (Inr) initiator element; (DPE) downstream promoter element; and (DCE) downstream core element. Another element, MTE (motif ten element), described in the text, is not shown in this figure but is located just upstream of the DPE. Also shown are the consensus sequences for each element (determined in the same way as described for the bacterial promoter elements; see Box 13-1) and (above) the name of the general transcription factor that recognizes each element.

most common element, found in combination with both TATA and DPEs. The consensus sequence for each element and the general transcription factor that binds it are also shown, and these features are described in more detail in coming sections.

Beyond—and typically upstream of—the core promoter, there are other sequence elements required for accurate and efficient transcription *in vivo*. Together, these elements constitute the **regulatory sequences** and can be grouped into various categories, reflecting their location, and the organism in question, as much as their function. These elements include promoter proximal elements, upstream activator sequences (UASs), enhancers, and a series of other elements called silencers, boundary elements, and insulators. All of these DNA elements bind regulatory proteins (activators and repressors), which help or hinder transcription from the core promoter (these are the subject of Chapter 19). Some of these regulatory sequences can be located many tens or even hundreds of kilobases from the core promoters on which they act.

### RNA Polymerase II Forms a Preinitiation Complex with General Transcription Factors at the Promoter

The general transcription factors collectively perform the functions performed by  $\sigma$  in bacterial transcription. Thus, the general transcription factors help polymerase bind to the promoter and melt the DNA (comparable to the transition from the closed to the open complex in the bacterial case). They also help polymerase escape from the promoter and embark on the elongation phase. The complete set of general transcription factors and polymerase, bound together at the promoter and poised for initiation, is called the **preinitiation complex**.

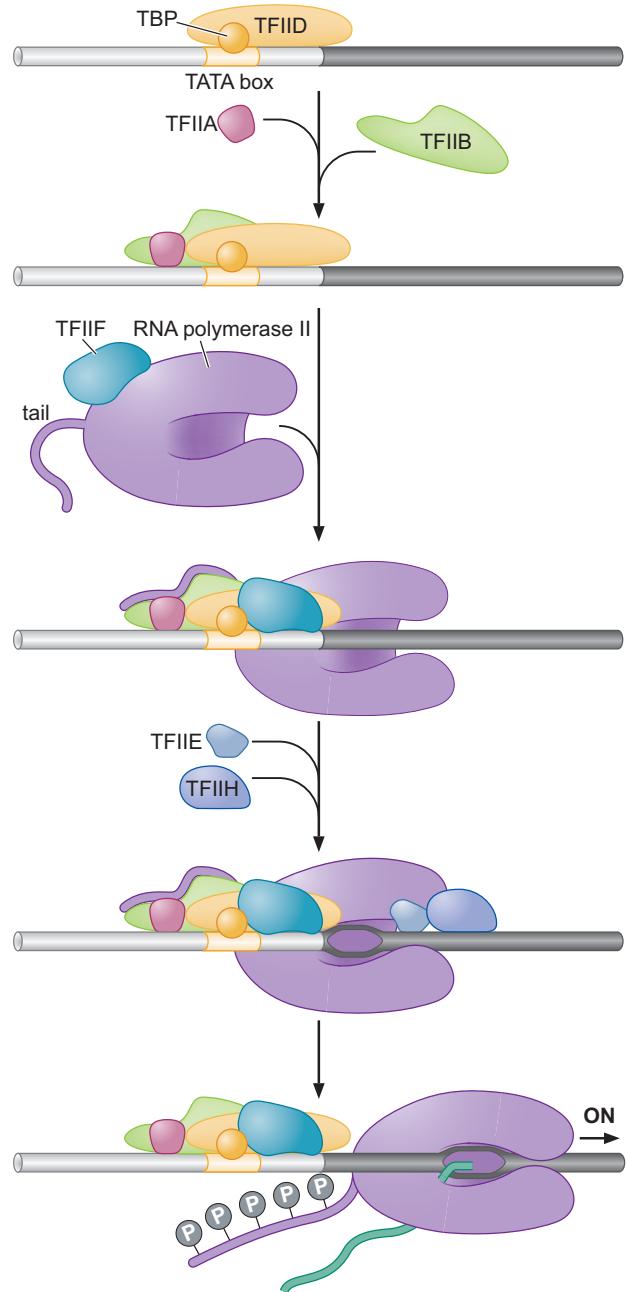
As we described above (and in Fig. 13-15), many Pol II promoters contain a so-called TATA element (some 30 bp upstream of the transcription start site). This is where preinitiation complex formation begins. The TATA element is recognized by the general transcription factor called **TFIID**. (The nomenclature “TFII” denotes a transcription factor for Pol II, with individual factors distinguished as A, B, and so on.) Like many of the general transcription factors, TFIID is, in fact, a multi-subunit complex. The component of TFIID that binds to the TATA DNA sequence is called **TBP** (TATA-binding protein). The other subunits in this complex are called **TAFs**, for TBP-associated factors. Some TAFs recognize other core promoter elements such as the Inr, DPE, and DCE, although the strongest binding is between TBP and TATA. Thus, TFIID is a critical factor in promoter recognition and preinitiation complex establishment.

Upon binding DNA, TBP extensively distorts the TATA sequence (we discuss this event in more detail later). The resulting TBP–DNA complex provides a platform to recruit other general transcription factors and polymerase itself to the promoter. *In vitro*, these proteins assemble at the promoter in the following order (Fig. 13-16): TFIIA, TFIIB, TFIIIF together with polymerase, and then TFIIIE and TFIIH. Formation of the preinitiation complex containing these components is followed by promoter melting. In contrast to the situation in bacteria, promoter melting in eukaryotes requires hydrolysis of ATP and is mediated by TFIIH.

### Promoter Escape Requires Phosphorylation of the Polymerase “Tail”

Just as we have seen in the bacterial case, there now follows a period of abortive initiation before the polymerase escapes the promoter and enters

**FIGURE 13-16** Transcription initiation by RNA Pol II. The stepwise assembly of the Pol II preinitiation complex is shown here and described in detail in the text. Once assembled at the promoter, Pol II leaves the preinitiation complex upon addition of the nucleotide precursors required for RNA synthesis and after phosphorylation of serine residues within the enzyme's "tail." The tail contains multiple repeats of the heptapeptide sequence: Tyr-Ser-Pro-Thr-Ser-Pro-Ser (see Fig. 13-21).



the elongation phase. Recall that during abortive initiation, the polymerase synthesizes a series of short transcripts. In eukaryotes, promoter escape involves two steps not seen in bacteria: one is ATP hydrolysis (in addition to the earlier ATP hydrolysis needed for DNA melting), and the other is phosphorylation of the polymerase, as we now describe.

The large subunit of Pol II has a carboxy-terminal domain (CTD), which is referred to as the "tail" (see Fig. 13-16). The CTD contains a series of repeats of the heptapeptide sequence: Tyr-Ser-Pro-Thr-Ser-Pro-Ser. There are 27 of these repeats in the yeast Pol II CTD, 32 in the worm *Caenorhabditis elegans*, 45 in the fly *Drosophila*, and 52 in humans. Indeed, the number of repeats seems to correlate with the complexity of the genome. Each repeat contains sites for phosphorylation by specific kinases, including one that is a subunit of TFIH.

The form of Pol II recruited to the promoter initially contains a largely unphosphorylated tail, but the species found in the elongation complex bears multiple phosphoryl groups on its tail. Addition of these phosphates helps polymerase shed most of the general transcription factors used for initiation, and which the enzyme leaves behind as it escapes the promoter.

As we will see, regulating the phosphorylation state of the CTD of Pol II controls subsequent steps—elongation and even processing of the RNA—as well. Indeed, in addition to TFIIH, several other kinases have been identified that act on the CTD, as well as a number of phosphatases that remove the phosphates added by those kinases.

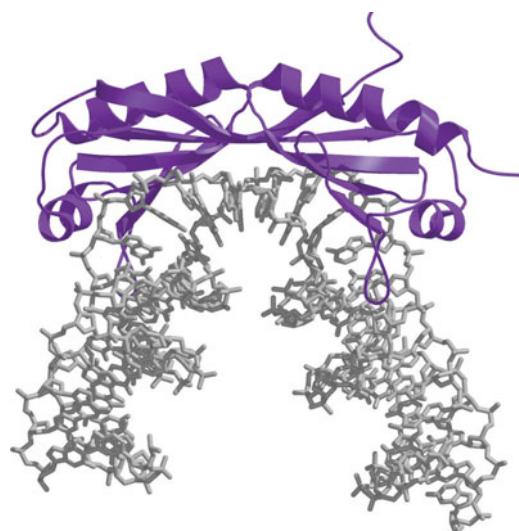
### TBP Binds to and Distorts DNA Using a $\beta$ Sheet Inserted into the Minor Groove

TBP uses an extensive region of  $\beta$  sheet to recognize the minor groove of the TATA element (Fig. 13-17). This is unusual. More typically, proteins recognize DNA using  $\alpha$  helices inserted into the major groove of DNA, as we have seen in Chapter 6 and also for the  $\sigma$  factor in this chapter. The reason for TBP's unorthodox recognition mechanism is linked to the need for that protein to distort the local DNA structure. But this mode of recognition raises a problem: how is specificity achieved?

We have seen in Chapter 6 that, compared with the major groove, the minor groove of DNA is less rich in the chemical information that would enable base pairs to be distinguished. Instead, to select the TATA sequence, TBP relies on the ability of that sequence to undergo a specific structural distortion, as we now describe.

When it binds DNA, TBP causes the minor groove to be widened to an almost flat conformation; it also bends the DNA by an angle of  $\sim 80^\circ$ . The interaction between TBP and DNA involves only a limited number of hydrogen bonds between the protein and the edges of the base pairs in the minor groove. Instead, much of the specificity is imposed by two pairs of phenylalanine side chains that intercalate between the base pairs at either end of the recognition sequence and drive the strong bend in the DNA.

A:T base pairs are thus favored because they are more readily distorted to allow the initial opening of the minor groove. There are also extensive interactions between the phosphate backbone and basic residues in the  $\beta$  sheet, adding to the overall binding energy of the interaction.



**FIGURE 13-17** TBP–DNA complex. TBP (purple) is complexed with the DNA TATA sequence (gray) found at the start of many Pol II genes. The details of this interaction are described in the text. (From Nikolov D.B. et al. 1995. *Nature* 377: 119.) Image prepared with MolScript, BobScript, and Raster3D. Extended DNA on either side of image was modeled by Leemor Joshua-Tor.

**TABLE 13-2** The General Transcription Factors of RNA Polymerase II

GTGs	Number of Subunits
TBP	1
TFIIA	2
TFIIB	1
TFIIE	2
TFIIF	3
TFIIH	10
TAFs	11

The numbers shown are for yeast but are similar for other eukaryotes, including humans. There are some differences, however—for example, human TFIIF has only two subunits, and its TFIIA has three.

### The Other General Transcription Factors Also Have Specific Roles in Initiation

We do not know in detail the functions of all of the other general transcription factors. As we have noted, some of these factors are in fact complexes made up of two or more subunits (shown in Table 13-2). Later we shall comment on a few structural and functional characteristics.

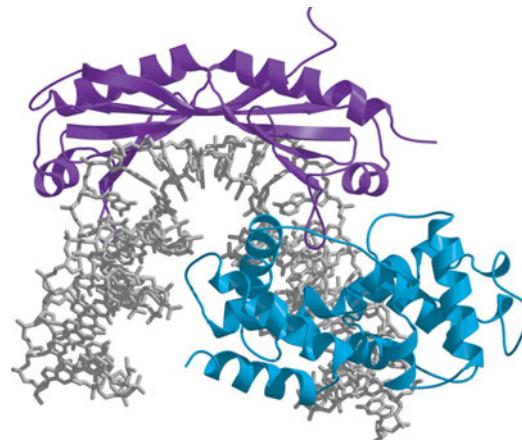
**TAFs** TBP is associated with about 10 TAFs. Two of the TAFs bind DNA elements at the promoter, for example, the initiator element (Inr) and the downstream promoter elements (see Fig. 13-15). Several of the TAFs have structural homology with histone proteins, and it has been proposed that they might bind DNA in a similar manner, although evidence for such a form of DNA binding has not been obtained. For example, TAF42 and TAF62 from *Drosophila* have been shown to form a structure similar to that of the H3·H4 tetramer (see Chapter 8). These histone-like TAFs are found not only in the TFIID complex, but also associated with some histone modification enzymes, such as the yeast SAGA complex (see Chapter 8, Table 8-7).

Another TAF appears to regulate the binding of TBP to DNA. It does this using an inhibitory flap that binds to the DNA-binding surface of TBP, another example of molecular mimicry. This flap must be displaced for TBP to bind TATA.

**TFIIB** This protein, a single polypeptide chain, enters the preinitiation complex after TBP (Fig. 13-16). The crystal structure of the ternary complex of TFIIB–TBP–DNA shows specific TFIIB–TBP and TFIIB–DNA contacts (Fig. 13-18). These include base-specific interactions with the major groove upstream (to the BRE) (see Fig. 13-15) and the minor groove downstream of the TATA element. The asymmetric binding of TFIIB to the TBP–TATA complex accounts for the asymmetry in the rest of the assembly of the preinitiation complex and the unidirectional transcription that results. TFIIB also contacts Pol II in the preinitiation complex. Thus, this protein appears to bridge the TATA-bound TBP and polymerase. Structural studies suggest that segments of TFIIB insert into the RNA-exit channel and active center cleft of Pol II in a manner analogous to the  $\sigma$  region 3/4 linker in the bacterial case. These regions of TFIIB (called the **linker** and **reader**) aid in open complex formation, perhaps by stabilizing the melted DNA until the RNA:DNA hybrid takes over that role.

**TFIIF** This two-subunit (in humans) factor associates with Pol II and is recruited to the promoter together with that enzyme (and other factors).

**FIGURE 13-18** TFIIB–TBP–promoter complex. This structure shows the TBP protein bound to the TATA sequence, just as in the previous figure. Here, the general transcription factor TFIIB (turquoise) has been added. This tripartite complex forms the platform to which other general transcription factors, and Pol II itself, are recruited during preinitiation complex assembly. (From Nikолов D.B. et al. 1995. *Nature* 377: 119.) Image prepared with MolScript, BobScript, and Raster3D. Extended DNA on either side of image was modeled by Leemor Joshua-Tor.



Binding of Pol II–TFIIF stabilizes the DNA–TBP–TFIIB complex and is required before TFIIIE and TFIIH are recruited to the preinitiation complex (Fig. 13-16). In yeast, this factor includes a third subunit (as shown in Table 13-2), but the function of the third subunit is not known.

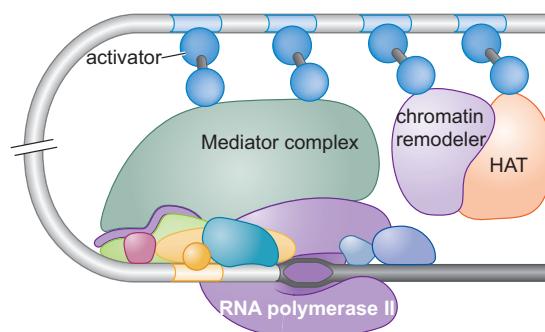
**TFIIE and TFIIH** TFIIE, which, like TFIIF, consists of two subunits, binds next and has roles in the recruitment and regulation of TFIIH. TFIIH controls the ATP-dependent transition of the preinitiation complex to the open complex. It is also the largest and most complex of the general transcription factors having 10 subunits and a molecular mass comparable to that of the polymerase itself! Within TFIIH are two subunits that function as ATPases and another that is a protein kinase, with roles in promoter melting and escape, as described above. Together with other factors, the ATPase subunits are also involved in nucleotide excision repair (see Chapter 10).

How does TFIIH mediate promoter melting? We saw in the bacterial case that melting of the –10 element in the promoter is mediated by bases on the non-coding DNA strand being flipped out and bound within pockets in the  $\sigma$  subunit. This requires no ATP hydrolysis and is driven simply by binding reactions that favor the melted conformation (see Fig. 13-8). In eukaryotes, things are more complicated. It is now believed that a subunit of TFIIH acts as an ATP-driven translocator of double-stranded DNA. This subunit binds to DNA downstream from polymerase (as was shown in Fig. 13-16) and feeds double-stranded DNA, with a right-handed threading, into the cleft of the polymerase. This action drives the melting of the DNA because the upstream promoter DNA is held in a fixed position by TFIID and the rest of the GTFs.

### In Vivo, Transcription Initiation Requires Additional Proteins, Including the Mediator Complex

Thus far, we have described what is needed for Pol II to initiate transcription from a naked DNA template *in vitro*. But we have already noted that high, regulated levels of transcription *in vivo* require, additionally, transcriptional regulatory proteins, the Mediator complex, and nucleosome-modifying enzymes (which are themselves often parts of large protein complexes) (Fig. 13-19). (For characteristics of various modifying complexes, see Chapter 8, Table 8-7.)

One reason for these additional requirements is that the DNA template *in vivo* is packaged into chromatin, as we discussed in Chapter 8. This condition complicates binding to the promoter of polymerase and its associated factors. Transcriptional regulatory proteins called **activators** help recruit polymerase to the promoter, stabilizing its binding there. This recruitment is mediated through interactions between DNA-bound activators, chromatin-modifying and -remodeling factors, and parts of the



**FIGURE 13-19** Assembly of the pre-initiation complex in the presence of Mediator, nucleosome modifiers and remodelers, and transcriptional activators. In addition to the general transcription factors shown in Figure 13-16, transcriptional activators bound to sites near the gene recruit nucleosome-modifying and -remodeling complexes and the Mediator complex, which together help form the preinitiation complex.

transcription machinery. One such interaction is with the Mediator complex (hence, its name). Mediator is associated with the basic transcription machinery, most likely touching the CTD “tail” of the large polymerase subunit through one surface, while presenting other surfaces for interaction with DNA-bound activators. This explains the need for Mediator to achieve significant transcription *in vivo*.

Despite this central role in transcriptional activation, deletion of individual subunits of Mediator often leads to loss of expression of only a small subset of genes, different for each subunit (it is made up of many subunits). This result likely reflects the fact that different activators are believed to interact with different Mediator subunits to bring polymerase to different genes. In addition, Mediator aids initiation by regulating the CTD kinase in TFIH.

The need for nucleosome modifiers and remodelers also differs at different promoters or even at the same promoter under different circumstances. When and where required, these complexes are also typically recruited by the DNA-bound activators, or sometimes by regulatory RNAs.

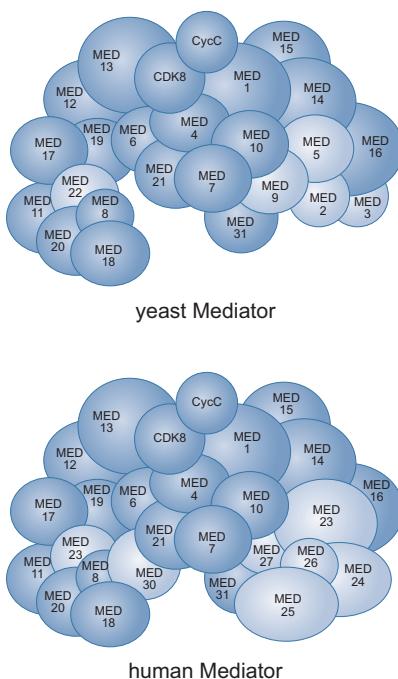
We discuss the role of Mediator and modifiers in stimulating transcription in Chapter 19. We now consider some of the structural and functional properties of Mediator.

### Mediator Consists of Many Subunits, Some Conserved from Yeast to Human

As shown in Figure 13-20, the yeast and human Mediators each include more than 20 subunits, of which seven show significant sequence homology between the two organisms. (The names of the subunits were initially different in each case, reflecting the experimental approaches that led to their identification, but subsequently a convention was established so that equivalent subunits in different organisms take the same name. It is these that are given in Fig. 13-20.) Very few of these subunits have any identified function. Only one (Srb4/Med17) is essential for transcription of essentially all Pol II genes *in vivo*. Low-resolution structural comparisons suggest that both Mediators have a similar shape, and both are very large, even bigger than RNA polymerase itself.

The Mediator from both yeast and humans is organized in modules, each containing a subset of the subunits shown in Fig. 13-20. These modules—called head, middle (or arm), and tail—can be dissociated from one another under certain conditions *in vitro*. This observation, together with the fact that human Mediator varies in its composition (and size) depending on how it is isolated, has led to the idea that there are various forms of Mediator (particularly in metazoans), each containing subsets of Mediator subunits. Furthermore, it has been argued that the different forms are involved in regulating different subsets of genes or responding to different groups of regulators (activators and repressors). It is equally possible, however, that the variations seen in subunit composition are artifacts, simply reflecting different methods of isolation.

Attempts to ascertain the structure of Mediator have recently benefited from the solution of a crystal structure of part of the complex—the head module of yeast Mediator. This module contains seven subunits (Med17/Srb4, Med11, Med22/Srb6, Med6, Med8, Med18/Srb5, and Med20/Srb2) and forms a three-domain structure that binds the transcription complex in such a way as to juxtapose TFIH and the CTD tail of RNA polymerase, promoting phosphorylation of the latter by the former. Phosphorylation of serine residues within the tail is required for initiation and promoter escape, as we discuss later. And, in particular, phosphorylation of serine 5 by TFIH itself leads to Mediator dissociation from polymerase during that process.

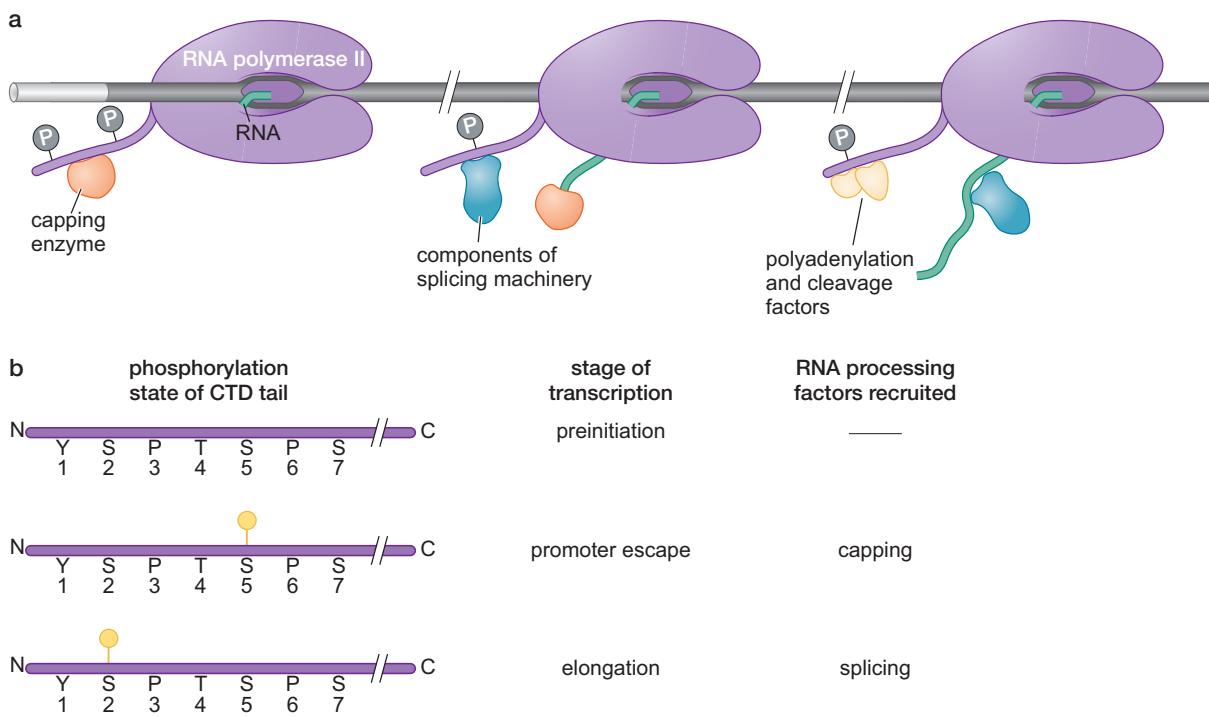


**FIGURE 13-20** Comparison of the yeast and human Mediators. The majority of the subunits are found in both cases, but differences are indicated by paler shading. (Yeast Mediator: modified from Guglielmi B. et al. 2004. *Nucleic Acids Res.* **32**: 5379–5391, Fig. 8B; human Mediator: modified, with permission, from Malik S. and Roeder R.G. 2005. *Trends Biochem. Sci.* **30**: 256–263, Fig. 1a. © Elsevier.)

### A New Set of Factors Stimulates Pol II Elongation and RNA Proofreading

Once polymerase has escaped the promoter and initiated transcription, it shifts into the elongation phase, as we have discussed. This transition involves the Pol II enzyme shedding most of its initiation factors—for example, the general transcription factors and Mediator. In their place, another set of factors is recruited. Some of these (such as TFIIS and SPT5) are **elongation factors** (i.e., factors that stimulate elongation). Others are required for RNA processing. The enzymes involved in RNA processing (described in detail later) are, like several of the initiation factors we have discussed, recruited to the carboxy-terminal (CTD) tail of the large subunit of Pol II (Fig. 13-21). In this case, however, the factors favor the phosphorylated form of the CTD. Thus, phosphorylation of the CTD leads to an exchange of initiation factors for those factors required for elongation and RNA processing.

As is evident from the crystal structure of yeast Pol II, the polymerase CTD lies directly adjacent to the channel through which the newly synthesized RNA exits the enzyme. The CTD tail is also very long (it could potentially extend ~800 Å from the body of the enzyme—that is, about seven times the length of the rest of the enzyme). Together, these features allow the



**FIGURE 13-21** RNA processing enzymes are recruited by the CTD tail of polymerase. (a) Various factors involved in RNA processing recruited by the CTD tail of polymerase. Different factors are recruited depending on the phosphorylation state of the tail. Those factors are then transferred to the RNA as they are needed (see next section in text). (b) A schematic of the tail, with the sequence of one copy of the heptapeptide repeat shown in the top line. The positions of serine residues that get phosphorylated are indicated in lines 2 and 3. Phosphorylation of serine at position 5 is seen upon promoter escape and is associated with recruitment of capping factors, whereas phosphorylation of serine at position 2 is seen during elongation and is associated with recruitment of splicing factors. Recruitment of factors involved in elongation of transcription and in RNA processing overlaps. Thus, elongation factor hSPT5 is recruited to the tail phosphorylated on Ser-5.

tail to bind several components of the elongation and processing machinery and deliver them to the emerging RNA.

Various proteins are thought to stimulate elongation by Pol II. One of these, the kinase P-TEFb, is recruited to polymerase by transcriptional activators. Once bound to Pol II, this protein phosphorylates the serine residue at position 2 of the CTD repeats. That phosphorylation event correlates with elongation (Fig. 13-21). In addition, P-TEFb phosphorylates and thereby activates another protein, called SPT5, itself an elongation factor. Finally, TAT-SF1, yet another elongation factor, is recruited by P-TEFb. Thus, P-TEFb stimulates elongation in three separate ways.

SPT5 is comparable to the bacterial elongation factor NusG that we encountered above. Indeed, this is the only universally conserved transcription factor across all three kingdoms of life—from bacteria, through Archaea, to eukaryotes. NusG/SPT5 factors bind to their respective RNA polymerases at the tip of the clamp, overlapping the region contacted by  $\sigma$  region 4 (in bacteria) and TFIIB (in eukaryotes). This overlapping—and presumably mutually exclusive—binding raises the interesting possibility that displacing initiation factors may be part of the function of these elongation regulators. This also suggests that regulating the rate of elongation is an ancient mechanism of regulating gene expression. As we discuss in Chapter 19, there are some promoters in higher eukaryotes where the preinitiation complex is recruited effectively, but polymerase remains paused just after initiating transcription. Such promoters seem to be associated with genes poised to be expressed either rapidly or in a highly coordinated fashion, and their expression is regulated through recruitment by specific activators of the P-TEFb kinase, which then releases them from their pause (see Chapter 19).

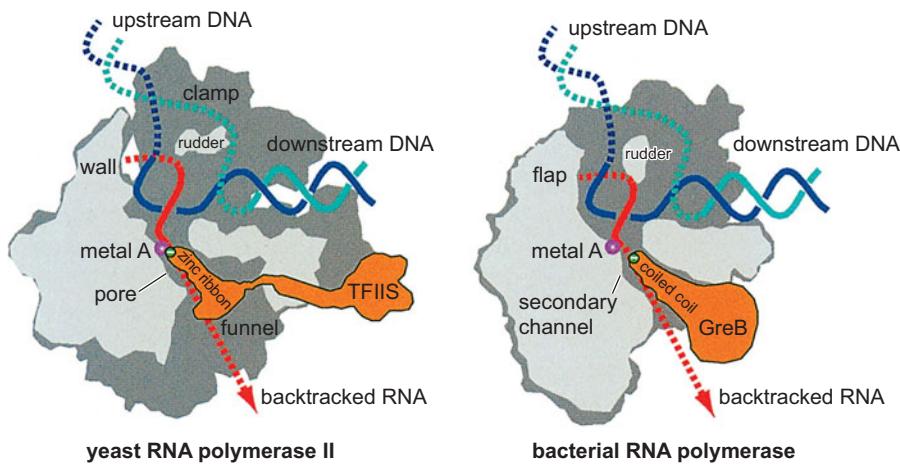
Yet another class of elongation factor is the so-called ELL family. These also bind to elongating polymerase and suppress transient pausing by the enzyme; such pausing otherwise occurs at many sites along the DNA. The first human ELL protein was originally identified as the product of a gene that undergoes translocations in acute myeloid leukemia (see Box 19-3).

Another factor that does not affect initiation, but stimulates elongation, is TFIIS. This factor, like ELL, stimulates the overall rate of elongation by limiting the length of time that polymerase pauses when it encounters sequences that would otherwise tend to slow the enzyme's progress. It is a feature of polymerase that it does not transcribe through all sequences at a constant rate. Rather, it pauses periodically, sometimes for rather long periods, before resuming transcription. In the presence of TFIIS, the length of time that polymerase pauses at any given site is reduced.

TFIIS also contributes to proofreading by polymerase. We saw earlier in the chapter how polymerases are able, inefficiently, to remove misincorporated bases using the active site of the enzyme to perform the reverse reaction to nucleotide incorporation. In addition, TFIIS stimulates an inherent RNase activity in polymerase (not part of the active site), allowing an alternative approach to removing misincorporated bases through local limited RNA degradation. This feature is comparable to the hydrolytic editing in the bacterial case stimulated by the Gre factors we discussed there. Figure 13-22 shows how TFIIS and GreB, although structurally unrelated (and unrelated in sequence, too), nevertheless interact with the yeast and bacterial polymerases, respectively, in comparable ways, to stimulate the same reactions.

### Elongating RNA Polymerase Must Deal with Histones in Its Path

As with initiation of transcription, elongation also takes place in the presence of histones, because the DNA template is incorporated into nucleosomes. How does RNA polymerase transcribe through these potential barriers?



**FIGURE 13-22** TFIIS and GreB act in analogous ways. Cutaway views of the major features of the complexes of arrested RNA polymerase II and TFIIS (left) and bacterial RNA polymerase and GreB (right). TFIIS (orange) is inserted into the RNA polymerase II core, and GreB (orange) is inserted into the bacterial RNA polymerase channel. In each case, the primary catalytic magnesium ion is designated as Metal A (pink), and the positions of the two conserved acidic residues are indicated (green circles). Thus we see that although the two proteins are so different, they act in essentially the same way. (Dashed arrows) The presumed locations of the backtracked RNAs (see also Fig. 13-11). (Reprinted, with permission, from Conaway R.C. et al. 2003. *Cell* 114: 272–274, Fig. 1. © Elsevier.)

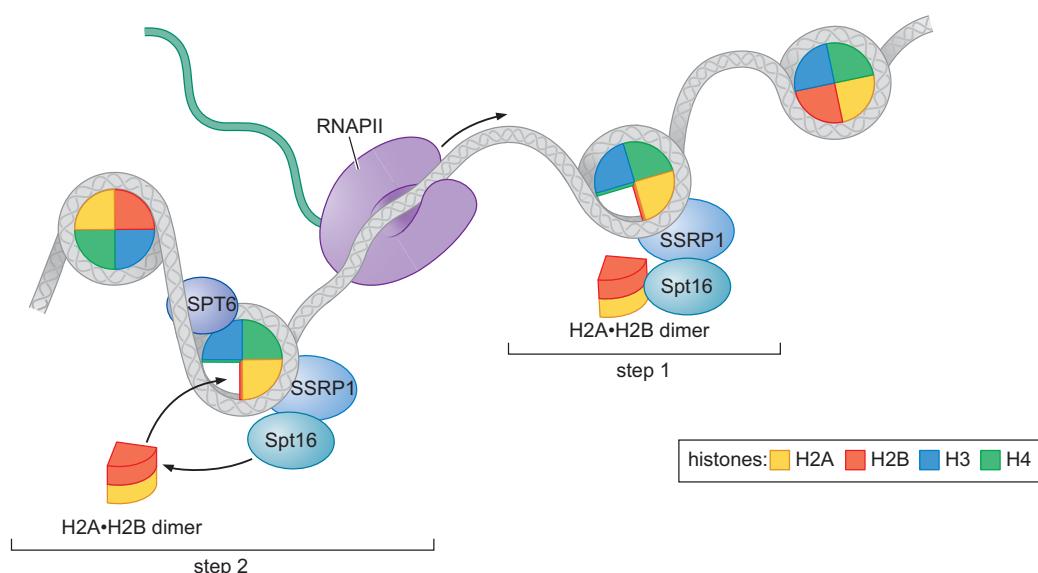
Experiments *in vitro* comparing transcription on naked DNA and on DNA incorporated in chromatin revealed that chromatin greatly impedes transcription. This experimental setup provided the assay for identifying factors that facilitate transcription in the presence of chromatin. In this way, a factor called **FACT** (facilitates chromatin transcription) was identified in human cell extracts. As its name suggests, this factor makes transcription on chromatin templates much more efficient. FACT is a heterodimer of two well-conserved proteins, Spt16 and SSRP1. The yeast homolog of the former had already been linked to chromatin modulation from genetic studies, and a role for FACT in elongation was established through genetic interactions between this complex and known elongation factors, including TFIIS.

How does FACT work? Recall from Chapter 8 that nucleosomes are octamers, made up of H2A, H2B, H3, and H4 histone subunits and DNA (see Chapter 8, Fig. 8-20). These histones are arranged in two modules: the H2A-H2B dimers and the H3-H4 tetramer. Spt16 binds to the former, SSRP1 to the latter. Strikingly, FACT can both dismantle histones, by removing one H2A-H2B dimer, and reassemble them by restoring that dimer.

Thus there evolved a picture of how FACT works during elongation (Fig. 13-23). Ahead of a transcribing RNA polymerase, FACT removes one H2A-H2B dimer. This allows polymerase to pass that nucleosome (*in vitro*, it has been shown that removing H2A-H2B from a template allows transcription). FACT also has histone chaperone activity, which allows it to restore the H2A-H2B dimer to the histone hexamer immediately behind the processing polymerase. In this way, FACT allows polymerase to elongate and at the same time maintains the integrity of the chromatin through which the enzyme is transcribing.

### Elongating Polymerase Is Associated with a New Set of Protein Factors Required for Various Types of RNA Processing

Once transcribed, eukaryotic RNA has to be processed in various ways before being exported from the nucleus where it can be translated. These processing events include capping of the 5' end of the RNA, splicing, and polyadenylation of the 3' end of the RNA. The most complicated of these is splicing—the process whereby non-coding introns are removed from RNA to generate the mature mRNA. The mechanisms and regulation of that process and others, such as RNA editing, are the subject of Chapter 14. Here we consider the other two processes—capping and polyadenylating the transcript.



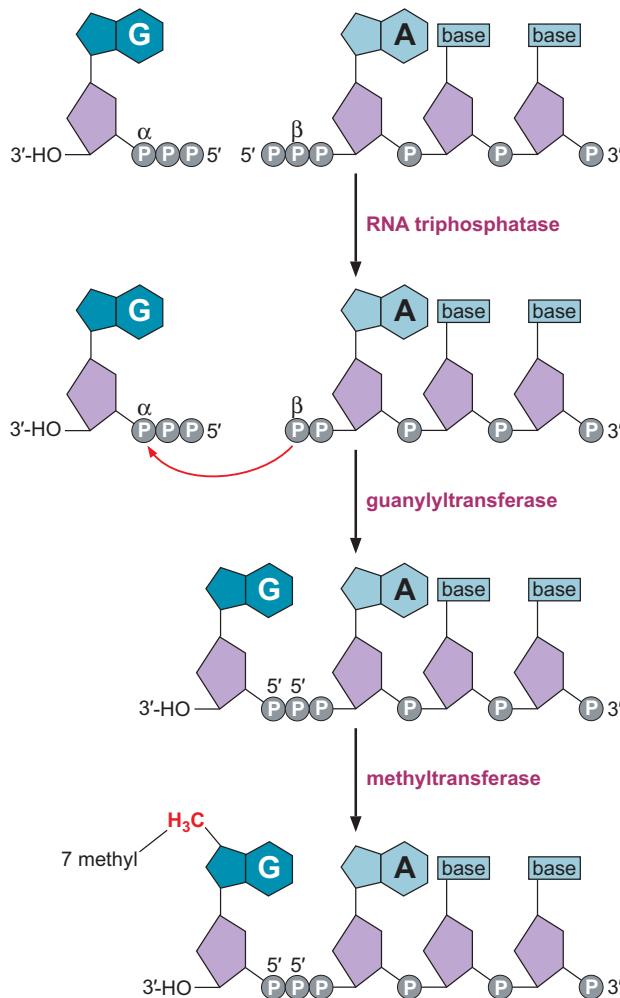
**FIGURE 13-23** A model for FACT-aided elongation through nucleosomes. As described in the text, FACT, shown as the heterodimer of Spt16 and SSRP1, is able to dismantle nucleosomes ahead of the transcribing RNA polymerase (Step 1) and reassemble them behind (Step 2). Specifically, it removes the H2A-H2B dimer. SPT6 binds histone H3 and is believed to aid in nucleosome reassembly. (Adapted, with permission, from Reinberg D. and Sims R. 2006. *J. Biol. Chem.* **281**: 23297–23301, Fig. 2b. © American Society for Biochemistry and Molecular Biology.)

Strikingly, there is an overlap in proteins involved in elongation and those required for RNA processing. In one case, for example, an elongation factor mentioned above (SPT5) also helps to recruit the 5'-capping enzyme to the CTD tail of polymerase (phosphorylated at serine position 5) (Fig. 13-21b). The hSPT5 stimulates the 5'-capping enzyme activity. In another case, elongation factor TAT-SF1 recruits components of the splicing machinery to polymerase with a Ser-2 phosphorylated tail (Fig. 13-21b). Thus, elongation, termination of transcription, and RNA processing are interconnected, presumably to ensure their proper coordination.

The first RNA processing event is **capping**. This involves the addition of a modified guanine base to the 5' end of the RNA. Specifically, it is a methylated guanine, and it is joined to the RNA transcript by an unusual 5'-5' linkage involving three phosphates (this structure is shown in the last step at the bottom of Fig. 13-24).

The 5' cap is created in three enzymatic steps, as detailed in Figure 13-24 and described in detail in the legend. In the first step, a phosphate group is removed from the 5' end of the transcript. Then, in the second step, the GMP moiety is added. In the final step, that nucleotide is modified by the addition of a methyl group. The RNA is capped as soon as it emerges from the RNA-exit channel of polymerase. This happens when the transcription cycle has progressed only as far as the transition from the initiation to elongation phases. After capping, dephosphorylation of Ser-5 within the tail repeats may be responsible for dissociation of the capping machinery, and further phosphorylation (this time of Ser-2 within the tail repeats) causes recruitment of the machinery needed for RNA splicing (Chapter 14) (see Fig. 13-21b).

The final RNA processing event, **Polyadenylation** of the 3' end of the mRNA, is intimately linked with the termination of transcription (Fig. 13-25). Just as with capping and splicing, the polymerase CTD tail is involved in recruiting some of the enzymes necessary for polyadenylation (Fig. 13-21). Once polymerase has reached the end of a gene, it encounters



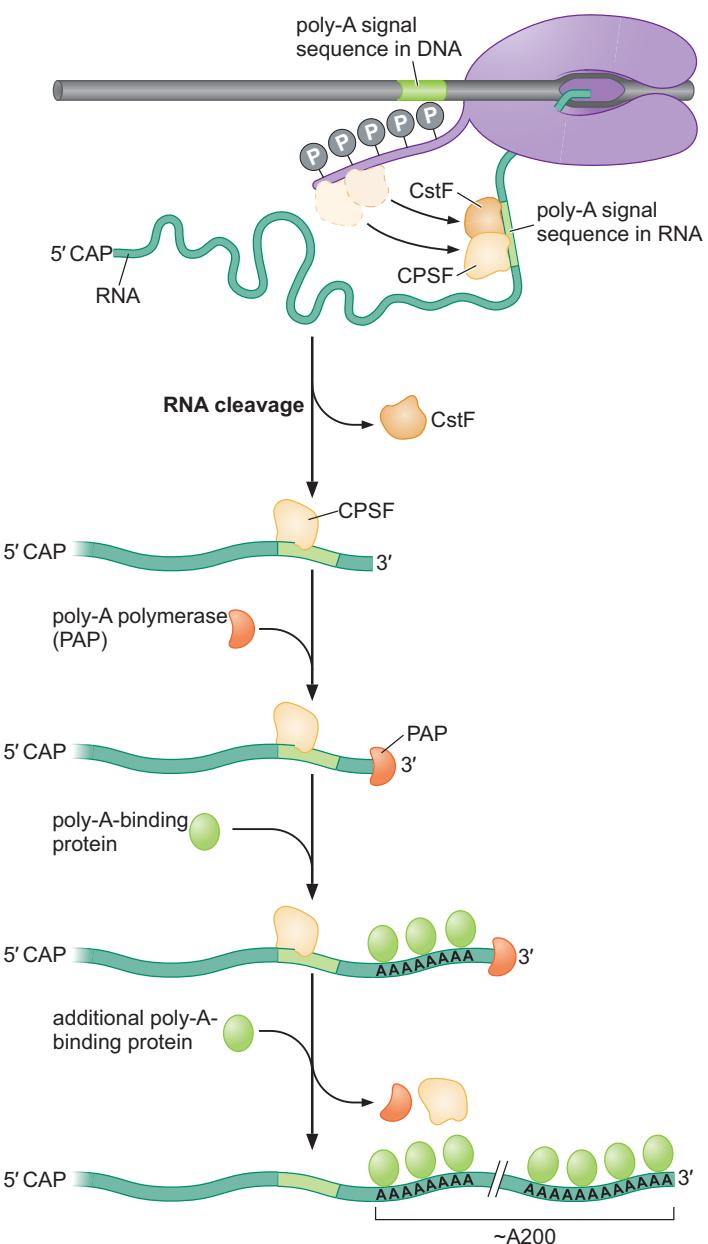
**FIGURE 13-24** The structure and formation of the 5' RNA cap. In the first step, the γ-phosphate at the 5' end of the RNA is removed by an enzyme called RNA triphosphatase (the initiating nucleotide of a transcript initially retains its α-, β-, and γ-phosphates). In the next step, the enzyme guanylyltransferase adds a GMP moiety to the resulting terminal β-phosphate of the 5' end of the RNA. This is a two-step process: first, an enzyme–GMP complex is generated from GTP with release of the β- and γ-phosphates of that GTP, and then the GMP from the enzyme is transferred to the β-phosphate of the 5' end of the RNA. Once this linkage is made, the newly added guanine and the purine at the original 5' end of the mRNA are further modified by the addition of methyl groups by methyltransferase. The resulting 5' cap structure subsequently recruits the ribosome to the mRNA for translation to begin (see Chapter 15).

specific sequences that, after being transcribed into RNA, trigger the transfer of the polyadenylation enzymes to that RNA, leading to four events: cleavage of the message; addition of many adenine residues to its 3' end; degradation of the RNA remaining associated with RNA polymerase by a 5'-to-3' ribonuclease; and, subsequently, termination of transcription. This series of events unfolds as follows.

Two protein complexes are carried by the CTD of polymerase as it approaches the end of the gene: CPSF (cleavage and polyadenylation specificity factor) and CSTF (cleavage stimulation factor). The sequences that, once transcribed into RNA, trigger transfer of these factors to the RNA are called **poly-A signals**, and their operation is shown in Figure 13-25. Once CPSF and CSTF are bound to the RNA, other proteins are recruited as well, leading initially to RNA cleavage and then polyadenylation.

Polyadenylation is mediated by an enzyme called poly-A polymerase, which adds approximately 200 adenines to the RNA's 3' end produced by the cleavage. This enzyme uses ATP as a precursor and adds the nucleotides using the same chemistry as RNA polymerase. But it does so without a template. Thus, the long tail of As is found in the RNA but not the DNA. It is not clear what determines the length of the poly-A tail, but this process involves other proteins that bind specifically to the poly-A sequence. The mature mRNA is then transported from the nucleus, as we discuss in Chapter 14. It is noteworthy that the long tail of As is unique to transcripts made by

**FIGURE 13-25** Polyadenylation and termination. The various steps in this process are described in the text.



Pol II, a feature that allows experimental isolation of protein-coding mRNAs by affinity chromatography.

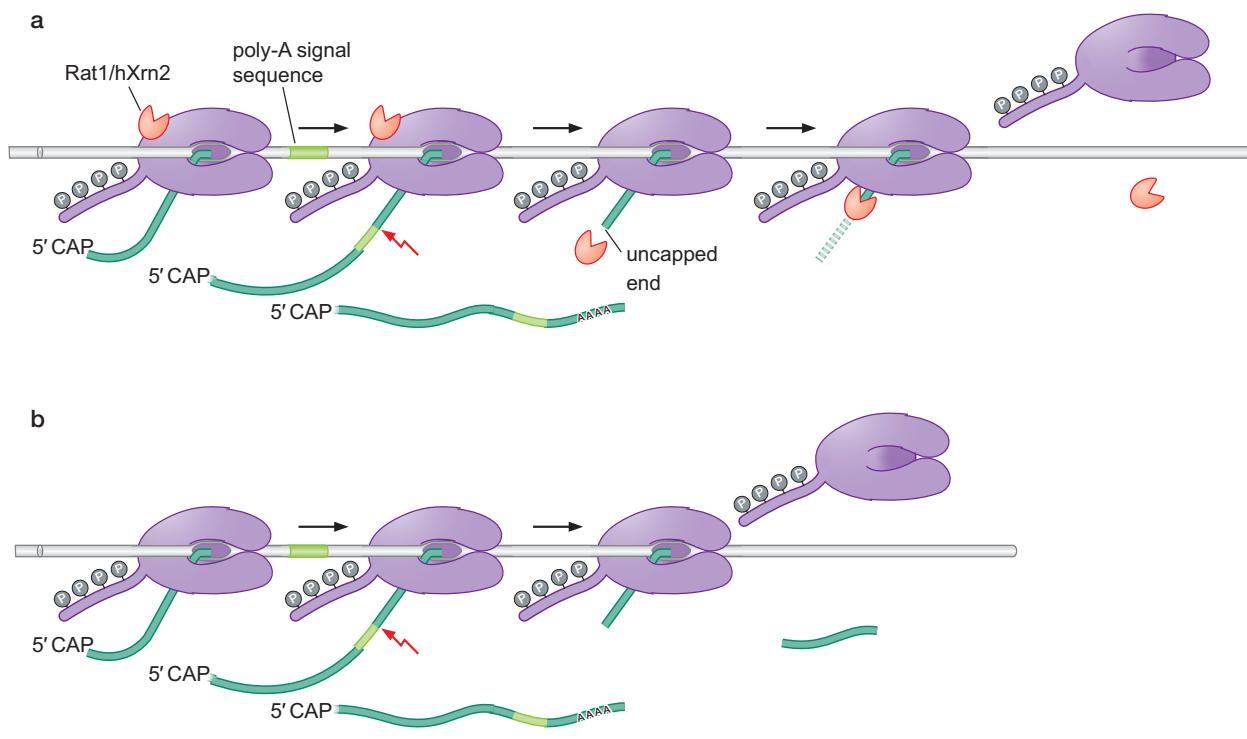
We thus see how a mature mRNA is released from polymerase once the gene has been transcribed. But what terminates transcription by polymerase? In fact, the enzyme does not terminate immediately after the RNA is cleaved and polyadenylated. Rather, it continues to move along the template, generating a second RNA molecule. The polymerase can continue transcribing for several thousand nucleotides before terminating and dissociating from the template. We now describe current models for how termination might happen.

#### Transcription Termination Is Linked to RNA Destruction by a Highly Processive RNase

Polyadenylation is linked to termination, although exactly how is still not quite clear. Recently, however, an enzyme that degrades the second RNA

as it emerges from the polymerase has been identified, and this enzyme may itself trigger termination. This is called the torpedo model of termination (Fig. 13-26a).

The free end of the second RNA is uncapped and thus can be distinguished from genuine transcripts. This new RNA is recognized by an RNase called, in yeast, Rat1 (in humans, Xrn2) that is loaded onto the end of the RNA by another protein (Rtt103) that binds the CTD of RNA polymerase. The Rat1 enzyme is very processive and quickly degrades the RNA in a 5'-to-3' direction, until it catches up to the still-transcribing polymerase from which the RNA is being spewed. Termination may not require any very specific interaction between Rat1 and polymerase and might, in fact, be triggered in a manner rather similar to that described above in the chapter for Rho-dependent termination in bacteria—that is, the highly processing RNase polymerase either pushes polymerase forward and/or pulls the remains of the nascent RNA transcript from the enzyme. It is also possible that other factors are needed in addition to Rat1 to dislodge polymerase as, *in vitro*, Rat1 is alone insufficient to carry out this function, even after it has degraded the transcript.



**FIGURE 13-26** Models of termination: torpedo and allosteric. As described in the text, there are two proposed models for how transcription by eukaryotic RNA Pol II terminates after transcribing a gene. In the figure, the poly-A site is marked by the light green stretch in the DNA and is located just downstream from the gene. It is also light green in the transcript. (The dotted green line) Degraded transcript. (a) In the torpedo model, RNA transcribed downstream from the poly-A site is attacked by the 5'-to-3' RNase (the torpedo), which is loaded onto this transcript from polymerase itself. When this exonuclease catches up with polymerase, it triggers dissociation from the DNA template and termination of transcription. (b) In the allosteric model, the polymerase is highly processive within the gene, and then, once the poly-A signal is passed, becomes less processive. This alteration could be due to a modification or a conformational change. Even in the allosteric model, the second RNA would be degraded by the RNase, but that would not be the cause of termination. In this case, RNA degradation is not shown in the figure to emphasize the different mechanisms of termination in these two models. (Adapted, with permission, from Luo W. and Bentley D. 2004. *Cell* 119: 911–914, Fig. 1. © Elsevier.)

Although the torpedo model for termination is now the favored one, there is an alternative called the allosteric model (Fig. 13-26b). According to this model, termination depends on a conformational change in the elongating polymerase that reduces the processivity of the enzyme leading to spontaneous termination soon afterward. This conformational change would be linked to polyadenylation and could, for example, be triggered by the transfer of the 3'-processing enzymes from the CTD tail of polymerase to the RNA or by the subsequent binding to the CTD tail of other factors that induce a conformational change.

## TRANSCRIPTION BY RNA POLYMERASES I AND III

### RNA Pol I and Pol III Recognize Distinct Promoters but Still Require TBP

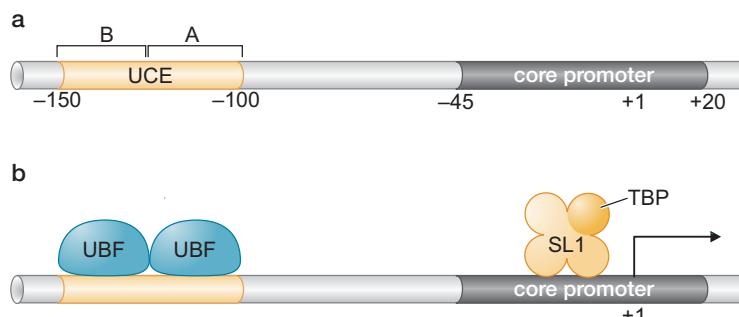
We have already mentioned that all eukaryotes have two other RNA polymerases—Pol I and Pol III—in addition to Pol II. These enzymes are related to Pol II and even share several subunits (Table 13-2), but they initiate transcription from distinct promoters and transcribe distinct genes. Those genes encode specialized RNAs rather than proteins. Each of these enzymes also works with its own unique set of general transcription factors. TBP, however, is universal—it is involved in initiating transcription by Pol I and Pol III, as well as Pol II.

Although TBP is the only GTF that is used by Pol I and Pol III as well as by Pol II, it has emerged recently that some of the other GTFs discussed above in the Pol II case do, in fact, have structurally and functionally equivalent components in the other systems. Thus, for example, TFIIF seems to have a counterpart in two subunits within Pol I (A49/34.5), and also in Pol II (C37/53). Likewise, TFIIE-like subunits are found in Pol I and Pol III enzymes. In addition, both these other systems include additional factors comparable to TFIIB: the TAF1B factor in the Pol I system, and the Brf1 subunit of TFIIB in the case of Pol III.

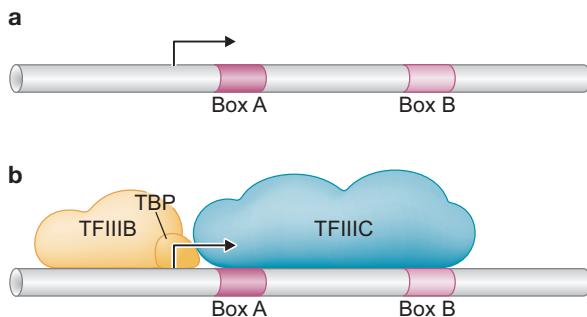
### Pol I Transcribes Just the rRNA Genes

Pol I is required for the expression of only one gene, that encoding the rRNA precursor. There are many copies of that gene in each cell, and, indeed, it is expressed at far higher levels than any other gene, perhaps explaining why it has its own dedicated polymerase.

The promoter for the rRNA gene comprises two parts: the core element and the UCE (upstream control element) as shown in Figure 13-27. The former is located around the start site of transcription, and the latter between



**FIGURE 13-27** Pol I promoter region.  
(a) Structure of the Pol I promoter. (b) Pol I transcription factors. The case shown here is for humans. The set of proteins involved in helping Pol I transcription in yeast is rather different.



**FIGURE 13-28** Pol III core promoter. Shown here is the promoter for a yeast tRNA gene. The order of events leading to transcription initiation is described in the text. For other Pol III genes (such as that for the 5S rRNA), another factor (TFIIIA) is required as well as TFIIB and TFIIC. TFIIIA binds to Box A.

100 and 150 bp upstream (in humans). In addition to Pol I, initiation requires two other factors, called SL1 and UBF. SL1 comprises TBP and three TAFs specific for Pol I transcription. This complex binds to the core element. SL1 binds DNA only in the presence of UBF. This factor binds to UCE, bringing in SL1 and stimulating transcription from the core promoter by recruiting Pol I.

### Pol III Promoters Are Found Downstream from the Transcription Start Site

Pol III promoters come in various forms, and the vast majority have the unusual feature of being located *downstream* from the transcription start site (i.e., within the coding region of the gene). Some Pol III promoters (e.g., those for the tRNA genes) consist of two regions, called Box A and Box B, separated by a short element (Fig. 13-28); others contain Box A and Box C (e.g., the 5S rRNA gene); and still others contain a TATA element like those of Pol II.

Just as with Pol II and Pol I, transcription by Pol III requires transcription factors in addition to polymerase. In this case, the factors are called TFIIB and TFIIC for the tRNA genes and those plus TFIIIA for the 5S rRNA gene.

Figure 13-28 shows the tRNA promoter. Here, the TFIIC complex binds to the promoter region. This complex recruits TFIIB to the DNA just upstream of the start site, where it, in turn, recruits Pol III to the start site of transcription. The enzyme then initiates, presumably displacing TFIIC from the DNA template as it goes. As with the other two classes of polymerase, Pol III uses TBP. In this case, that ubiquitous factor is found within the TFIIB complex.

## SUMMARY

Gene expression is the process by which the information in the DNA double helix is converted into the RNAs and proteins whose activities bestow upon a cell its morphology and functions. Transcription is the first step in gene expression and involves copying DNA into RNA. This process is catalyzed by the enzyme RNA polymerase.

RNA polymerases from bacteria to humans are highly conserved. Eukaryotes have at least three different RNA polymerases each; bacteria have just one. The three ubiquitous eukaryotic enzymes are called RNA Pol I, Pol II, and Pol III. Of these, in this chapter, we focused primarily on Pol II, because this is the enzyme that transcribes the vast majority of genes in the cell and all of the protein-coding genes.

Plants contain two additional RNA polymerases, Pol IV and Pol V.

The basic enzyme from *E. coli*, called the core enzyme, has one copy of each of three subunits— $\beta$ ,  $\beta'$ , and  $\omega$ —and two copies of  $\alpha$ . All of these subunits have homologs in the eukaryotic enzymes. The structures of the bacterial and yeast Pol II enzymes are also similar. Both resemble a crab claw in shape, the pincers being made up of the largest subunits,  $\beta$  and  $\beta'$  in the case of the bacterial enzyme. The active site is at the base of the pincers, and access to and from the active site is afforded through five channels: one channel allows double-stranded DNA to enter between the pincers at the front of the enzyme; two other channels allow the two single

strands—the template and nontemplate strands—to leave the enzyme behind the active site; another channel provides the route by which NTPs enter the active site; and the RNA product, which peels off the DNA template a short distance behind the site of polymerization, exits the enzyme through the fifth channel.

Pol II differs from the bacterial enzyme in one important way. The former has a so-called tail at the carboxy-terminal end of the large subunit, and this is absent from the bacterial enzyme. This tail is made up of multiple repeats of a heptapeptide sequence.

A round of transcription proceeds through three phases called initiation, elongation, and termination. Although RNA polymerases can synthesize RNA unaided, other proteins—called initiation factors—are required for accurate and efficient initiation. These factors ensure that the enzyme initiates transcription only from appropriate sites on the DNA, called promoters. In bacteria, there is only one initiation factor,  $\sigma$ , whereas in eukaryotes there are several, collectively called the general transcription factors. In eukaryotes, the DNA is wrapped within nucleosomes, and, *in vivo*, efficient initiation requires additional proteins, including the Mediator complex and nucleosome-modifying enzymes. Transcriptional activator proteins are also needed (see Chapter 19).

During initiation, RNA polymerase (together with the initiation factors) binds to the promoter in a closed complex. In that state, the DNA remains in a double-stranded form. This closed complex then undergoes isomerization to the open complex. In that form, the DNA around the transcription start site is unwound, disrupting the base pairs and forming a bubble of single-stranded DNA. This transition allows access to the template strand, which determines the order of bases in the new RNA strand. This phase of initiation is followed by promoter escape: once the enzyme has synthesized a series of short RNAs, called abortive initiation, it manages to make a transcript that grows beyond 10 bp. At this point, the enzyme leaves the promoter and enters the elongation phase. During this phase, polymerase moves along the gene while the enzyme performs several functions: it opens the DNA downstream and reseals it upstream (behind) the active site; it adds ribonucleotides to the 3' end of the growing transcript; it peels the newly formed RNA off the template some 8 or 9 bp behind the point of polymerization; and it also proofreads the transcript, checking for (and replacing) incorrectly inserted nucleotides.

Transcription in both bacteria and eukaryotes follows these same steps. There are differences in the two cases, however. For example, in bacteria, isomerization to the open complex occurs spontaneously and does not require ATP hydrolysis. In eukaryotes, this step *does* require ATP hydrolysis. More strikingly, in eukaryotes, promoter escape is regulated by the phosphorylation state of the CTD tail. Thus, the form of Pol II that binds the promoter in the preinitiation complex has an unphosphorylated CTD. This domain becomes phosphorylated by one or more kinases, including the kinase that is part of one of the general transcription factors, TFIH.

Once phosphorylated, the CTD tail of the Pol II frees itself from the other proteins at the promoter, releasing polymerase into the elongation phase. The CTD then binds factors involved in transcriptional elongation and RNA processing. Thus, there is an exchange of initiation for elongation and processing factors as the polymerase moves away from the promoter and starts transcribing the gene. There are also interactions between the elongation factors and those involved in processing, ensuring proper coordination of these events. Another difference between bacteria and eukaryotes is that the latter must deal with nucleosomes during elongation. This requires yet another complex that can dismantle nucleosomes ahead of, and reassemble them behind, the advancing polymerase.

Termination also works differently in bacteria and eukaryotes. Thus, in bacteria, there are two kinds of terminators: intrinsic (Rho-independent) and Rho-dependent. Intrinsic terminators consist of two sequence elements that operate once transcribed into RNA. One element is an inverted repeat that forms a stem-loop in the RNA, disrupting the elongating polymerase. In combination with a string of U nucleotides (which bond only weakly with the template strand), this leads to release of the transcript. Rho-dependent terminators require the ATPase Rho, a protein that hops on elongating transcripts and translocates along them until they reach polymerase, triggering termination. In eukaryotes, termination is closely linked to an RNA processing event called 5' polyadenylation. But in these organisms, too, termination is believed to involve another protein—in this case, an RNase enzyme—traveling along a nascent transcript until it collides with polymerase, triggering termination.

In this chapter, we considered capping of the 5' end of the RNA transcripts, polyadenylation of the 3' end, and the link between the last of these and transcriptional termination. Splicing is described in the next chapter.

## BIBLIOGRAPHY

### Books

*Cold Spring Harbor Symposia on Quantitative Biology*. 1998. Volume 63: Mechanisms of transcription. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

### RNA Polymerase and Transcription Initiation

Brueckner F., Ortiz J., and Cramer P. 2009. A movie of the RNA polymerase nucleotide addition cycle. *Curr. Opin. Struct. Biol.* **19**: 294–299.

Campbell E.A., Westblade L.F., and Darst S.A. 2008. Regulation of bacterial RNA polymerase  $\sigma$  factor activity: A structural perspective. *Curr. Opin. Microbiol.* **11**: 121–127.

Conaway R.C. and Conaway J.W. 2011. Origins and activity of the Mediator complex. *Semin. Cell Dev. Biol.* **22**: 729–734.

Cramer P., Armache K.J., Baumli S., Benkert S., Brueckner F., Buchen C., Damsma G.E., Dengl S., Geiger S.R., Jasiak A.J., et al. 2008. Structure of eukaryotic RNA polymerases. *Annu. Rev. Biophys.* **37**: 337–352.

Ebright R.H. 2000. RNA polymerase: Structural similarities between bacterial RNA polymerase and eukaryotic RNA polymerase II. *J. Mol. Biol.* **304**: 687–698.

- Hahn S. and Young E.T. 2011. Transcriptional regulation in *Saccharomyces cerevisiae*: Transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. *Genetics* **189**: 705–736.
- Kornberg R.D. and Young E.T. 2007. The molecular basis of eukaryotic transcription. *Proc. Natl. Acad. Sci.* **104**: 12955–12961.
- Krishnamurthy S. and Hampsey M. 2009. Eukaryotic transcription initiation. *Curr. Biol.* **19**: R153–R156.
- Malik S. and Roeder R.G. 2005. Dynamic regulation of Pol II transcription by the mammalian Mediator complex. *Trends Biochem. Sci.* **30**: 256–263.
- Roberts J.W. 2006. RNA polymerase, a scrunching machine. *Science* **314**: 1097–1098.
- Saecker R.M., Record M.T. Jr., and Dehaseth P.L. 2011. Mechanism of bacterial transcription initiation: RNA polymerase-promoter binding, isomerization to initiation-competent open complexes, and initiation of RNA synthesis. *J. Mol. Biol.* **412**: 754–771.
- Sekine S., Tagami S., and Yokoyama S. 2012. Structural basis of transcription by bacterial and eukaryotic RNA polymerases. *Curr. Opin. Struct. Biol.* **22**: 110–118.
- Promoters**
- Juven-Gershon T., Hsu J.Y., Theisen J.W., and Kadonaga J.T. 2008. The RNA polymerase II core promoter—The gateway to transcription. *Curr. Opin. Cell Biol.* **20**: 253–259.
- Elongation and RNA Processing**
- Herbert K.M., Greenleaf W.J., and Block S.M. 2008. Single-molecule studies of RNA polymerase: Motoring along. *Annu. Rev. Biochem.* **77**: 149–176.
- Maniatis T. and Reed R. 2002. An extensive network of coupling among gene expression machines. *Nature* **416**: 499–506.
- Perales R. and Bentley D. 2009. “Cotranscriptionality”: The transcription elongation complex as a nexus for nuclear transactions. *Mol. Cell.* **36**: 178–191.
- Petesch S.J. and Lis J.T. 2012. Overcoming the nucleosome barrier during transcript elongation. *Trends Genet.* **28**: 285–294.
- Reinberg D. and Smith R.J. III 2006. de FACTo nucleosome dynamics. *J. Biol. Chem.* **281**: 23297–23301.
- Zhou Q., Li T., and Price D.H. 2021. RNA polymerase II elongation control. *Annu. Rev. Biochem.* **81**: 119–143.

## Termination

- Luo W. and Bartley D. 2004. A ribonucleolytic rat torpedo RNA polymerase II. *Cell* **119**: 911–914.
- Peters J.M., Vangeloff A.D., and Landick R. 2011. Bacterial transcription terminators: The RNA 3'-end chronicles. *J. Mol. Biol.* **412**: 793–813.
- Richardson J.P. 2006. How Rho exerts its muscle as RNA. *Mol. Cell* **23**: 711–712.
- Rosonina E., Kaneko S., and Manley J.L. 2006. Terminating the transcript: Breaking up is hard to do. *Genes Dev.* **20**: 1050–1056.

## RNA Polymerases I and III

- Schramm L. and Hernandez N. 2002. Recruitment of RNA polymerase III to its target promoters. *Genes Dev.* **16**: 2593–2620.
- Vannini A. and Cramer P. 2012. Conservation between the RNA polymerase I, II, and III transcription initiation machineries. *Mol. Cell* **45**: 439–446.
- White R.J. 2008. RNA polymerases I and III, non-coding RNAs and cancer. *Trends Genet.* **24**: 62.

## QUESTIONS

### MasteringBiology®

For instructor-assigned tutorials and problems, go to MasteringBiology.

For answers to even-numbered questions, see Appendix 2: Answers.

**Question 1.** Except for the replacement of Ts with Us, the RNA transcript is identical in sequence to which DNA strand? Choose one or more of the following terms: template strand, non-template strand, coding strand, non-coding strand. Explain your choice.

**Question 2.** Explain why regulation of transcription frequently involves the promoter and protein interactions with the promoter.

**Question 3.** Describe the three steps of transcription initiation that occur before the elongation phase begins focusing on the key features of RNA polymerase at each step.

**Question 4.** Consider a bacterial promoter with –35 and –10 elements. What assay is best to show that RNA polymerase binds at regions centered on the –35 and –10 positions upstream of the start site of transcription? (Reviewing Chapter 7 may help.)

**Question 5.** State whether the following statement is true or false, and explain your conclusion. The sequence of the –35 element is always 5'-TTGACA-3'.

**Question 6.** Given the three models for initial transcription in bacteria (transient excursion, inchworming, and scrunching),

which model represents the hypothesis most supported by data? Describe the general conclusions of these experiments.

**Question 7.** Describe the two proofreading functions of RNA polymerase in prokaryotes.

**Question 8.** Consider the Rho-independent terminator sequence 5'-CCCAGCCCCCUAAUGAGCGGGCUUUUUUUU-3'. Why does a point mutation at any one of the bolded nucleotides disrupt termination of transcription? How would you test your conclusion?

**Question 9.** Explain why the mediator and nucleosome modifiers are required for high levels of transcription in eukaryotic cells but not *in vitro*.

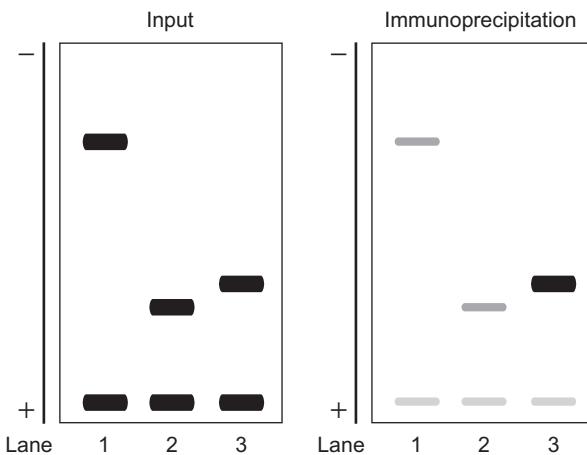
**Question 10.** What steps in the eukaryotic transcription cycle are stimulated by phosphorylation of the carboxyl terminal (CTD) of the large subunit of RNA polymerase II and beyond?

**Question 11.** You want to radiolabel the 5' end of an mRNA through the formation of the 5' RNA cap. Would you want to use α-, β-, or γ-<sup>32</sup>P GTP in your capping reaction? Explain why.

**Question 12.** How does the function of poly-A polymerase differ from RNA polymerase?

**Question 13.** What purposes do capping and poly-A tail addition serve for eukaryotic mRNAs?

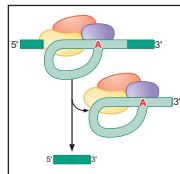
**Question 14.** Researchers studying the torpedo model of eukaryotic termination wanted to test Rtt103 and Rat1 positioning on transcribed genes. To do this, they performed a chromatin immunoprecipitation assay (ChIP) assay using tagged Rat1 protein. (See Chapter 7 for a review of ChIP.) After shearing chromatin from wild-type cells, they immunoprecipitated Rat1 using antibodies specific to the tag. They PCR-amplified the DNA of interest associated with Rat1 using different sets of primers specific for highly transcribed genes. We will show the results for one gene. For the gene *ADH1*, the researchers chose primers specific for amplification of the TATA box region upstream of the open reading frame (ORF) (reaction in lane 1), primers specific for amplification of the 3' region of the ORF (reaction in lane 2), or primers specific for amplification of the DNA just 3' of the sequence encoding the poly-A signal sequence (reaction in lane 3). They compared the PCR results from the immunoprecipitations to the PCR results using the same primers with the input chromatin sample before immunoprecipitation. They included a reaction using primers specific for amplification of a nontranscribed region on chromosome V in every lane (lower band in each reaction). The data is shown below for the PCR of input or immunoprecipitation samples.



- Explain why all the bands are roughly equal in intensity for the input PCR.
- What is the main conclusion from the ChIP results?
- The ChIP data for the other highly transcribed genes looked similar to the data for Rat1 at *ADH1*. Explain how these data support the torpedo model.

Data adapted from Kim et al. (2004. *Nature* **432**: 517–522).

CHAPTER 14



# RNA Splicing

THE CODING SEQUENCE OF A PROTEIN-CODING gene is a series of three-nucleotide codons that specifies the linear sequence of amino acids in its polypeptide product. Thus far, we have tacitly assumed that the coding sequence is contiguous: the codon for one amino acid is immediately adjacent to the codon for the next amino acid in the polypeptide chain. This is true in the vast majority of cases in bacteria and their phage. But it is rarely so for eukaryotic genes. In those cases, the coding sequence is periodically interrupted by stretches of non-coding sequence.

Many eukaryotic genes are thus mosaics, consisting of blocks of coding sequences separated from each other by blocks of non-coding sequences. The coding sequences are called **exons** and the intervening sequences are called **introns**. Once transcribed into an RNA transcript, the introns must be removed and the exons joined together to create the mRNA for that gene. In fact, technically, the term *exon* applies to any region retained in a mature RNA, whether or not it is coding. Non-coding exons include the 5' and 3' untranslated regions of an mRNA; all portions of spliced, stable non-coding RNAs such as the X-chromosome inactivation regulator *Xist* (Chapter 20); and regions that give rise to functional RNAs such as the microRNAs we shall also encounter in Chapter 20.

Figure 14-1 shows a typical eukaryotic gene in which the coding region is interrupted by three introns, splitting it into four exons. The number of introns found within a gene varies enormously—from one in the case of most intron-containing yeast genes (and a few human genes), to 50 in the case of the chicken *proα2* collagen gene, to as many as 363 in the case of the *Titin* gene of humans. Figure 14-2 shows the average number of introns per gene for a range of organisms. Clearly, the average number increases as one looks from simple single-celled eukaryotes, such as yeast, through higher organisms such as worms and flies, all the way up to humans.

The sizes of the exons and introns vary as well. Indeed, introns are very often much longer than the exons they separate. Thus, for example, exons are typically on the order of 150 nucleotides, whereas introns—although they too can be short—can be as long as 800,000 nucleotides (800 kb). As another example, the mammalian gene for the enzyme dihydrofolate reductase is more than 31 kb long, and within it are dispersed six exons that correspond to 2 kb of mRNA. Thus, in this case, the coding portion of the gene is <10% of its total length.

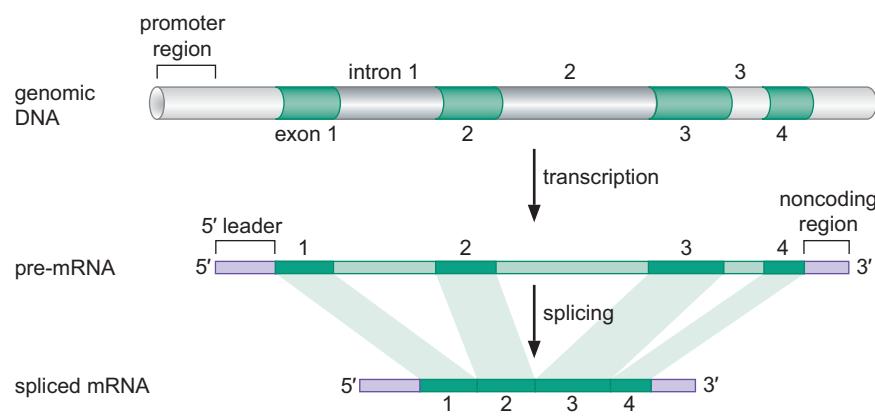
Like the uninterrupted genes of prokaryotes, the split genes of eukaryotes are transcribed into a single RNA copy of the entire gene—the primary transcript for a typical eukaryotic gene contains introns as well as exons. This is shown in the middle part of Figure 14-1. Because of the length and number

## O U T L I N E

The Chemistry of RNA Splicing, 469	•
The Spliceosome Machinery, 473	•
Splicing Pathways, 474	•
Variants of Splicing, 482	•
Alternative Splicing, 483	•
Exon Shuffling, 497	•
RNA Editing, 500	•
mRNA Transport, 503	•

Visit Web Content for Structural Tutorials and Interactive Animations

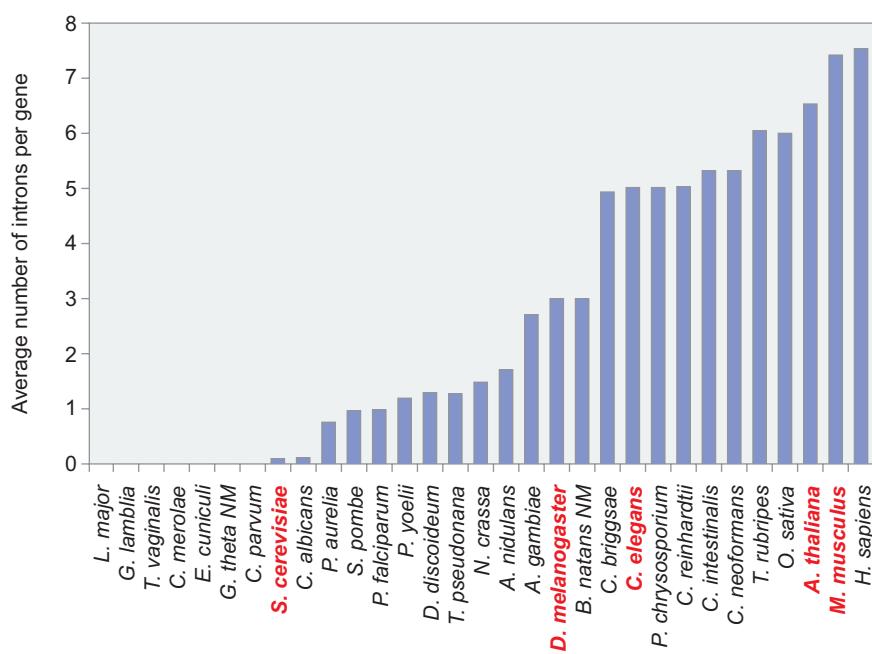
**FIGURE 14-1** A typical eukaryotic gene. The depicted gene contains four coding exons separated by three introns. Transcription from the promoter generates a pre-mRNA, shown in the middle line, that contains all of the exons and introns. Splicing removes the introns and fuses the exons to generate the mature mRNA that, once processed further (see polyadenylation, Chapter 13) and exported from the nucleus, can be translated to give a protein product. Technically, the 5' leader and 3' non-coding regions are also exons because they are retained in the mature mRNA. They are shown here in purple to indicate their status as non-coding exons.



of introns, the primary transcript (or **pre-mRNA**) can be very long indeed. In the extreme case of the human *dystrophin* gene, RNA polymerase must traverse 2400 kb of DNA to copy the entire gene into RNA. (Given that transcription proceeds at a rate of 40 nucleotides per second, it can readily be seen that it takes a staggering 17 h to make a single transcript of this gene!) This raises the possibility that exon abundance and length could have a significant effect on the expression rate of genes, a matter we return to when we consider gene regulation during development in Chapter 21.

As we have said, the primary transcripts of intron-containing genes must have their introns removed before they can be translated into proteins. The process of intron removal, called **RNA splicing**, converts the pre-mRNA into mature mRNA and must occur with great precision to avoid the loss, or addition, of even a single nucleotide at the sites at which the exons are joined. As we shall see in Chapters 15 and 16, the triplet-nucleotide codons of mRNA are translated in a fixed reading frame that is set by the first codon in the protein-coding sequence. Lack of precision in splicing—if, for example, a base were lost or gained at the boundary between two exons—would throw

**FIGURE 14-2** Number of introns per gene in various eukaryotic species. The average number of introns per gene is shown for a selection of eukaryotic species. The names in red are those of the common model organisms (Appendix 1): the yeast (*Saccharomyces cerevisiae*), the fruit fly (*Drosophila melanogaster*), the roundworm (*Caenorhabditis elegans*), the plant (*Arabidopsis thaliana*), and the mouse (*Mus musculus*). The other species shown are *Anopheles gambiae*; *Aspergillus nidulans*; *Bigelovella natans* nucleomorph; *Caenorhabditis briggsae*; *Candida albicans*; *Chlamydomonas reinhardtii*; *Ciona intestinalis*; *Cryptococcus neoformans*; *Cryptosporidium parvum*; *Cyanidioschyzon merolae*; *Dictyostelium discoideum*; *Encephalitozoon cuniculi*; *Giardia lamblia*; *Guillardia theta* nucleomorph; *Homo sapiens*; *Leishmania major*; *Neurospora crassa*; *Oryza sativa*; *Paramecium aurelia*; *Phanerochaete chrysosporium*; *Plasmodium falciparum*; *Plasmodium yoelii*; *Schizosaccharomyces pombe*; *Takifugu rubripes*; *Thalassiosira pseudonana*; and *Trichomonas vaginalis*. (Redrawn, with permission, from Roy S.W. and Gilbert W. 2006. *Nat. Rev. Genet.* 7: 212, Fig. 1. © Macmillan.)



the reading frames of exons out of register: downstream codons would be incorrectly selected and the wrong amino acids incorporated into proteins.

Some pre-mRNAs can be spliced in more than one way. Thus, mRNAs containing different selections of exons can be generated from a given pre-mRNA. Called **alternative splicing**, this strategy enables a gene to give rise to more than one polypeptide product. These alternative products are called **isoforms**. It is estimated that 90% or more of the protein-coding genes in the human genome are spliced in alternative ways to generate more than one isoform.

The number of different variants a given gene can encode in this way varies from two to hundreds or even thousands. For example, the *Slo* gene from rat, which encodes a potassium channel expressed in neurons, has the potential to encode 500 alternative versions of that product. And, as we shall see, one particular *Drosophila* gene can encode as many as 38,000 possible products as a result of alternative splicing. Alternative splicing is often a regulated process, with different isoforms being produced in response to different signals or in different cell types.

In this chapter, we discuss not only the mechanisms and regulation of RNA splicing, but also ideas about why eukaryotic genes have interrupted coding regions. We also describe RNA editing, another way initial transcripts can be altered to change what they encode.

Splicing was discovered in studies of gene expression in the mammalian adenovirus, as described in Box 14-1, Adenovirus and the Discovery of Splicing.

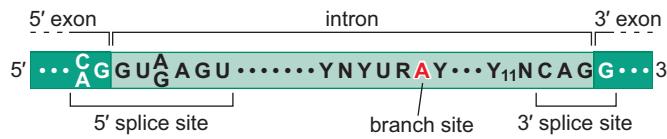
## THE CHEMISTRY OF RNA SPLICING

### Sequences within the RNA Determine Where Splicing Occurs

We now consider the molecular mechanisms of the splicing reaction (see Interactive Animation 14-1). How are the introns and exons distinguished from each other? How are introns removed? How are exons joined with high precision? The borders between introns and exons are marked by specific nucleotide sequences within the pre-mRNAs. These sequences delineate where splicing will occur. Thus, as shown in Figure 14-3, the exon–intron boundary—that is, the boundary at the 5' end of the intron—is marked by a sequence called the **5' splice site**. The intron–exon boundary at the 3' end of the intron is marked by the **3' splice site**. (The 5' and 3' splice sites were sometimes referred to as the **donor** and **acceptor** sites, respectively, but this nomenclature is rarely used today.)



The figure shows a third sequence necessary for splicing. This is called the **branchpoint site** (or branchpoint sequence). It is found entirely within the intron, usually close to its 3' end, and is followed by a polypyrimidine tract (Py tract).



**FIGURE 14-3** Sequences at intron–exon boundaries. The consensus sequences for both the 5' and 3' splice sites, and also the conserved A at the branch site. As in other cases of consensus sequences, where two alternative bases are similarly favored, those bases are both indicated at that position. In this figure, the consensus sequences shown are for humans. This is true for all other figures in this chapter, unless otherwise stated.

The consensus sequence for each of these elements is shown in Figure 14-3. The most highly conserved sequences are the GU in the 5' splice site, the AG in the 3' splice site, and the A at the branch site. These highly conserved nucleotides are all found within the intron itself—perhaps not surprisingly, because the sequence of most exons, in contrast to the introns, is constrained by the need to encode the specific amino acids of the protein product.

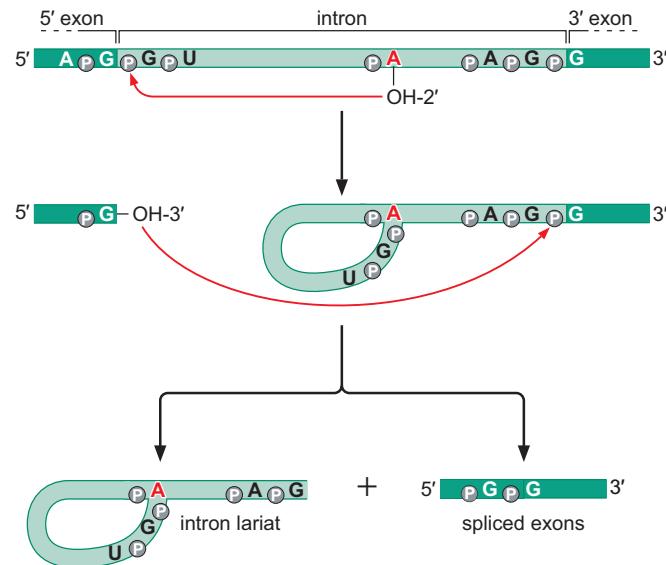
### The Intron Is Removed in a Form Called a Lariat as the Flanking Exons Are Joined

Let us begin by considering the chemistry of splicing. An intron is removed through two successive **transesterification** reactions in which phosphodiester linkages within the pre-mRNA are broken and new ones are formed (Fig. 14-4). The first reaction is triggered by the 2'-OH of the conserved A at the branch site. This group acts as a nucleophile to attack the phosphoryl group of the conserved G in the 5' splice site. (This is an  $S_N2$  reaction that proceeds through a pentavalent phosphorous intermediate.)

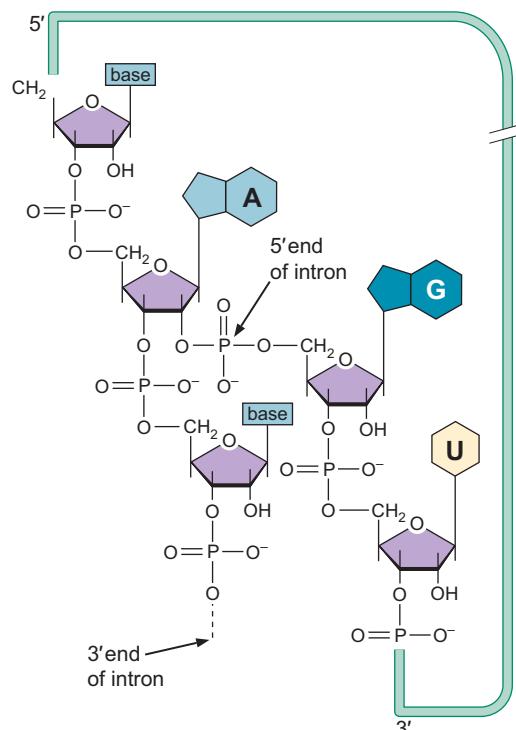
As a consequence of this first reaction, the phosphodiester bond between the sugar and the phosphate at the 5' junction between the intron and the exon is cleaved. The freed 5' end of the intron is joined to the A within the branch site. Thus, in addition to the 5' and 3' backbone linkages, a third phosphodiester extends from the 2'-OH of that A to create a three-way junction (hence its description as a branchpoint). The structure of the three-way junction is shown in Figure 14-5.

Note that the 5' exon is a leaving group in the first transesterification reaction. In the second reaction, the 5' exon (more precisely, the newly liberated 3'-OH of the 5' exon) reverses its role and becomes a nucleophile that attacks the phosphoryl group at the 3' splice site (Fig. 14-4). This second reaction has two consequences. First, and most importantly, it joins the 5' and 3' exons; thus, this is the step in which the two coding sequences are actually “spliced” together. Second, this same reaction liberates the intron, which serves as a leaving group. Because the 5' end of the intron had been joined to branchpoint A in the first transesterification reaction, the newly liberated intron has the shape of a **lariat**.

In the two reaction steps, there is no net gain in the number of chemical bonds—two phosphodiester bonds are broken, and two new ones made.



**FIGURE 14-4** The splicing reaction. The two steps of the splicing reaction described in the text. In the first step, the RNA forms a loop structure, which is shown in detail in the next figure.



**FIGURE 14-5** The structure of the three-way junction formed during the splicing reaction.

## ► KEY EXPERIMENTS

### Box 14-1 Adenovirus and the Discovery of Splicing

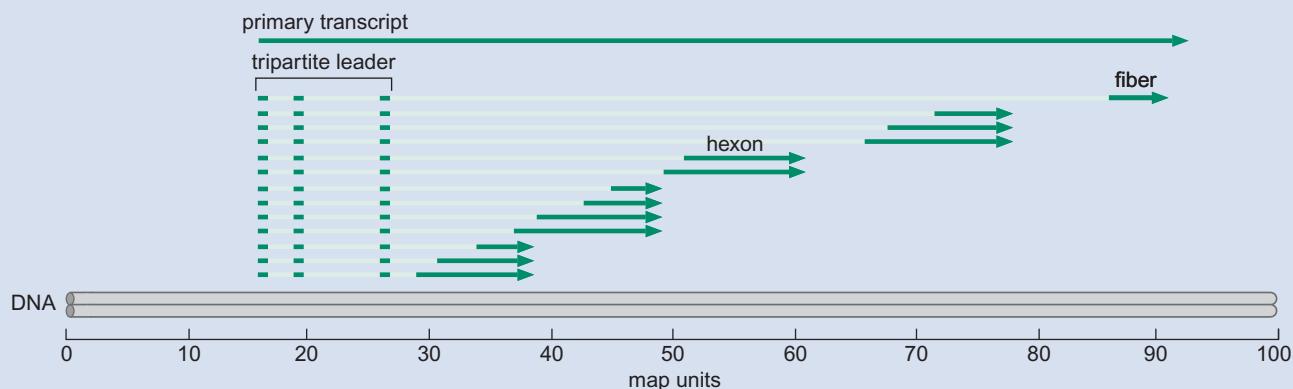
Studies with bacteria and their phage led to the view that the mRNA is an exact replica in terms of nucleotide sequence of the gene from which it is transcribed (see Chapter 16). It therefore came as a shock when, in 1977, it was discovered that certain (and, as we now know, most) eukaryotic mRNAs are spliced together in patchwork fashion from much longer primary transcripts. How was this startling discovery made?

In an effort to understand gene transcription in eukaryotes, scientists focused on the human DNA virus called **adenovirus**. This virus was intended to serve as a model for understanding the molecular biology of the eukaryotic gene just as phage T4 and  $\lambda$  had done for the prokaryotic gene (see Appendix 1). The virion of adenovirus is composed of several different virus-encoded proteins, and the mRNAs for these proteins were purified with the hope that their 5' termini would pinpoint the transcription initiation sites for each gene on the viral genome. Instead, all of the mRNAs, even though they encoded different proteins, were found to have identical 5' sequences. We now know that all of the mRNAs for the virion proteins of adenovirus arise from a single promoter known as the major late promoter. Initiation from this promoter generates long transcripts that span the coding sequences for multiple proteins (Box 14-1 Fig. 1). This transcript then undergoes alternative splicing to generate separate mRNAs for individual virion components such as the hexon and fiber proteins. All of the mRNAs share the same 5' sequence, which is stitched together from three short non-protein-coding sequences known as the

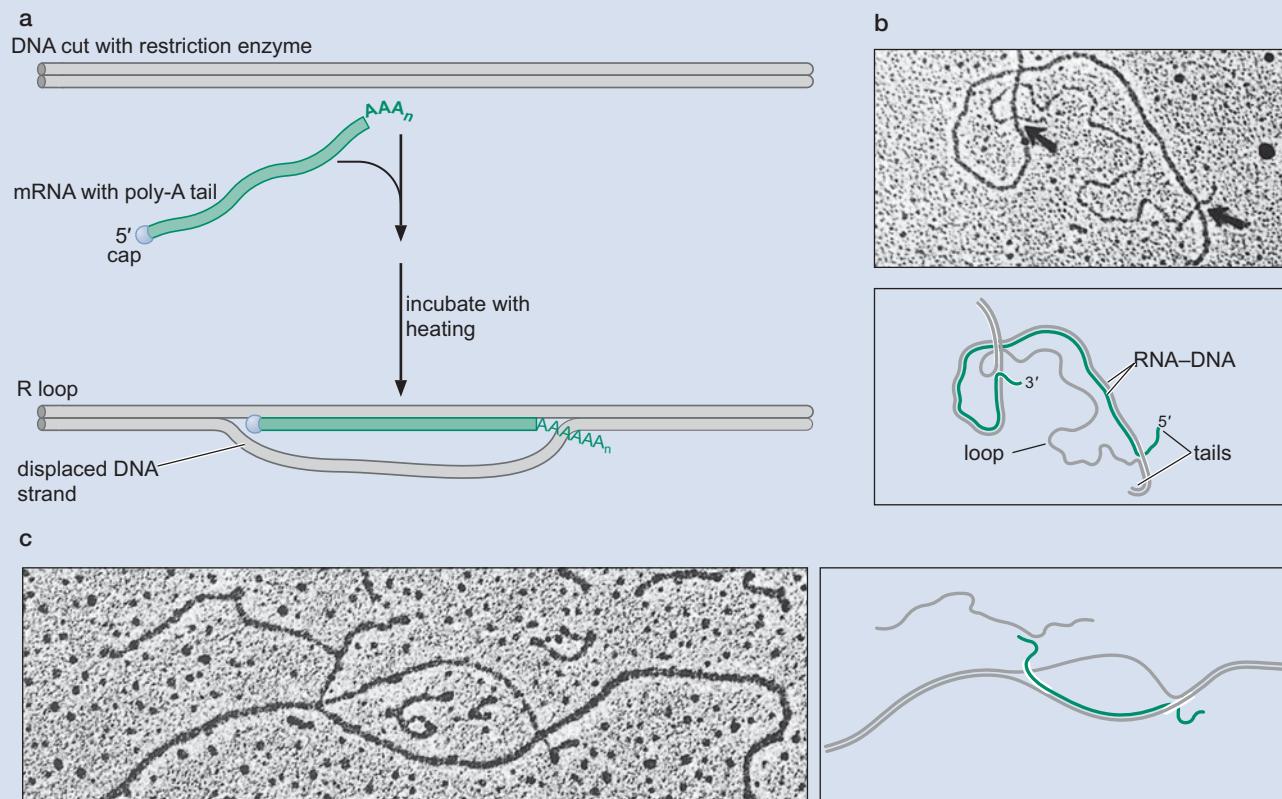
"tripartite leader." The leader is then alternatively spliced to the coding sequences for the hexon, fiber, and other virion proteins to generate each of the late viral mRNAs.

That these messengers are spliced together from RNAs arising from several regions of the genome emerged from a variety of experiments—one of which is known as R-loop mapping (Box 14-1 Fig. 2). When RNA is incubated, under the appropriate conditions, with a double-stranded DNA containing a stretch of sequence identical to that of the RNA, the RNA anneals to its complement, displacing a stretch of the noncomplementary strand in the form of a loop (Box 14-1 Fig. 2a). Following the staining procedure used to visualize nucleic acids, this R loop can be observed in the electron microscope, because RNA–DNA and DNA–DNA duplexes appear thicker than single-stranded nucleic acids. When such an experiment was performed with adenovirus messengers, the resulting R loops were found not to be fully contiguous with a single region of DNA. Instead, and depending on which fragment of viral DNA was used, one or both ends of the RNA were found to protrude from the RNA loops as single-strand tails (Box 14-1 Fig. 2b). In other cases, one of the tails is seen to anneal with a DNA fragment from a different region of the viral genome (Box 14-1 Fig. 2c). Clearly, these mRNAs were composite molecules that had been joined together from sequences complementary to noncontiguous regions of the genome. These and other kinds of DNA–RNA annealing experiments were used to deduce the pattern of alternative splicing shown in Box 14-1 Figure 1.

**Box 14-1** (Continued)



**BOX 14-1 FIGURE 1** Map of the human adenovirus-2 genome. The map shows the transcription patterns of the late mRNAs, including the primary transcript (the long, dark-green arrow at the top); the tripartite leader sequences found at positions 16.6, 19.6, and 26.6 (green bars); and the map positions of the DNA sequences that encode the various late mRNAs (the late mRNAs are shown as short, dark-green arrows).



**BOX 14-1 FIGURE 2** R-loop mapping of the adenovirus-2 late messenger RNAs. (a) The schematic shows the formation of an R-loop structure. A double-stranded DNA fragment generated by digestion with a restriction endonuclease is incubated with mRNA and heated to just above the melting temperature of the DNA in 80% formamide. The hybrid formed between the messenger and its complementary DNA sequence results in displacement of the second DNA strand. The poly-A tail of the mRNA (not encoded by DNA) (see Chapter 12) is seen projecting from the end of the hybrid duplex. (b) Electron micrograph and schematic diagram of an R loop observed after incubating hexon mRNA with a complementary DNA sequence from the late region of the adenovirus-2 genome. Note the extensions of both the 5' and 3' ends of the messenger. (Gray lines) The DNA; (green lines) the RNA. (Reprinted, with permission, from Berget S.M. et al. 1977. *Proc. Natl. Acad. Sci.* **74**: 3171–3175. © National Academy of Sciences.) (c) Electron micrograph and schematic diagram of an R loop observed after incubating fiber mRNA with two DNAs, the complete adenovirus genome, and a restriction endonuclease fragment derived from the early region of the genome. (Reprinted, with permission, from Chow L.T. et al. 1977. *Cell* **12**: 1–8, p. 2. © Elsevier.)

Because it is just a question of shuffling bonds, no energy input is demanded by the chemistry of this process. But, as we shall see later, a large amount of ATP is consumed during the splicing reaction. This energy is required, not for the chemistry, but to properly assemble and operate the splicing machinery.

Another point regarding the splicing reaction is direction: what ensures that splicing only goes forward—that is, toward the products shown in Figure 14-4? In principle, the reactions could go in the other direction, and indeed this can be forced to happen under special circumstances. But in practice, this does not happen in the cell, and we will describe presently how this is ensured.

## THE SPLICEOSOME MACHINERY

### RNA Splicing Is Performed by a Large Complex Called the Spliceosome

The transesterification reactions just described are mediated by a huge molecular “machine” called the **spliceosome** (see Interactive Animation 14-1). This complex comprises about 150 proteins and five RNAs and is similar in size to a ribosome, the machine that translates mRNA into protein (Chapter 15). In performing even a single splicing reaction, the spliceosome hydrolyzes several molecules of ATP. Strikingly, it is believed that many of the functions of the spliceosome are performed by its RNA components rather than the proteins, again reminiscent of the ribosome. Thus, RNAs locate the sequence elements at the intron–exon borders and likely participate in catalysis of the splicing reaction itself.



The five RNAs (U1, U2, U4, U5, and U6) are collectively called **small nuclear RNAs (snRNAs)**. Each of these RNAs is between 100 and 300 nucleotides long in most eukaryotes and is complexed with several proteins. These RNA–protein complexes are called **small nuclear ribonuclear proteins (snRNPs**—pronounced “snurps”). In Chapter 6, we saw the crystal structure of a section of the U1 snRNA bound to one of the proteins of the U1 snRNP (Fig. 6-18).

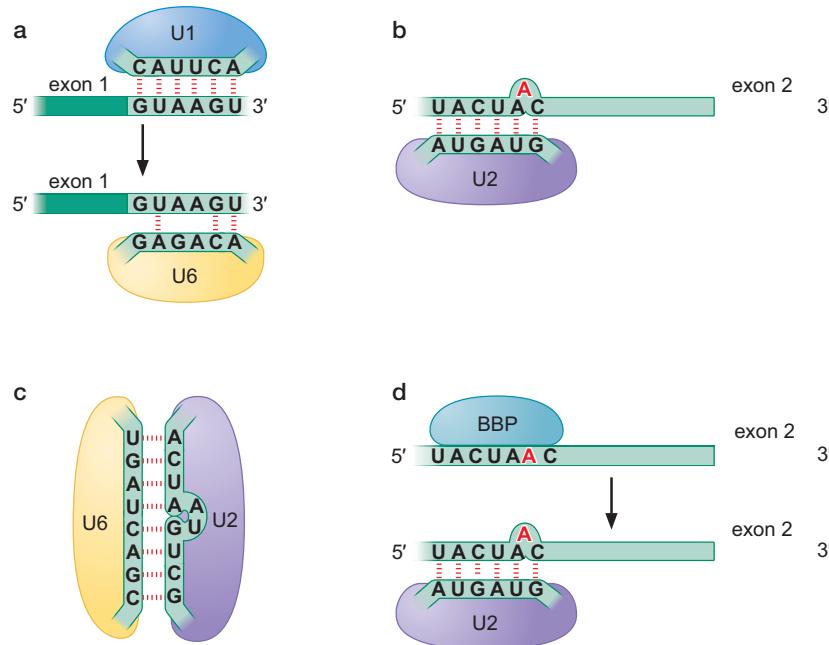
The spliceosome is the large complex made up of these snRNPs, but the exact makeup differs at different stages of the splicing reaction: different snRNPs come and go at different times, each performing particular functions in the reaction. There are also many proteins within the spliceosome that are not part of the snRNPs, and others besides that are only loosely bound to the spliceosome.

The snRNPs have three roles in splicing: they recognize the 5' splice site and the branch site; they bring those sites together as required; and they catalyze (or help to catalyze) the RNA cleavage and joining reactions. To perform these functions, RNA–RNA, RNA–protein, and protein–protein interactions are all important. We start by considering some of the RNA–RNA interactions. These operate within individual snRNPs, between different snRNPs, and between snRNPs and the pre-mRNA.

Thus, for example, Figure 14-6a shows the interaction, through complementary base pairing, of the U1 snRNA and the 5' splice site in the pre-mRNA. Subsequently in the reaction, that splice site is recognized by the U6 snRNA. In another example, shown in Figure 14-6b, the branch site is recognized by the U2 snRNA. A third example, in Figure 14-6c, shows an interaction between U2 and U6 snRNAs. This brings the 5' splice site and the branch site together. It is these and other similar interactions, and the rearrangements they lead to, that drive the splicing reaction and contribute to its precision, as we shall see a little later.

**FIGURE 14-6** Some RNA–RNA hybrids formed during the splicing reaction.

In some cases, (a) different snRNPs recognize the same (or overlapping) sequences in the pre-mRNA at different stages of the splicing reaction, as shown here for U1 and U6 recognizing the 5' splice site. (b) snRNP U2 is shown recognizing the branch site. (c) The RNA:RNA pairing between the snRNPs U2 and U6. Finally, (d), the same sequence within the pre-mRNA is recognized by a protein (not part of an snRNP) at one stage and displaced by an snRNP at another. Each of these changes accompanies the arrival or departure of components of the spliceosome and a structural rearrangement that is required for the splicing reaction to proceed. The sequences in this figure are from yeast.



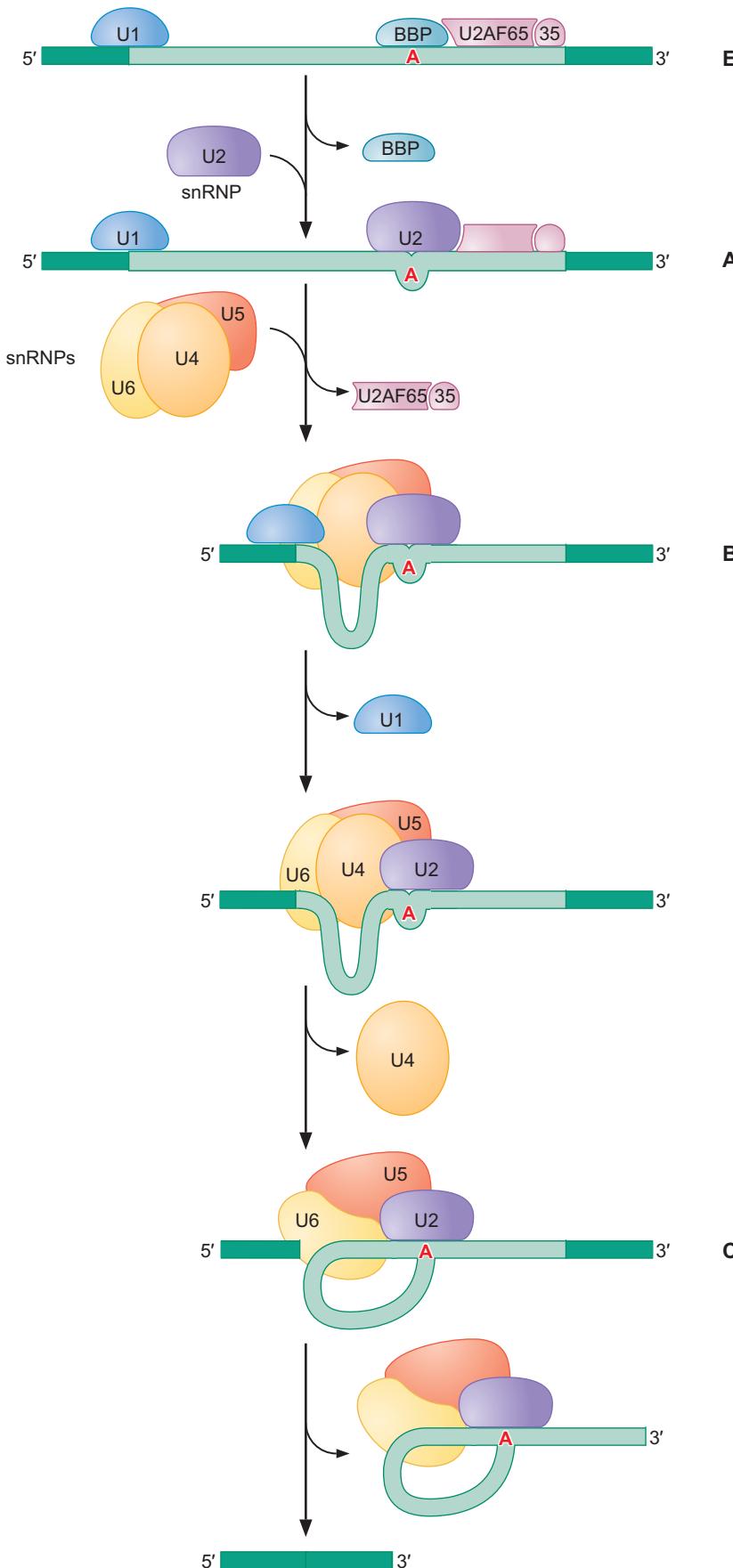
Some non-snRNPs are involved in splicing as mentioned above. One example, U2AF (U2 auxiliary factor), recognizes the polypyrimidine (Py) tract/3' splice site and, in the initial step of the splicing reaction, helps another protein, branchpoint-binding protein (BBP), bind to the branch site. BBP (also called SF1 in mammalian systems) is then displaced by the U2 snRNP, as shown in Figure 14-6d. Other proteins involved in the splicing reaction include RNA-annealing factors, which help load snRNPs onto the mRNA, and DEAD-box helicase proteins. The latter use their ATPase activity to dissociate given RNA–RNA interactions, allowing alternative pairs to form and thereby driving the rearrangements that occur through the splicing reaction. They are also required to remove spliced mRNA from the spliceosome and trigger spliceosome disassembly.

## SPLICING PATHWAYS

### Assembly, Rearrangements, and Catalysis within the Spliceosome: The Splicing Pathway

The steps of splicing are shown in Figure 14-7. What is shown is a canonical pathway, and in any given case a number of the steps may differ slightly in their order or might even on occasion reverse, matters we return to later. But the pathway as we first present it reveals the extraordinary series of events undertaken by the dynamic spliceosome to drive the splicing reaction in the cell.

Initially, the 5' splice site is recognized by the U1 snRNP (using base pairing between its snRNA and the pre-mRNA, shown in Fig. 14-6). U2AF is made up of two subunits, the larger of which (65) binds to the Py tract and the smaller (35) binds to the 3' splice site. The former subunit interacts with BBP (SF1) and helps that protein bind to the branch site. This arrangement of proteins and RNA is called the early (E) complex.



**FIGURE 14-7** Steps of the spliceosome-mediated splicing reaction. The assembly and action of the spliceosome; the details of each step are described in the text. Components of the splicing machinery arrive or leave the complex at each step, changes that are associated with structural rearrangements necessary for the splicing reaction to proceed. Note that the name of each complex is shown to the right. There is evidence to suggest that some of the components shown do not arrive or leave precisely when indicated in this figure; they may, for example, remain present but weaken their association with the complex rather than dissociating completely. It is also not possible to be sure of the order of some changes shown, particularly the two steps involving changes in U6 pairing: when it takes over from U1 at the 5' splice site, compared with when it takes over from U4 in binding U2. Despite these uncertainties, the critical involvement of different components of the machinery at different stages of the splicing reaction and the general dynamic nature of the spliceosome are as shown.

U2 snRNP then binds to the branch site, aided by U2AF and displacing BBP (SF1). This arrangement is called the A complex. The base pairing between the U2 snRNA and the branch site is such that the branch site A residue is extruded from the resulting stretch of double-helical RNA as a single-nucleotide bulge, as shown in Figure 14-6b. This A residue is thus unpaired and available to react with the 5' splice site.

The next step is a rearrangement of the A complex to bring together all three splice sites. This is achieved as follows: the U4 and U6 snRNPs, along with the U5 snRNP, join the complex. Together, these three snRNPs are called the **tri-snRNP particle**, within which the U4 and U6 snRNPs are held together by complementary base pairing between their RNA components, and the U5 snRNP is more loosely associated through protein–protein interactions. With the entry of the tri-snRNP, the A complex is converted into the B complex.

In the next step, U1 leaves the complex, and U6 replaces it at the 5' splice site. This requires that the base pairing between the U1 snRNA and the pre-mRNA be broken, allowing the U6 RNA to anneal with the same region (in fact, to an overlapping sequence, as shown in Fig. 14-6a).

Those steps complete the assembly pathway. The next rearrangement triggers catalysis and occurs as follows: U4 is released from the complex, allowing U6 to interact with U2 (through the RNA:RNA base pairing shown in Fig. 14-6c). This arrangement, called the C complex, produces the active site. That is, the rearrangement brings together within the spliceosome those components—believed to be solely regions of the U2 and U6 RNAs—that together form the active site. The same rearrangement also ensures that the substrate RNA is properly positioned to be acted upon. It is striking not only that the active site is primarily formed of RNA, but also that it is only formed at this stage of spliceosome assembly. Presumably, this strategy lessens the chance of aberrant splicing. Linking the formation of the active site to the successful completion of earlier steps in spliceosome assembly makes it highly likely that the active site is available only at legitimate splice sites.

Formation of the active site juxtaposes the 5' splice site of the pre-mRNA and the branch site, facilitating the first transesterification reaction. The second reaction, between the 5' and 3' splice sites, is aided by the U5 snRNP, which helps to bring the two exons together. The final step involves release of the mRNA product and the snRNPs. The snRNPs are initially still bound to the lariat, but they get recycled after rapid degradation of that piece of RNA.

### Spliceosome Assembly Is Dynamic and Variable and Its Disassembly Ensures That the Splicing Reaction Goes Only Forward in the Cell

It is important to emphasize that the pathway we have just described—the order of the steps required to assemble the spliceosome—is the canonical version. In truth, the process can be less tightly regimented than this description suggests. For one thing, the picture we have presented shows the machinery assembling around the intron to be removed. In fact, it is possible that more often the machinery initially assembles around an exon, a process often called **exon definition** (we describe more on this presently, when we consider the actions of splicing enhancers). In addition, the precise order of events likely varies to some extent—for example, splice site pairing can occur either before or after tri-snRNP recruitment: the details will depend on the RNA sequences and rate-limiting step in any given case. Furthermore, many of the steps during spliceosome assembly can be reversed.

Above we mentioned that the two reactions at the heart of splicing could in principle go backward as well as forward, but that in the cell that is not seen. This directionality is ensured because the spliceosome rapidly disassembles immediately after the second reaction takes place. Disassembly is driven by one of the DEAD-box helicase proteins we mentioned, this one called Prp22. This protein is required for the second catalytic step of splicing and also for stripping the spliced mRNA from the spliceosome. Mutations that eliminate this latter function also block spliceosome disassembly, and in such a situation the splicing reactions can be seen to go backward and forward in purified spliceosomes.

It might seem odd that the machinery and mechanism of splicing is so complicated. How did it evolve that way? Would it not have been simpler to fuse the exons in a single reaction, rather than undergo the two reactions just described? To consider this question, we turn to a group of introns that—unlike those we have considered thus far—can splice *themselves* out of pre-mRNA without the need for the spliceosome. They are called **self-splicing introns**.

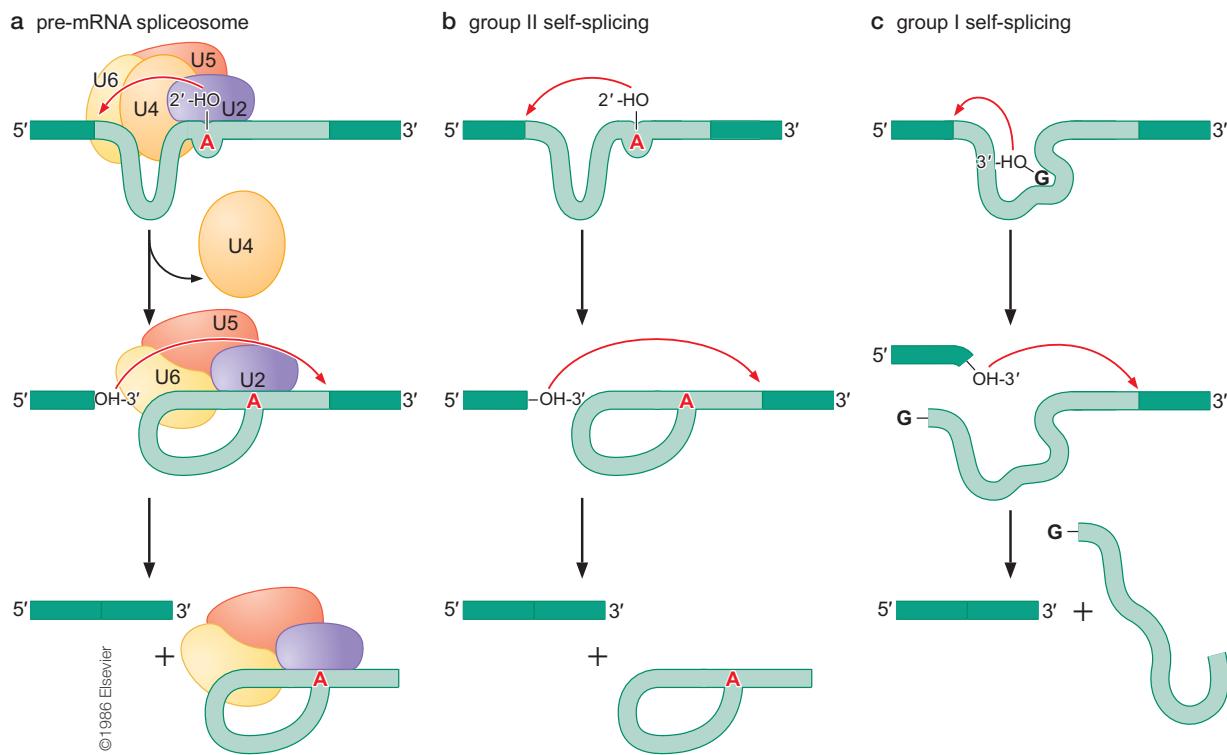
### Self-Splicing Introns Reveal That RNA Can Catalyze RNA Splicing

The three classes of splicing found in cells (not including tRNA processing, which we discuss in Chapter 15) are shown in Table 14-1. Thus far, we have dealt only with nuclear pre-mRNA splicing, that mediated by the spliceosome found in all eukaryotes. Also shown in Table 14-1 are the so-called **group I** and **group II** self-splicing introns. By “self-splicing,” we mean that the intron itself folds into a specific conformation within the precursor RNA and catalyzes the chemistry of its own release (recall that we discussed the general features of RNA enzymes in Chapter 5). In terms of a practical definition, “self-splicing” refers to introns that can remove themselves from RNAs in the test tube in the absence of any proteins or other RNA molecules. The self-splicing introns are grouped into two classes on the basis of their structure and splicing mechanism. Strictly speaking, self-splicing introns are not enzymes (“catalysts”) because they mediate only one round of RNA processing (as we consider in Box 14-2, Converting Group I Introns into Ribozymes).

In the case of group II introns, the chemistry of splicing and the RNA intermediates produced are the same as those for nuclear pre-mRNAs. For example, as shown in Figure 14-8, the intron uses an A residue within the branch site to attack the phosphodiester bond at the boundary between its 5' end and the end of the 5' exon—that is, at the 5' splice site. This reaction produces the branched lariat, as seen above, and is followed by a second reaction in which the newly freed 3'-OH of the exon attacks the 3' splice site, releasing the intron as a lariat and fusing the 3' and 5' exons.

**TABLE 14-1** Three Classes of RNA Splicing

Class	Abundance	Mechanism	Catalytic Machinery
Nuclear pre-mRNA	Very common; used for most eukaryotic genes	Two transesterification reactions; branch site A	Major and minor spliceosomes
Group II introns	Rare; some eukaryotic genes from organelles and prokaryotes	Same as pre-mRNA	RNA enzyme encoded by intron (ribozyme)
Group I introns	Rare; nuclear rRNA in some eukaryotes, organelle genes, and a few prokaryotic genes	Two transesterification reactions; branch site G	Same as group II



**FIGURE 14-8** Group I and group II introns. This figure compares the reaction of the self-splicing group I and II introns and the spliceosome-mediated reaction already described. The chemistry in the case of group II introns is essentially the same as in the spliceosome case, with a highly reactive adenine within the intron initiating splicing and leading to the formation of a lariat product. In the case of the group I intron, the RNA folds in a way that forms a guanine-binding pocket, which allows the molecule to bind a free guanine nucleotide and use that to initiate splicing. Although these introns can splice themselves out of RNA molecules unaided by proteins *in vitro*, *in vivo* they typically do require protein components to stimulate the reaction. (Adapted, with permission, from Cech T.R. 1986. *Cell* 44: 207–210, Fig. 1. © Elsevier.)



### Group I Introns Release a Linear Intron Rather Than a Lariat

Group I introns splice by a different pathway (Fig. 14-8c and Interactive Animation 14-2). Instead of a branchpoint A residue, they use a free G nucleotide or nucleoside. This G species is bound by the RNA, and its 3'-OH group is presented to the 5' splice site. The same type of transesterification reaction that leads to the lariat formation in the earlier examples here fuses the G to the 5' end of the intron. The second reaction now proceeds just as it does in the earlier examples: the freed 3' end of the exon attacks the 3' splice site. This fuses the two exons and releases the intron, although, in this case, the intron is linear rather than a lariat structure.

Group I introns, which are smaller than group II introns, share a conserved secondary structure (RNA folding is discussed in Chapter 5). The structure of group I introns includes a binding pocket that will accommodate any guanine nucleotide or nucleoside as long as it is a ribose form. In addition to the nucleotide-binding pocket, group I introns contain an “internal guide sequence” that base-pairs with the 5' splice site sequence and thereby determines the precise site at which nucleophilic attack by the G nucleotide takes place (see Box 14-2).

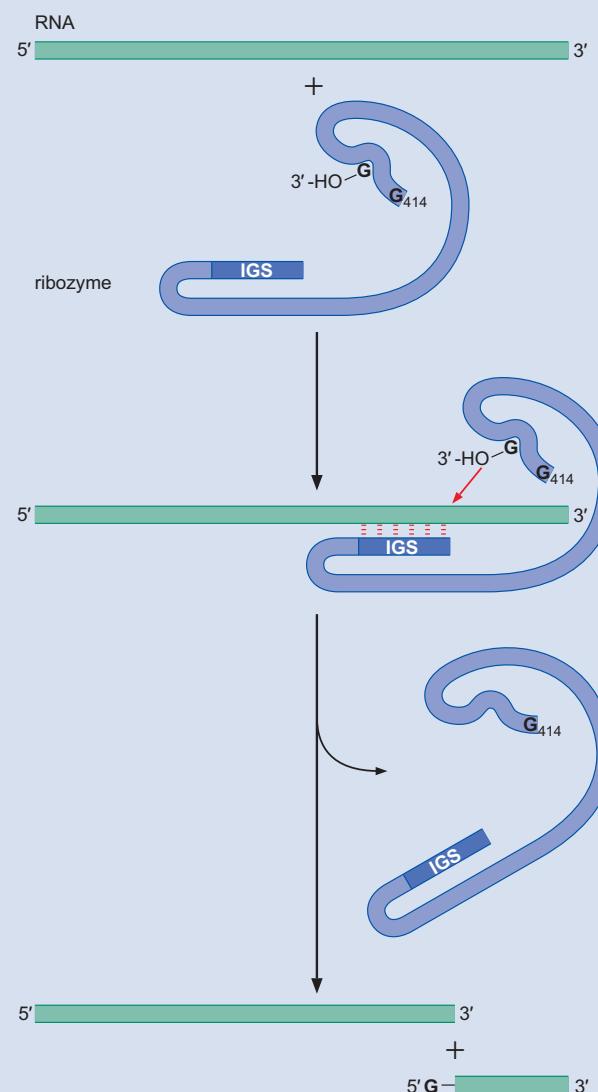
A typical self-splicing intron is between 400 and 1000 nucleotides long, and, in contrast to introns removed by spliceosomes, much of the sequence

## ► KEY EXPERIMENTS

**Box 14-2** Converting Group I Introns into Ribozymes

Once a group I self-splicing intron has been spliced out, the active site it contains remains intact. So what prevents this splicing reaction from reversing itself? One factor is the high cellular concentration of G nucleotides—this strongly favors the forward reaction. But in addition, the intron undergoes a further reaction that effectively prevents it from participating in the back reaction. Conveniently, at the extreme 3' end of the intron is a G, which can bind in the G-binding pocket. Meanwhile, the 5' end of the intron can bind along the internal guide sequence. Thus, a third transesterification reaction can occur to cyclize the intron. The new bond formed with the terminal G is labile and hydrolyzes spontaneously. As a consequence, the intron is relinearized, but it is truncated and thus precluded from the back-splicing reaction.

As explained above, group I (and II) introns are not enzymes because they have a turnover number of only 1. But they can be readily converted into enzymes (ribozymes) in the following way (Box 14-2 Fig. 1): the relinearized intron described above retains its active site. If we provide it with free G and a substrate that includes a sequence complementary to the internal guide sequence, it will repeatedly catalyze cleavage of substrate molecules. We will have converted a group I intron into a ribozyme, similar to the way that the self-cleaving Hammerhead could be converted to a ribozyme by separating the active site from the substrate (Chapter 5). We can go a step further by changing the sequence of the internal guide sequence and thereby generate tailor-made ribonucleases that cleave RNA molecules of our choice.

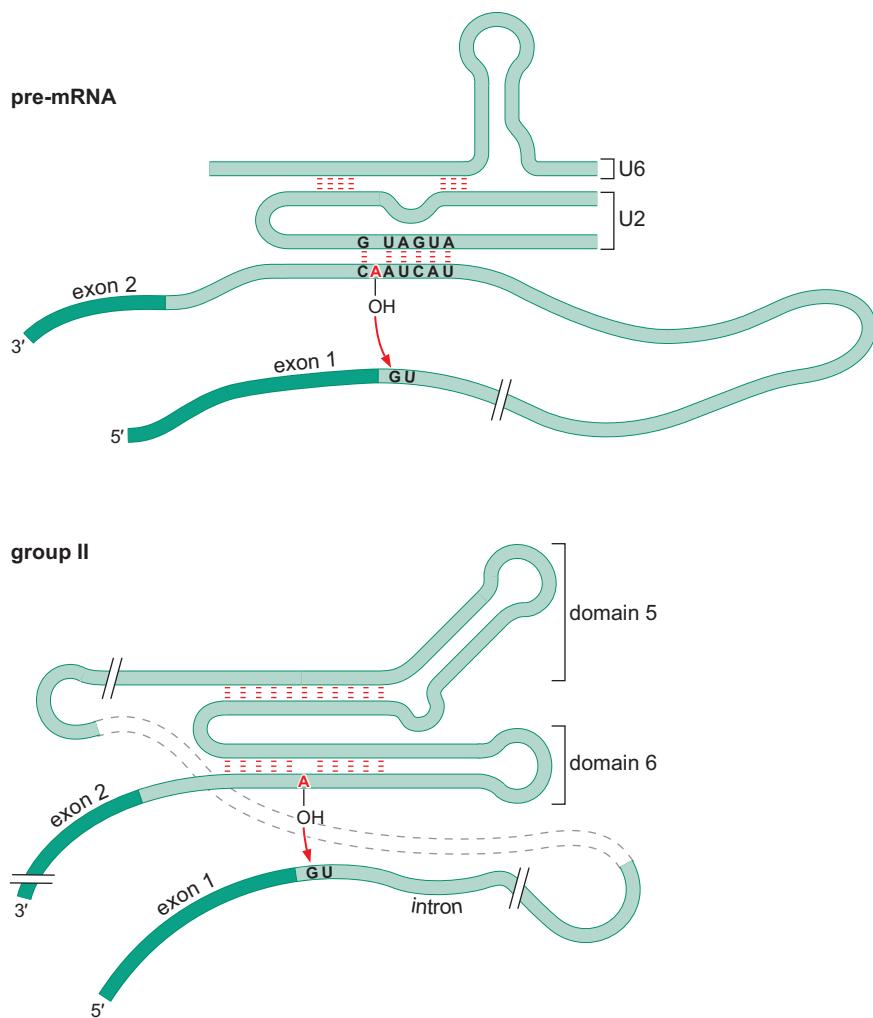


**BOX 14-2 FIGURE 1** Group I introns can be converted into true ribozymes.

of a self-splicing intron is critical for the splicing reaction. This sequence requirement holds because the intron must fold into a precise structure to perform the reaction chemistry. In addition, *in vivo*, the intron is complexed with several proteins that help stabilize the correct structure—partly by shielding regions of the backbone from each other. Thus, the folding requires certain sections of the RNA backbone to be in close proximity to other sections, and the negative charges provided by the phosphates in those backbone regions would repel each other if not shielded. *In vitro*, high salt concentrations (and thus positive ions) compensate for the absence of these proteins. This is how we know that the proteins are not needed for the splicing reaction itself.

The similar chemistry seen in self- and spliceosome-mediated splicing is believed to reflect an evolutionary relationship. Perhaps ancestral group II-like self-splicing introns were the starting point for the evolution of modern pre-mRNA splicing. The catalytic functions provided by the RNA were retained, but the requirement for extensive sequence specificity within

**FIGURE 14-9** Proposed folding of the RNA catalytic regions for splicing of group II introns and pre-mRNAs. The dotted regions of the RNA in the group II case replace an additional four folded domains not shown in this depiction. The proposed striking similarities in these structures have since been confirmed through X-ray crystallographic studies.

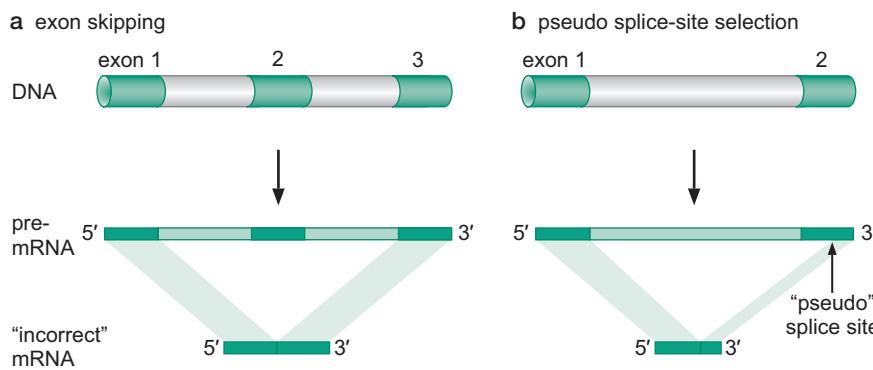


the intron itself was relieved by having the snRNAs and their associated proteins provide most of those functions in *trans*. In this way, introns had only to retain the minimum of sequence elements required to target splicing to the correct places. Thus, many more and varied sizes and sequences of introns were permitted.

The structure of the catalytic region that performs the first transesterification reaction is very similar in the group II intron and the pre-mRNA/snRNP complex (Fig. 14-9). This observation, confirmed through detailed X-ray crystallographic studies in recent years, fuels the broader speculation (discussed in Chapter 17) that early in the evolution of modern organisms, many catalytic functions in the cell were performed by RNAs, and that these functions have, on the whole, since been replaced by proteins. In the case of the spliceosome and the ribosome, however, these activities have not been entirely replaced by proteins. Rather, the vestigial RNA-catalyzed mechanisms remain at the heart of the present complex machinery.

### How Does the Spliceosome Find the Splice Sites Reliably?

We have already seen one mechanism that guards against inappropriate splicing: the active site of the spliceosome is only formed on RNA sequences that pass the test of being recognized by multiple elements during



**FIGURE 14-10** Errors produced by mistakes in splice-site selection. (a) The consequence of skipping an exon. This happens if the spliceosome components bound at the 5' splice site of one exon interact with spliceosome components bound at the 3' splice site of not the next exon, but one beyond. (b) The effect of spliceosome components recognizing pseudo-splice sites—sequences that resemble (but are not) legitimate splice sites. In the case shown, the pseudo-site is within an exon and leads to regions near the 5' end of that exon being mistakenly spliced out along with the intron.

spliceosome assembly. Thus, for example, the 5' splice site must be recognized initially by the U1 snRNP and then by the U6 snRNP. It is unlikely both would recognize an incorrect sequence, and thus selection is stringent. Yet, the problem of appropriate splice-site recognition in the pre-mRNA remains formidable.

Consider the following: the average human gene has seven or eight exons and can be spliced in three alternative forms. But there is one human gene with 363 exons and one *Drosophila* gene that can be spliced in 38,000 alternative ways, a case we describe in detail in the next section. If the snRNPs had to find the correct 5' and 3' splice sites on a complete RNA molecule and bring them together in the correct pairs, unaided, it seems inevitable that many errors would occur. Remember also that the average exon is only some 150 nucleotides long, whereas the average intron is ~3000 nucleotides long (and as we have seen, some introns can be as long as 800,000 nucleotides). Thus, the exons must be identified within a vast ocean of intronic sequences.

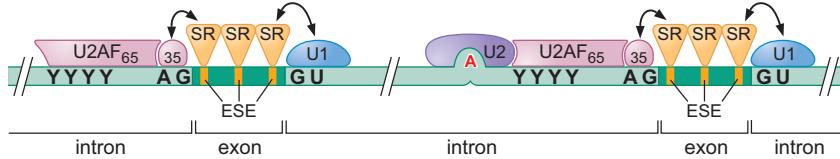
Splice-site recognition is prone to two kinds of errors (Fig. 14-10). First, splice sites can be skipped, with components bound at, for example, a given 5' splice site pairing with those at a 3' site beyond the correct one.

Second, other sites, close in sequence but not legitimate splice sites, could be mistakenly recognized. This is easy to appreciate when one recalls that the splice site consensus sequences are rather loose. Therefore, for example, components at a given 5' splice site might pair with components bound incorrectly at such a “pseudo” 3' splice site (see Fig. 14-10b).

Two ways in which the accuracy of splice-site selection can be enhanced are as follows: first, as we saw in Chapter 13, while transcribing a gene to produce the RNA, RNA polymerase II carries with it various proteins with roles in RNA processing (see Chapter 13, Fig. 13-19). These include proteins involved in splicing. When a 5' splice site is encountered in the newly synthesized RNA, the factors that recognize that site are transferred from the polymerase carboxy-terminal “tail” (that part of the enzyme where they hitch a ride) onto the RNA. Once in place, the 5' splice site components are poised to interact with those other factors that bind to the next 3' splice site to be synthesized. Thus, the correct 3' splice site can be recognized before any competing sites further downstream have been transcribed. This cotranscriptional loading process greatly diminishes the likelihood of exon skipping.

(It is worth noting that even though much of the splicing machinery assembles while the gene is being transcribed, this does not mean the introns are themselves spliced out in that order. Thus, in contrast to many other activities we have heard about—transcription, replication, and so on—there appears to be no “tracking” mechanism involved, whereby the machinery assembles at one end of the RNA and acts as it tracks to the other end.)

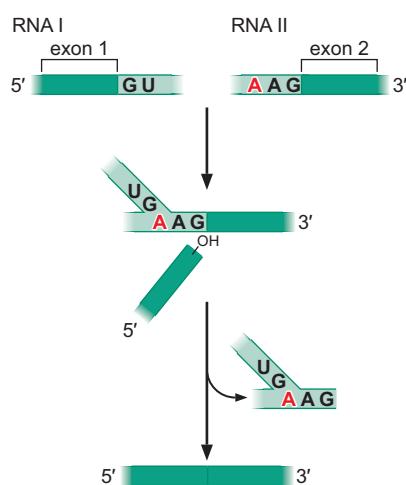
**FIGURE 14-11** SR proteins recruit spliceosome components to the 5' and 3' splice sites. Legitimate splice sites are recognized by the splicing machinery by virtue of being close to exons. Thus, SR proteins bind to sequences within the exons (exonic splicing enhancers, ESEs), and from there recruit U2AF and U1snRNP to the downstream 5' and upstream 3' splice sites, respectively—a process often called exon definition. This initiates the assembly of the splicing machinery on the correct sites, and splicing can proceed as outlined above. (Adapted, with permission, from Maniatis T. and Tasic B. 2002. *Nature* **418**: 236–243. © Macmillan.)



A second mechanism guards against the use of incorrect sites by ensuring that splice sites close to exons (and thus likely to be authentic) are recognized preferentially. So-called SR (serine–arginine-rich) proteins bind to sequences called **exonic splicing enhancers (ESEs)** within the exons. SR proteins bound to these sites recruit the splicing machinery to the nearby splice sites. In this way, the machinery binds more efficiently to those nearby splice sites than to incorrect sites not close to exons. Specifically, the SR proteins recruit the U2AF proteins to the 3' splice site and U1 snRNP to the 5' site (Fig. 14-11). As we saw above, these factors demarcate the splice sites for the rest of the machinery to assemble correctly (Fig. 14-7). This recruitment is either through direct interaction between the SR proteins and proteins within the spliceosome or through interaction with, and stabilization of, RNA:RNA hybrids formed during spliceosome assembly and action.

By recruiting splicing factors to each side of a given exon, this process encourages the so-called “exon definition” we alluded to above when discussing the order of events during spliceosome assembly. That is, spliceosome components are recruited around exons initially, rather than around the intron to be removed. Subsequently, components near one exon will pair with those near an adjacent exon to eliminate the intervening intron.

SR proteins are essential for splicing. They not only ensure the accuracy and efficiency of constitutive splicing (as we have just seen) but also regulate alternative splicing (as we shall see presently). They come in many varieties, some controlled by physiological signals, others constitutively active. Some are expressed preferentially in certain cell types and control splicing in cell-type-specific patterns. We discuss some specific examples of the roles of SR proteins in the section on Alternative Splicing.



**FIGURE 14-12** *Trans-splicing*. In *trans-splicing*, two exons, initially found in two separate RNA molecules, are spliced together into a single mRNA. The chemistry of this reaction is the same as that of the standard splicing reaction described above, and the spliced product is indistinguishable. The only difference is that the other product—the lariat in the standard reaction—is, in *trans-splicing*, a Y-shaped branch structure instead. This is because the initial reaction brings together two RNA molecules rather than forming a loop within a single molecule.

## VARIANTS OF SPLICING

Before turning to alternative splicing, we briefly describe two variants on the splicing machinery and splicing reactions discussed so far. In the first case, we consider examples in which the two exons being joined reside on different RNA molecules, and in the second, we consider a specialized version of the splicing machinery that is used to splice a subset of introns.

### Exons from Different RNA Molecules Can Be Fused by *Trans-Splicing*

In our description of splicing above, we assumed that the 5' splice site of one exon is joined to the 3' splice site of the exon that immediately follows it. This is not always the case. In **alternative splicing**, exons can be deliberately skipped, and a given exon is joined to one further downstream (as we shall see later). In some cases, two exons carried on different RNA molecules can be spliced together in a process called ***trans-splicing***. Although generally rare, *trans-splicing* occurs in almost all of the mRNAs of trypanosomes. In the nematode worm (*Caenorhabditis elegans*), all mRNAs undergo *trans-splicing* (to attach a 5' leader sequence), and many of them undergo *cis-splicing* as well. Figure 14-12 shows how the basic splicing reaction

just described is adapted to perform *trans*-splicing. *Trans*-splicing uses the same spliceosomal machinery as normal *cis*-splicing, except for U1, which, at least in worms, is not needed for *trans*-splicing. We now turn to cases of splicing in which the machinery is quite distinct.

### A Small Group of Introns Is Spliced by an Alternative Spliceosome Composed of a Different Set of snRNPs

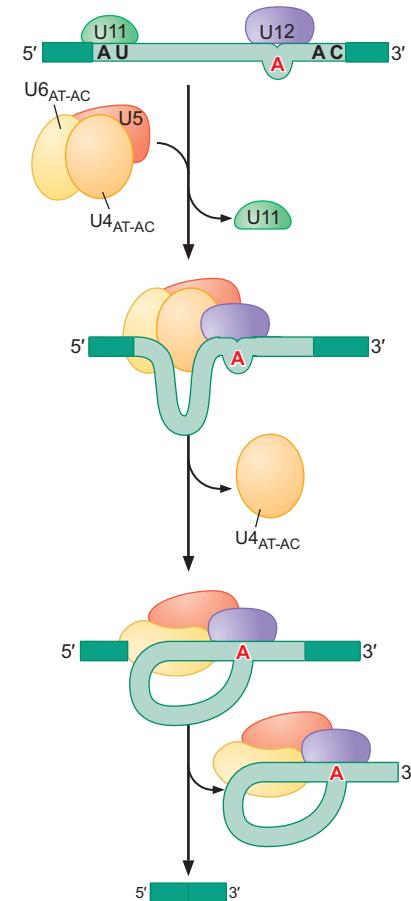
Higher eukaryotes (including mammals, plants, etc.) use the major splicing machinery we have discussed thus far to direct splicing of the majority of their pre-mRNAs. But in these organisms (unlike in yeast), some pre-mRNAs are spliced by an alternative, low-abundance form of the spliceosome. This rare form contains some components common to the major spliceosome, but it contains other unique components as well. Thus, U11 and U12 components of the alternative spliceosome have the same roles in the splicing reaction as U1 and U2 of the major form, but they recognize distinct sequences. U4 and U6 have equivalent counterparts in both spliceosome forms—although these snRNPs are distinct, they share the same names. Finally, the identical U5 component is found in both the major and the alternative—so-called minor—spliceosome.

The minor spliceosome recognizes rarely occurring introns having consensus sequences distinct from the sequences of most pre-mRNA introns. It should be emphasized that although these introns are rare, they are widely distributed—approximately 800 human genes contain at least one minor intron. Furthermore, mutations in minor snRNAs have recently been found to underlie some rare human genetic diseases.

The minor form of the spliceosome is also known as the AT-AC spliceosome, because the termini of the originally identified rare introns contain AU at the 5' splice site and AC at the 3' site (in RNA or AT and AC in DNA). Later it transpired that many introns spliced by this pathway have GT-AG termini (like mainstream introns), but otherwise their consensus sequences are distinct from those of the major pathway.

Despite the different splice site and branch site sequences recognized by the two systems, these major and minor forms of spliceosomes both remove introns using the same chemical pathway (Fig. 14-13). Consistent with this conserved mechanism, the differences in splice-site sequences recognized by these snRNPs are mirrored by complementary differences in the sequences of their snRNAs. Thus, it is the ability of the snRNAs and splice-site sequences to base-pair that is conserved, not any particular sequence within either.

It is also worth noting that AT-AC introns might fit into the evolutionary scheme discussed earlier. As we mentioned, it has been proposed that the group II introns represent the oldest form of introns. Furthermore, it is suggested that the AT-AC introns evolved from the group II introns and eventually gave rise to the major pre-mRNA introns.



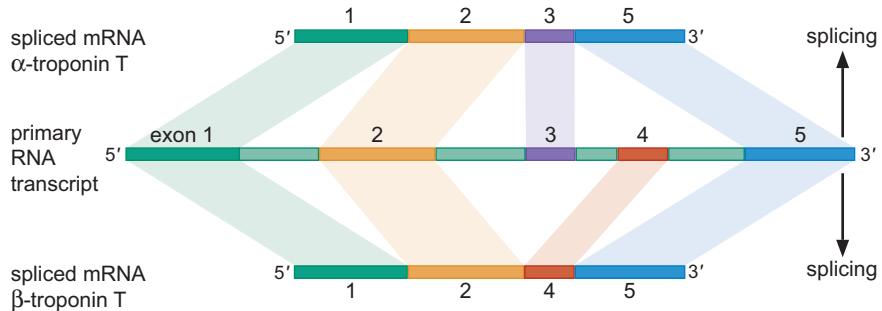
**FIGURE 14-13** The AT-AC (minor) spliceosome catalyzed splicing. This minor spliceosome works on a minority of exons (e.g., perhaps one in 1000 in humans), and those have distinct splice-site sequences. Regardless, the chemistry is the same, and so are some of the spliceosome components, and others are closely related.

## ALTERNATIVE SPlicing

### Single Genes Can Produce Multiple Products by Alternative Splicing

As described in the introduction to this chapter, many genes in higher eukaryotes encode RNAs that can be spliced in alternative ways to generate two or more different mRNAs and thus different protein products (or isoforms). It is now believed that at least 40% of *Drosophila* genes and as many as 90% of human genes undergo alternative splicing. Many alternatively

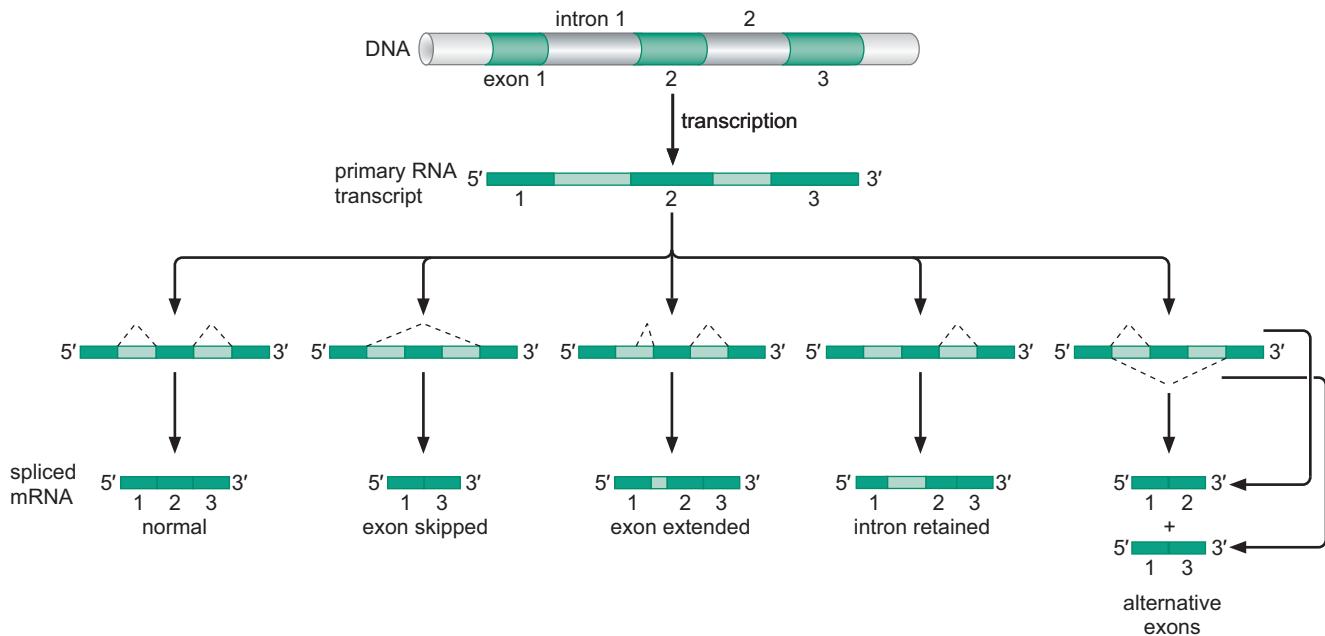
**FIGURE 14-14 Alternative splicing in the troponin T gene.** Shown is a region of the troponin T gene encoding five exons that generates two alternatively spliced forms as indicated. One contains exons 1, 2, 4, and 5; the other contains exons 1, 2, 3, and 5.



spliced genes generate only two alternative products, but in some cases, the number of potential alternatives that can be generated from a single gene is breathtaking—hundreds (e.g., in the human *Slo* gene) or even many thousands (for the *Drosophila Dscam* gene). Alternative splicing is sometimes used as a way of generating diversity, with alternative forms being generated stochastically. But in many cases, the process is regulated to ensure that different protein products are made in different cell types or in response to different conditions.

For a simple case of alternative splicing, consider the gene for the mammalian muscle protein troponin T. Shown in Figure 14-14 is a region of the pre-mRNA made from this gene that contains five exons. This pre-RNA is spliced to form two alternative mature mRNAs, each containing four exons. A different exon is eliminated from each of the two mRNAs, thus the two messages have three exons in common, as well as each carrying one unique exon.

But, as shown in Figure 14-15, alternative splicing can occur in a number of ways. Thus, in addition to alternative exons, exons can be extended (by



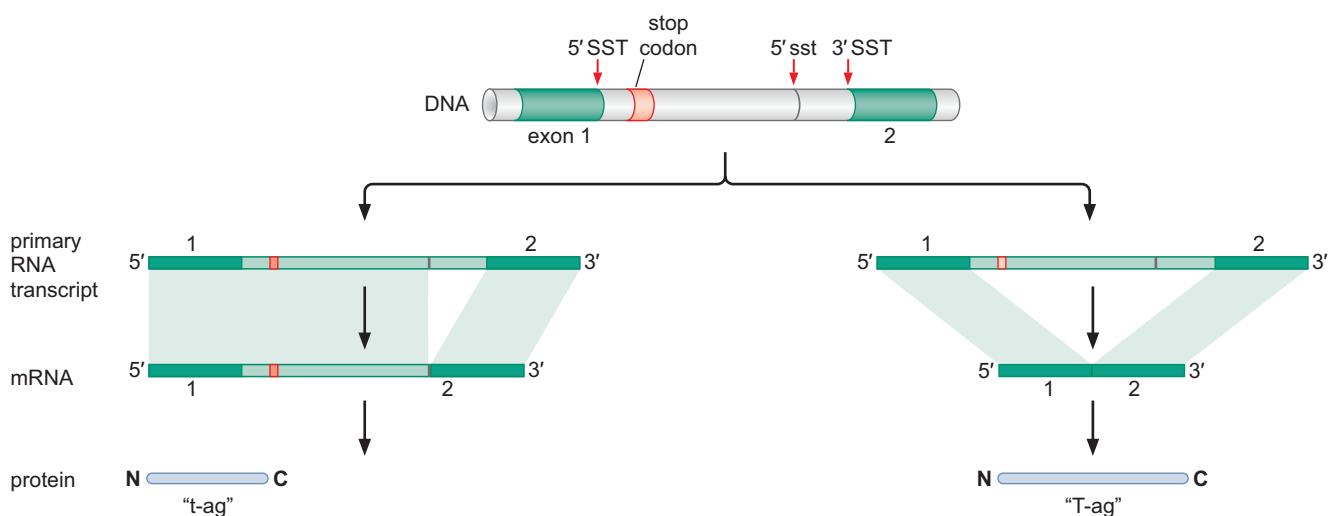
**FIGURE 14-15 Five ways to splice an RNA.** At the top is shown a gene encoding three exons. This is transcribed into a pre-mRNA, shown in the middle, and then spliced by five different alternative pathways. Thus, by including all exons, an mRNA containing all three exons is generated. Exon skipping gives an mRNA containing just exons 1 and 3. By exon extension, part of intron 1 is included together with the three exons. In another case, a complete intron is retained in the mature mRNA. Finally, exons 2 and 3 might be used as alternatives, generating a mixture of mRNAs, each including exon 1 and either exon 2 or 3.

selecting an alternative downstream 5', or upstream 3', splice site). In other cases, exons can be skipped (deliberately), or introns can be retained in the mature message. Some alternative splicing results from transcription of a gene from alternative promoters, allowing one transcript to include a 5' exon not present in the other. Similarly, alternative poly-A sites allow 3' terminal exons to be extended or alternative 3' terminal exons to be used in some transcripts of a given gene. There are even cases of alternative *trans-splicing* (see Fig. 14.12).

In an example of an extended exon, Figure 14-16 shows the case of the T antigen of the monkey virus SV40. The T-antigen gene encodes two protein products: the large T antigen (T-ag) and the small t antigen (t-ag). The two proteins result from alternative splicing of the pre-mRNAs from the same gene. Thus, as shown in Figure 14-16, the gene has two exons, and different mature mRNAs result from the use of two different 5' splice sites. In the mRNA encoding T-ag, exon 1 is spliced directly to exon 2, deleting the intron that lies between. The mRNA for t-ag, on the other hand, is formed using the alternative 5' splice site within the intron. Thus, in this case, the mRNA includes some of the intron as well. (It is therefore an example of the “extended exon” shown in Fig. 14-15.) The reason this larger message encodes the smaller protein is because there is an in-frame stop codon within the region of the intron retained in this mRNA.

Both forms of T antigens are made in a cell infected by SV40 but have different functions. Large T induces transformation and cell cycle reentry, whereas small t blocks the apoptotic response of cells forced down that path. The ratio of the two forms produced differs depending on the level of the splicing regulator SF2/ASF. When present at high levels, this protein directs the machinery to favor use of the 5' splice site that generates more of the t-ag mRNA. SF2/ASF is an SR protein and, when abundant, presumably binds sites within exon 2 and helps the spliceosome assemble there.

In genome-wide studies, the most commonly seen forms of alternative splicing are cases in which complete exons are included or excluded from the mature message. Such exons are often called **cassette exons**. In ~10%



**FIGURE 14-16** Alternative splicing of SV40 T antigen. Splicing of the SV40 T-antigen RNA. Both forms are typically produced, and both proteins are made, upon infection. The small t antigen is encoded by the longer of the two mRNAs; that message contains an in-frame stop codon upstream of exon 2. 5' SST refers to the 5' splice site used to generate the large T mRNA; 5' st refers to the 5' splice site used for small t. 3' SST is the 3' splice site used in generating both mRNAs.

of cases, cassette exons come in pairs, only one of which is included in the spliced message, just as we saw in the case of  $\alpha$ -troponin T (Fig. 14-14). In these cases, there must be mechanisms that ensure that the exons are spliced in a mutually exclusive fashion.

### Several Mechanisms Exist to Ensure Mutually Exclusive Splicing

There are several mechanisms to ensure that selection of alternative exons is mutually exclusive—that is, that when one is chosen, the other is not (or, to refer again to the  $\alpha$ -troponin T example, when exon 3 is chosen, exon 4 is always excluded, and vice versa). We deal with each of these mechanisms here and then, in the next section, discuss an extreme case in which a special mechanism is required.

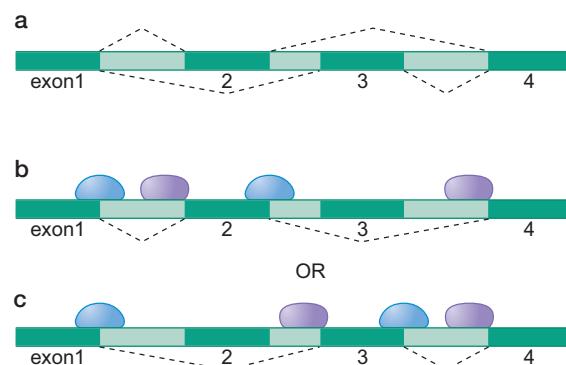
**Steric Hindrance** Consider two alternative exons separated by an intron. If the splice sites within the intron are too close together, splicing factors cannot bind to both sites at the same time. Thus, Figure 14-17 shows a case in which the binding of U1 snRNP to the 5' splice site of the intron between two alternative exons (exons 2 and 3) prevents the binding of U2 snRNP to the branchpoint within that same intron (Fig. 14-17b). Alternatively, binding of U2 snRNP to the branchpoint excludes use of the 5' splice site (Fig. 14-17c). The splicing of exons 3 and 4 of  $\alpha$ -troponin is made mutually exclusive by this mechanism.

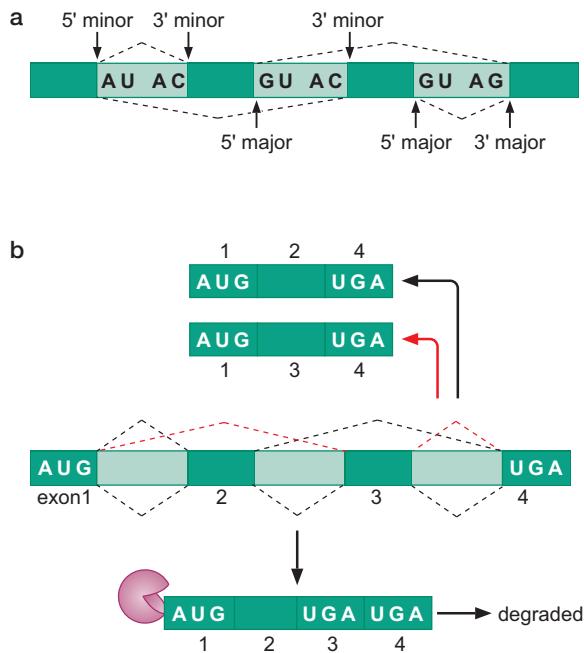
This arrangement can arise through the relative positions of the splice sites within an intron or because the intron is simply too small to work; in *Drosophila*, any intron under 59 nucleotides falls into that category.

**Combinations of Major and Minor Splice Sites** As we saw above, there is a form of the spliceosome called the minor spliceosome that recognizes splice sites distinct from those recognized by the major spliceosome. Neither spliceosome can remove an intron that contains a combination of sites (i.e., a 5' splice site of one type and a 3' of the other). Thus, by judicious arrangement of 5' and 3' splice sites recognized by these alternative spliceosomes, mutual exclusion can be achieved, as shown in Figure 14-18a. The human *JNK1* gene is an example of this.

**Nonsense-Mediated Decay** Rather than forcing the splicing machinery to splice in a mutually exclusive fashion, this mechanism instead ensures that only messages that have one or another exon (never both and never neither) survive. In other words, although not ensuring mutually exclusive splicing, the consequences of this mechanism amount to the same thing. Nonsense-mediated decay (NMD) results from the fact that including both

**FIGURE 14-17** Mutually exclusive splicing: Steric hindrance. (a) This view shows the alternative splicing possibilities. (b) Binding of U1 snRNP to the 5' splice site of the second intron excludes binding of the U2 snRNP to the branchpoint of the same intron; binding of U2 to the following intron results in exclusion of exon 3. (c) Here, binding of the U2 snRNP to the branchpoint of the second intron excludes binding of U1 to the 5' splice site of the same intron. In this case, binding of U1 to the 5' splice site of the first intron results in exclusion of exon 2.



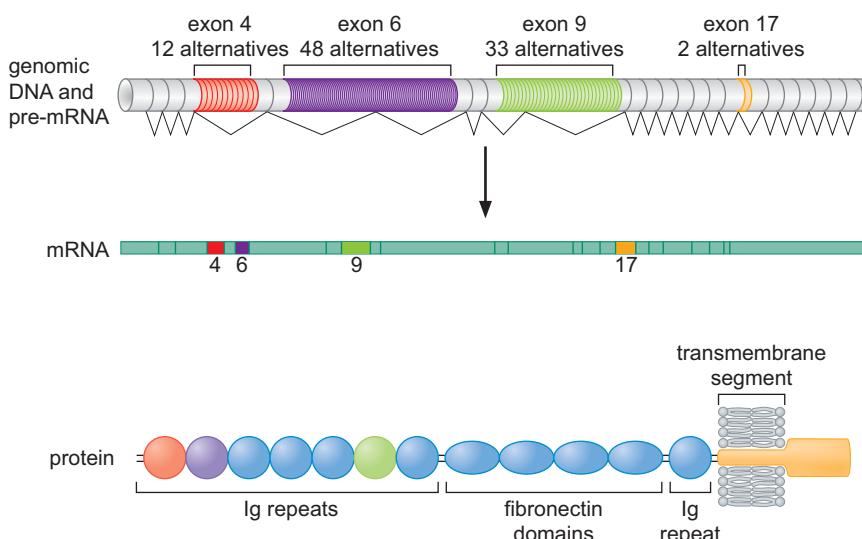


**FIGURE 14-18** Mutually exclusive splicing. (a) Splice sites recognized by the major and minor spliceosomes. (b) Nonsense-mediated decay.

exons produces an mRNA that contains a premature termination codon (Fig. 14-18b). These messages are destroyed by NMD, the details of which are described in Chapter 15 (see Fig. 15-50).

### The Curious Case of the *Drosophila Dscam* Gene: Mutually Exclusive Splicing on a Grand Scale

The *Drosophila Dscam* (Down syndrome cell-adhesion molecule) gene potentially encodes 38,016 protein isoforms. As shown in Figure 14-19, each possible mRNA made from this gene contains 24 exons, 20 of which are always the same, but 4 of which (exons 4, 6, 9, and 17) come in multiple alternative forms in the pre-mRNA. Thus, there are 12 possible versions of exon 4, 48 of exon 6, 33 of exon 9, and two of exon 17. The permutations these allow ( $12 \times 48 \times 33 \times 2$ ) give rise to the huge number of possible forms.



**FIGURE 14-19** The multiple exons of the *Drosophila Dscam* gene. The *Dscam* gene (shown at the top) is 61.2 kb long; once transcribed and spliced, it produces one or more versions of a 7.8-kb, 24-exon mRNA (the figure shows the generic structure of those mRNAs). As shown, there are several mutually exclusive alternatives for exons 4, 6, 9, and 17. Thus, each mRNA will contain one of 12 possible alternatives for exon 4 (in red), one of 48 for exon 6 (purple), one of 33 for exon 9 (green), and one of two for exon 17 (yellow). Exons 4, 6, and 9 encode parts of three Ig domains, depicted in the corresponding colors, and exon 17 encodes the transmembrane domain. If all possible combinations of these exons are used, the *Dscam* gene produces 38,016 different mRNAs and proteins. (Adapted, with permission, from Schmucker D. 2000. *Cell* 101: 671, Fig. 8. © Elsevier.)

The products of this gene are cell-surface proteins of the immunoglobulin (Ig) superfamily. A generic form of the protein is shown at the bottom of Figure 14-19. The molecule has a transmembrane segment (encoded by exon 17, and thus coming in two alternative forms); fibronectin domains that are identical in all isoforms; and Ig domains, parts of three of which are encoded by the highly variable exons 4, 6, and 9. Thus, it is in these Ig domains that the vast majority of the variation from isoform to isoform resides.

The *Dscam* protein has two disparate functions in the fly: it acts in neural patterning in the brain and also recognizes antigens as part of the innate immune system. In its neuronal function, the *Dscam* protein mediates specific cell–cell interactions. Any given isoform of the protein interacts with itself but not with other isoforms. This selectivity is believed to enable a given neurite to distinguish between other neurites it encounters on the basis of their being “self” or “non-self”—that is, derived from the same neuron or from a different one. During neural network formation in the developing brain, neurites exhibit “self-avoidance” behavior: neurites projecting from the same neuron avoid each other. It has been shown *in vivo* that this recognition of self is mediated by homophilic recognition of the particular set of DSCAM isoforms presented on the surface of a given neuron.

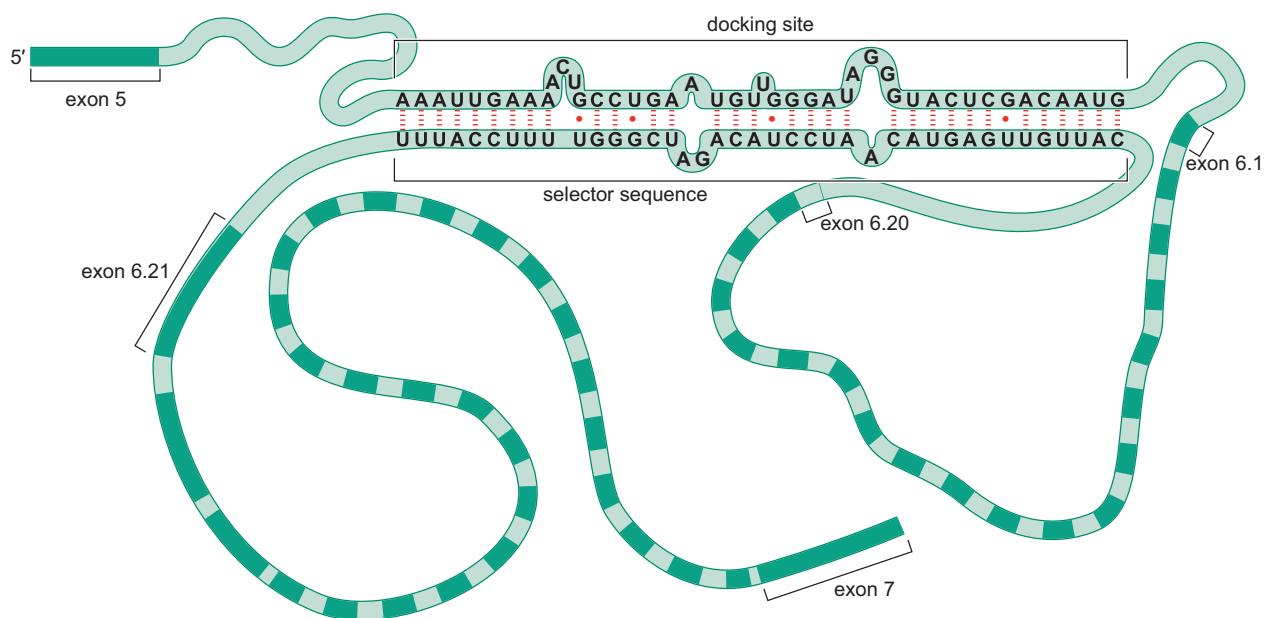
In the immune system, the different isoforms recognize different antigens, much as vertebrate antibodies do. The evolutionary pressure driving diversity is thought to come from selection on this function.

### Mutually Exclusive Splicing of *Dscam* Exon 6 Cannot Be Accounted for by Any Standard Mechanism and Instead Uses a Novel Strategy

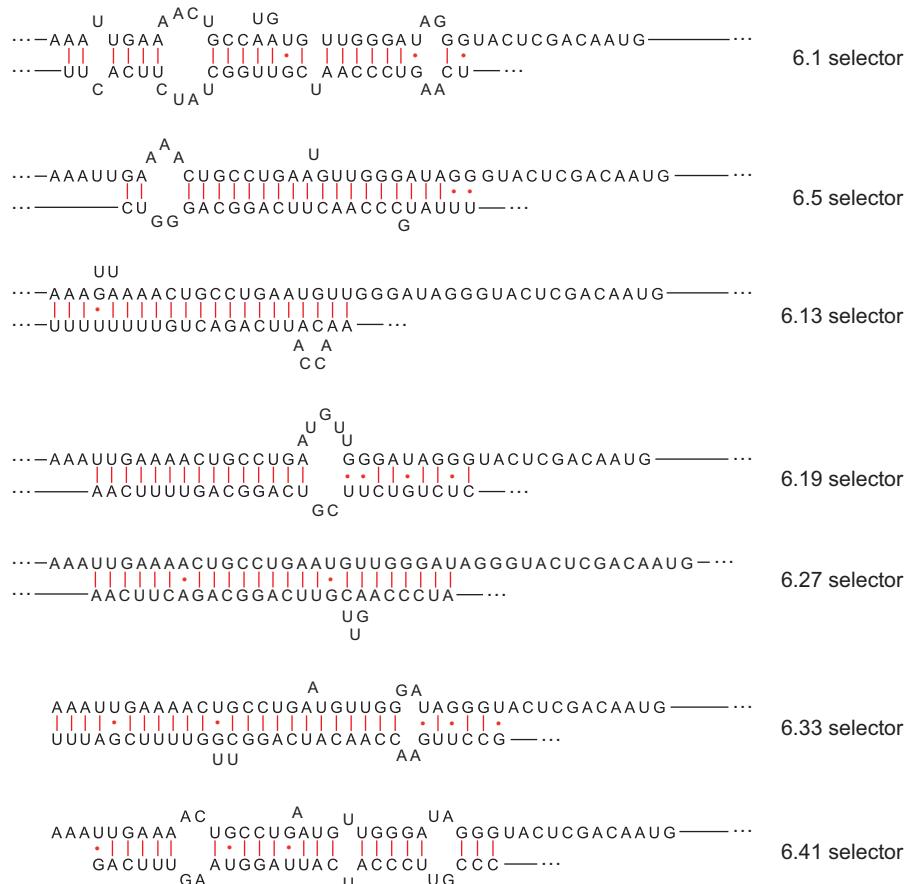
As we have seen, exon 6 is one of the four alternatively spliced exons of the *Dscam* gene—and in this case, there are 48 alternatives from which to choose. The scale of this selection is beyond the scope of the mechanisms discussed above. For example, although steric hindrance could be responsible for adjacent exons not being spliced, it cannot explain how others, further away, could also be excluded. In addition, all of the splice sites in the *Dscam* gene are for the major spliceosome, and thus the dual spliceosome mechanism is not an option. NMD also cannot explain the mutually exclusive splicing of exon 6: even if frameshifts resulted from the inclusion of no, two, or three exons, an mRNA with, say, four exons would have the same reading frame as the message with only one. The same would be true of an mRNA that included seven exons, and so on.

How, therefore, does the cell ensure that only one exon 6 variant is included in the mRNA? The novel mechanism hinges on the formation of alternative RNA:RNA base-paired structures within the pre-mRNA. Each alternative structure ensures that one, and only one, of the exon 6 variants is at any time protected from a general repression of splicing. We now consider how this mechanism works, and how it was discovered through sequence analysis of the *Dscam* gene of *Drosophila* and its various counterparts in other insect species.

The basic model is shown in Figure 14-20. Two classes of conserved sequence element are shown. One, the **docking site**, is located between exon 5 and the first alternative exon 6 variant (exon 6.1). A copy of the second type of element—the **selector sequence**—is found in front of each exon 6 variant (in the figure, exon 6.21 is shown as an example). Each selector sequence is different, but, as shown in Figure 14-21, each can base-pair with the docking site. The regions they each bind in the docking site overlap, and thus binding of the different selector sequences to the one docking site is mutually exclusive—only one selector can bind at a time. The selector



**FIGURE 14-20** The docking site: selector sequences. Base pairing of the selector sequence for exon 6.21 with the docking site. (Design courtesy of Brenton Graveley.)



**FIGURE 14-21** The selector sequences for six exon 6 variants, each bound to the exon 6 docking sequence. As is evident, each selector sequence base-pairs to a slightly different region of the docking sequence, but their binding to that docking sequence is nevertheless mutually exclusive. (Courtesy of Brenton Graveley.)

sequence that *does* bind brings its associated exon 6 variant close to exon 5, ensuring that it is the exon 6 variant chosen.

In addition to bringing the chosen exon 6 variant close to exon 5, the hybridization of the selector sequence and the docking site also ensures that the chosen exon 6 variant is free from a general repression mechanism that inhibits splicing of other possible exon 6 variants. A protein (Hrp36) acts as a general repressor of splicing by coating the other exons and inhibiting their inclusion in that mRNA. This local relief from inhibition afforded by the RNA hybridization may occur either as a direct result of the RNA secondary structure, by the creation of a structure in the RNA recognized by a protein that removes the repressor, or by bringing the chosen exon close to activators within exon 5 that can then overcome the repression.

The docking site and selector sequences were discovered through sequence comparisons in an example of bioinformatic analysis, as described in Box 14-3, Identification of Docking Site and Selector Sequences.

## ► KEY EXPERIMENTS

### Box 14-3 Identification of Docking Site and Selector Sequences

The docking site is 66 nucleotides long in *Drosophila melanogaster*. It is 90%–100% conserved in 10 other *Drosophila* species examined. Even when the comparison includes non-*Drosophila* insect species—mosquito, silkworm, and honeybee, for example—the central 24 nucleotides of the docking site are still very highly conserved. In fact, it is the most conserved sequence in the whole *Dscam* gene (which is more than 60 kb long)! Initial identification of the docking sequence was based entirely on this conservation (Box 14-3 Fig. 1).

The selector sequences were also discovered through sequence comparisons, even though they are less highly con-

served than the docking site. Thus, selector sequences turned up as relatively conserved sequences in the introns upstream of exon 6 variants. An alignment of the 48 selector sequences from the exon 6 variants of *D. melanogaster* revealed a 28-nucleotide consensus sequence that was complementary to the docking sequence (Box 14-3 Fig. 2). When each individual selector sequence was compared with the docking site, each was seen to base-pair with it, each in a unique, but overlapping, manner. Some examples are shown in Figure 14-21.



**BOX 14-3 FIGURE 1** The nucleotide sequence alignment of the docking sites of 15 insects. The insects analyzed include 10 species of *Drosophila*; two of mosquito, *Anopheles gambiae* (malaria mosquito) and *Aedes aegypti* (yellow fever mosquito); the Lepidopteran *Bombyx mori* (silkworm); the Hymenopteran *Apis mellifera* (honeybee); and the Coleopteran *Tribolium castaneum* (red flour beetle). The most common nucleotide at each position is shaded, and the docking site consensus sequence is represented below as a pictogram. The height of each letter represents the frequency of each nucleotide at that position. (Modified, with permission, from Graveley B.R. 2005. *Cell* 123: 65–73, Fig. 2. © Elsevier.)

**Box 14-3** (Continued)

6.24 GTCATTGTCGAGAGCTT-----TACATCCAATAC TCAGGCAGT  
 6.47 GGCTTTCCAGTACCCATTATCAGGTTAGTCAAC-----TCGGGCATAC---CAATTAGACAGAGG  
 6.22 CAGCTCAATCGTATCC-----AATCCCAGCTT TTAGG---AAACACTTAAGATTA  
 6.17 CAGCTGTCAAGGACTT-----G-TCCCGACC TTAGGCAGTAAATCG  
 6.34 TTCAG---CCC-----TTAGACCAACATTCAGGC  
 6.10 GTGGGTTTCCC-----TTTCCCAACATCATCAGACAGTTTT  
 6.4 GGTAAACC-----A-ACCCAACTTTAGGC  
 6.6 GTCAGTCCCT-----TCCCATCTTCAGGC  
 6.38 GGCATTCC-----GGTCCCAGTT TTAGGTTATACAAATTGTTGGTT  
 6.33 G---CC-----TTGAACCAACATTCAGGCTTGTTITCGATTTCCTTTA  
 6.15 TGCCC-----TCCACCAACATTTCAG  
 6.8 TCCTAGGCCAACATTTCAG  
 6.14 GTCGTTTCATTCT-ATCCCAGCATTTCAGATAGTAGATTTT  
 6.35 TCCT-----ATCCCATACATTTCAAATGTCGCCGATAGATT  
 6.9 TACTTTAAATTAATCAAACACATTCAGTCAGTTC-----AATAAGGGA  
 6.1 TCA-----AGTCCCAATCGTTTGGTATCTTCACTTCTTA  
 6.41 CTCAGGCCTTCCCCGTTCCATCATT-----AGGTAAGTTCAGCAA---CAGGCTTCTAGTT  
 6.3 CC-----TATCCCAAATCG-A-AGG-----TTTCT-CTTTCGA  
 6.31 TTGGGAATCAAGTGTCAAT-----TCAGGCAGTTGTATGGAGTTGAGAC  
 6.37 GGTAA-----GCCA-----CATTCCAGCAGTT-----AGTAA  
 6.29 GGTGATTCTGCTCAGA-----CATTCCAGGAGTTTTA-----AGTTATGGCT  
 6.32 TCTTTATGATTCCCACAT-----TCAGACAGTT  
 6.39 TGTGATAAACCCAAATT-----TCAGTCAGTTTCA  
 6.23 GAGTGCCTGGTTGCTTATTCATGTAGTTT-----CAT-----GGTCT  
 6.7 TATCCCTGACCTTCACTT-CGGCAGTTACAAT-----TGAGTTAGG  
 6.42 GCCTTGATCCGGTCTAATCAGGCAGTTTCATAGAGATT  
 6.12 TGCTCAACTTCCACATTGGCAGATTTC  
 6.26 CCCCATCCAATTCCACTCAGGCAGTTTC-ACTAGACTTCGGTT  
 6.2 ACCCAGACCA-----ACTTGCGCAGTTCCAAT-----TGAGATTGCTCGC  
 6.16 CCGCTG-----CCACACTTCAGGTATTCTTAGCAT-----GGC  
 6.19 CTCTGTCCTT-----CGTC-AGGCAGTTTCAAAGTICCTTAGCTGATAGGT  
 6.25 TGTCGAGTT-----CCTGGCAGTTCAAT-----CTCAG-CACGGTT  
 6.20 CATTGCTGAGTACAGTCCAGGCAGGTTTCATG-AGATTGGG  
 6.21 CATTGTTGAGTACATCAGACGGTTTCCATTACATAGAATGTTAGAAGC  
 6.5 TTTATGCCCAACT-----TCAGGCAG-GTCTAGA  
 6.36 ACCCCGCAAGCACATTCAGTCAGTTT-----TGTTTGCCTTAGCT  
 6.13 AACACCAA-----T-----TCAGACTGTTTTTATG  
 6.27 ATCCCAAGTTGTCAGGCAGA-----C-----TTCAAACTGA-----CTT  
 6.28 ATCCTTACGCATTTAGGCAGGTTTCCGTTTACTTAG  
 6.44 ACCCAAACCTTATTCAGCAGTT-----ATTAAGCGAC  
 6.40 ACGCTG-----T-----TCA-ACTGGTTCTGTTAGGGTTCAATAGA  
 6.43 TTAGCATCAGGCAGTTTC  
 6.30 CCACACTGGGGAGTTTCAAT  
 6.48 CCACACTAGGCAGCAAATAGCAT-----TTCAAATAGGATCTTA  
 6.45 GTTCAGGCAG-----CTTAGAAGGC-T  
 6.18 GGGCGTATTCGAA-----TTCAG-GAC  
 6.11 AGGTAGCCAATA-----AGTAGA-----CTTA

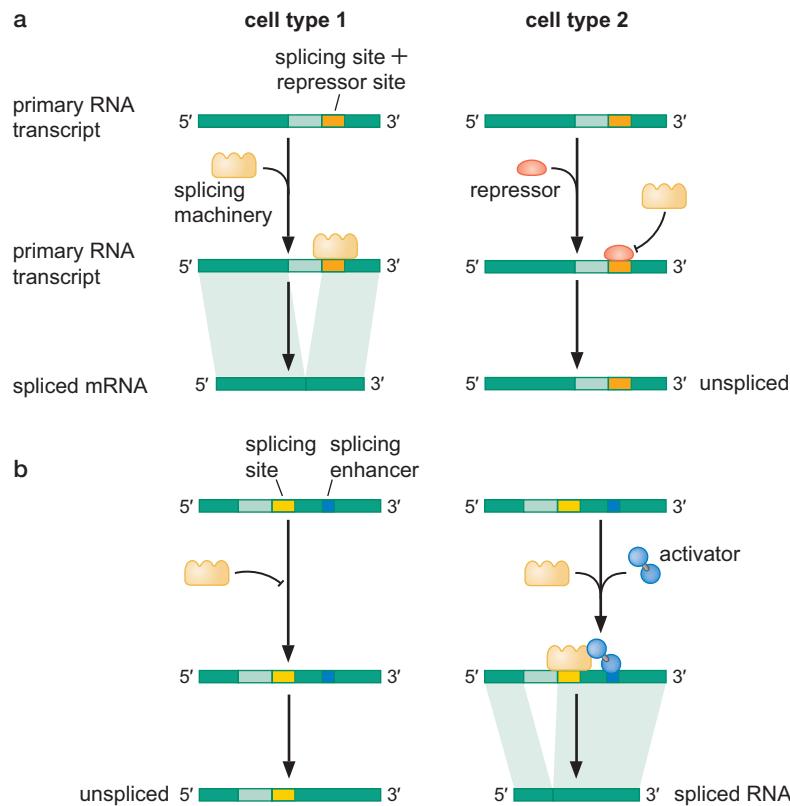


**BOX 14-3 FIGURE 2** *D. melanogaster* selector sequence consensus. (Top panel) The 48 selector sequences and flanking sequence were aligned. The most frequent nucleotides in the central portion of the alignment are highlighted. (Bottom panel) The alignment was used to generate a selector sequence consensus. (Reprinted, with permission, from Graveley B.R. 2005. *Cell* 123: 65–73, Fig. 4. © Elsevier.)

## Alternative Splicing Is Regulated by Activators and Repressors

Proteins that regulate splicing bind to specific sites called **exonic** (or **intronic**) **splicing enhancers** (ESE or ISE) or **silencers** (ESS and ISS). The former enhance, and the latter repress, splicing at nearby splice sites. We have already encountered enhancers and the SR proteins that bind to them (Fig. 14-11). Indeed, these elements and proteins are important in directing the splicing machinery to many exons, even when alternative splicing is not involved. In addition, in the example of T-antigen splicing we described (Fig. 14-16), it was an SR protein that ensured that alternative splicing occurred. But this protein family—which is large and diverse—has specific roles in *regulated* alternative splicing as well, directing the splicing machinery to different splice sites under different conditions. Thus, the presence or activity of a given SR protein can determine whether a particular splice site is used in a particular cell type or at a particular stage of development. Figure 14-22 shows hypothetical cases of regulated splicing by an activator bound to a splicing enhancer and a repressor bound to a splicing silencer.

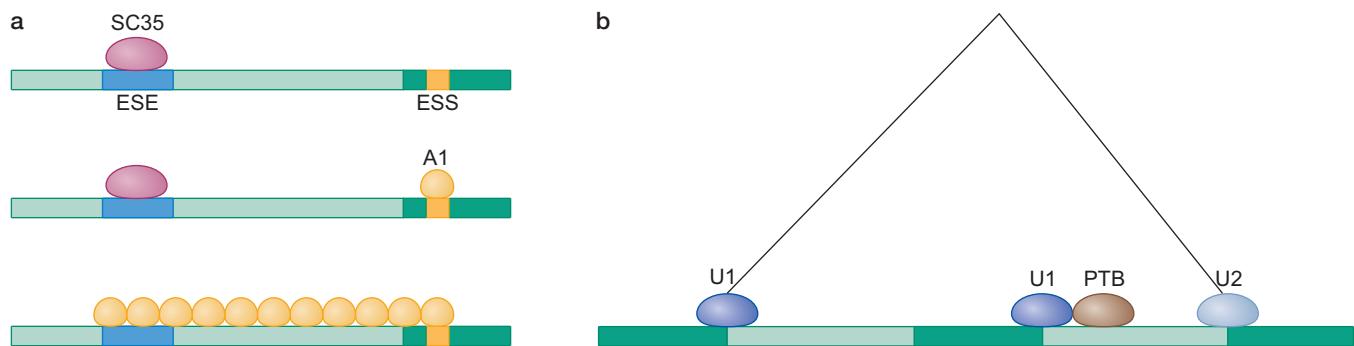
**FIGURE 14-22** Regulated alternative splicing. (a) Some alternatively spliced exons appear in mRNAs unless prevented from doing so by a repressor protein. (b) Others appear only if a specific activator promotes their inclusion. Either mechanism can be used to regulate splicing such that in one cell type a particular exon is included in an mRNA, whereas in another it is not.



The SR proteins bind RNA using one domain—for example, the well-characterized RNA-recognition motif (RRM) described in Chapter 6 (Fig. 6-18). Each SR protein has another domain, rich in arginine and serine, called an **RS domain**. The RS domain, found at the carboxy-terminal end of the protein, mediates interactions between the SR protein and proteins within the splicing machinery, recruiting that machinery to a nearby splice site.

An example of an activator that promotes a particular alternative splicing event in a specific tissue type is the *Drosophila* Half-pint protein. This activator regulates the alternative splicing of a set of pre-mRNAs in the fly ovary. It works by binding to sites near the 3' splice site of specific exons in those pre-mRNAs and recruiting the U2AF splicing factor.

Most silencers are recognized by members of the heterogeneous nuclear ribonucleoprotein (hnRNP) family. These bind RNA but lack the RS domains and thus cannot recruit the splicing machinery. Instead, by blocking specific splice sites, they repress the use of those sites. We saw a function like this in the *Dscam* example earlier where Hrp36 inhibits inclusion of exon 6 variants in the mRNA. Another example is hnRNPA1, which binds to an exonic silencer element within an exon of the human immunodeficiency virus (HIV) *tat* pre-RNA and represses the inclusion of that exon in the final mRNA. By binding to its site, the repressor blocks binding of the activator SC35 (an SR protein) to a nearby enhancer element. In this case, blocking is not direct—the two binding sites do not overlap—but hnRNPA1 promotes cooperative binding of additional molecules of hnRNPA1 to adjacent sequences, spreading over the enhancer site (Figs. 14-23). When present, another SR protein (SF2/ASF) can overcome this repression because it has a higher affinity for the enhancer sequence than does SC35 and therefore displaces the repressors bound there.



**FIGURE 14-23** Two mechanisms of silencer action. (a) Mechanism of HIV tat exon 3 exclusion by hnRNPA1. The splicing activator SC35 binds to the ESE and promotes exon inclusion. A1 binds to the ESS within the exon, and from there it spreads through cooperative binding until it occludes the ESE and competes off SC35 binding. (b) Mechanism of exon exclusion by hnRNPI (PTB) protein. PTB binds within an exon, and interacts with U1 at the 5' splice site, as described in the text. This interaction blocks the ability of U1 to interact with 3' splice site components, and so the U1 at the upstream exon pairs with the U2 at the downstream exon.

We will see similar themes of cooperative and competitive binding in examples of transcriptional regulation in Chapters 18 and 19.

Another mammalian splicing repressor is the hnRNPI (or Py tract binding, PTB) protein. This protein excludes a given exon from the mature mRNA by binding to sequences that flank that exon. But the mechanism by which this operates is not to compete with spliceosome components for binding to splice sites or of splicing activators to their sites. Rather, hnRNPI interacts with the splicing machinery and inhibits its function: after U1 binds to the 5' splice site, hnRNPI interacts with a region of U1 that would otherwise interact with other proteins to facilitate exon pairing. In this way, hnRNPI prevents exon pairing (Figs. 14-23b).

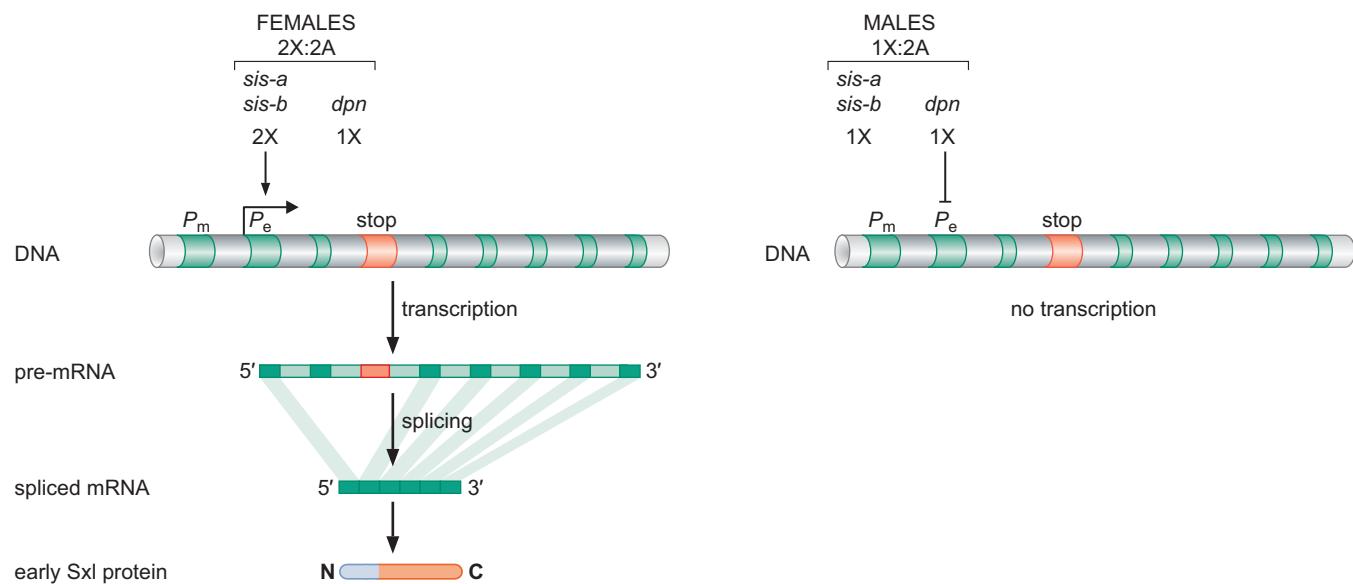
### Regulation of Alternative Splicing Determines the Sex of Flies

We now consider a particularly elaborate example of regulated alternative splicing—that involving the *double-sex* gene of *Drosophila*. The sex of a given fly depends on which of two alternative splicing variants of this mRNA it produces.

The sex of a fly is determined by the ratio of *X* chromosomes to autosomes. A female results from a ratio of 1 (two *X*s and two sets of autosomes) and a male from a ratio of 0.5. This ratio is initially measured at the level of transcriptional regulation using two activators, called SisA and SisB (we consider mechanisms of transcriptional regulation in detail in Chapters 18 and 19). Because the genes encoding these regulators are both on the *X* chromosome, in the early embryo, the prospective female makes twice as much of their products as does the male (Fig. 14-24).

These activators bind to sites in the regulatory sequence upstream of the gene *Sex-lethal* (*Sxl*). Another regulator that binds to and controls the *Sxl* gene is a repressor called Dpn (Deadpan); this is encoded by a gene found on one of the autosomes (chromosome 2). Thus, the ratio of activators to repressor differs in the two sexes, and this accounts for the difference between the *Sxl* gene being activated (in females) and repressed (in males).

The *Sxl* gene is expressed from two promoters,  $P_e$  and  $P_m$ . The former (promoter for establishment) is the one controlled by SisA and SisB (and hence expressed in females only). Subsequently in development, this



**FIGURE 14-24** Early transcriptional regulation of *Sxl* in male and female flies. The *sisA* and *sisB* genes are found on the X chromosome and encode transcriptional activators that control expression of the *Sxl* gene. *Dpn*, a repressor of *Sxl*, is encoded by a gene on chromosome 2. Although both males and females express the same amount of the autosomally encoded *Dpn*, females make twice as much of the activators as males (because females have two X chromosomes and males have only one). The difference in ratio of activators to repressor ensures that *Sxl* is expressed in females but not males. The *Sxl* protein then autoregulates its own expression as described in the text and the next figure. (Adapted from Estes P.A. et al. 1995. *Mol. Cell. Biol.* 15: 904–917.)

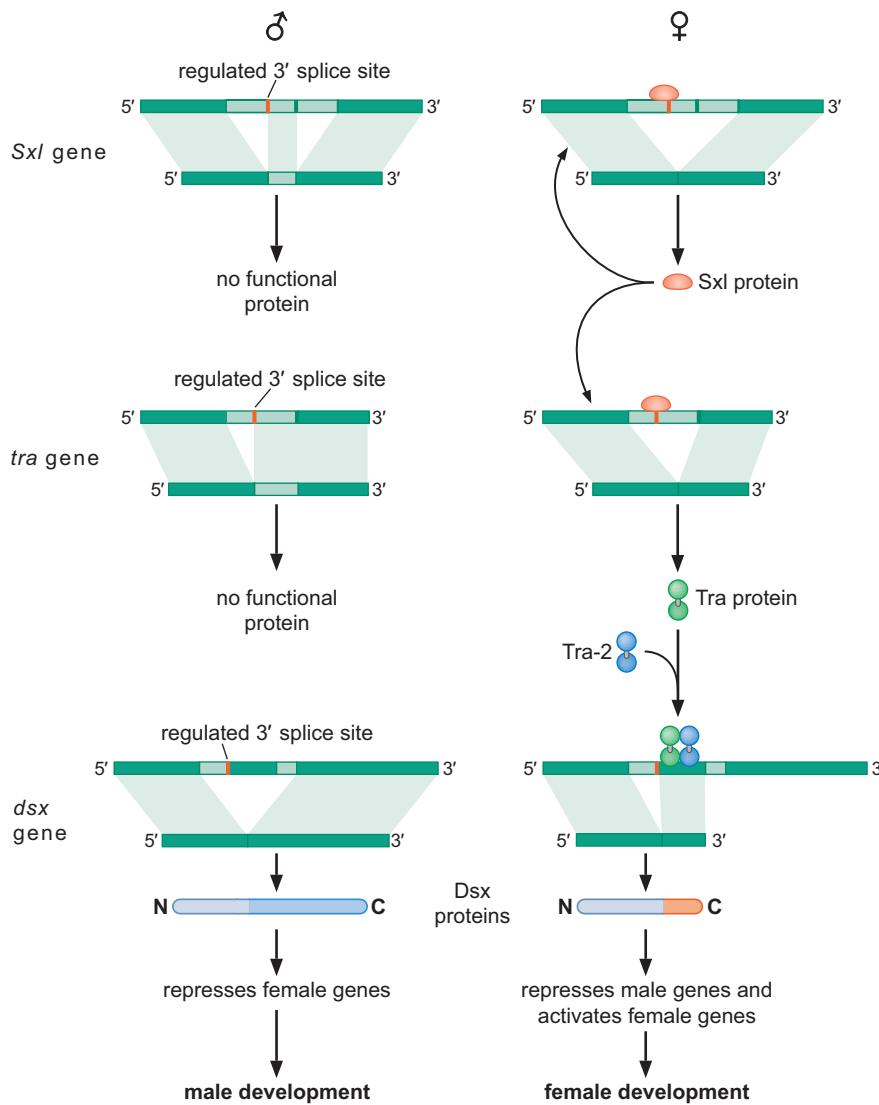
promoter is switched off permanently. In female embryos, expression of *Sxl* is maintained by expression from *P<sub>m</sub>* (promoter for maintenance).

Transcription from *P<sub>m</sub>* is constitutive in both females and males, but the RNA produced from this promoter contains one exon more than the transcript produced from *P<sub>e</sub>* (Fig. 14-23). If that exon remains in the mature message, it fails to produce an active protein, which is what happens in the male. But in the female, splicing removes that exon and functional *Sxl* protein continues to be produced.

As shown in Figure 14-25, it is the *Sxl* protein itself, present in the female but not the male (thanks to earlier expression from *P<sub>e</sub>*), that directs splicing of the RNA made from *P<sub>m</sub>* and ensures that the inhibitory exon is spliced out. *Sxl* does this by working as a splicing repressor.

Functional *Sxl* protein thus continues to be made in females. That protein regulates the splicing of other RNAs in the female as well as its own. One of these is the RNA made constitutively (in males and females) from the *tra* gene (Fig. 14-25). Again, in the absence of *Sxl*-directed splicing, this RNA fails to give protein (in males), but in the presence of *Sxl*, it is spliced to give functional *Tra* protein (in females).

*Tra* protein is also a splicing regulator. Whereas *Sxl* is a splicing repressor, *Tra* is an activator (Fig. 14-25). One of its targets is RNA made from the gene encoding Doublesex (*Dsx*). This RNA is spliced in two alternative forms, both encoding regulatory proteins but with different activities. Thus, in the presence of *Tra*, *dsx* RNA is spliced in a manner that gives rise to a protein that represses expression of male-specific genes. In the absence of *Tra* protein, the form of *Dsx* produced represses female-specific genes.



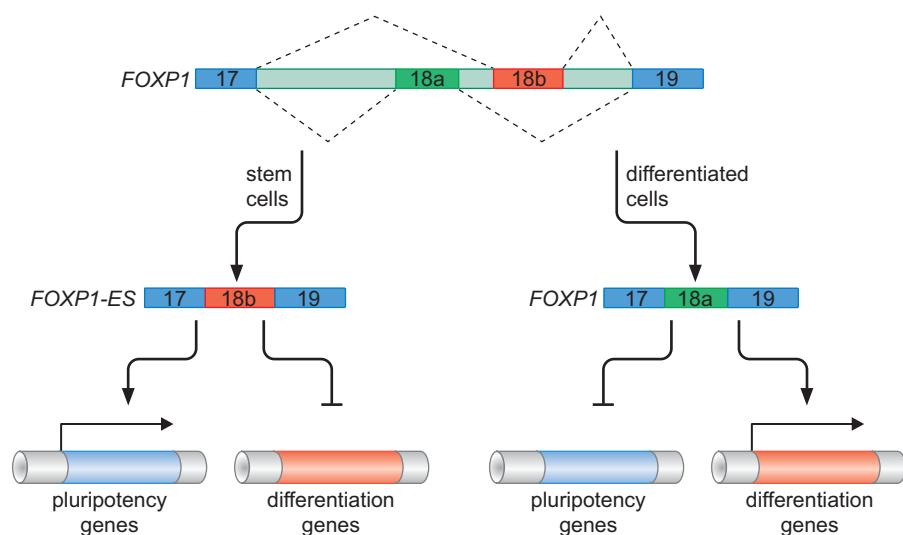
**FIGURE 14-25** A cascade of alternative splicing events determines the sex of a fly. As described in detail in the text, the Sex-lethal protein is produced in flies that will develop into females (shown on the right of the figure) but not those that will develop into males (shown on the left). The presence of that protein is maintained by autoregulation of the splicing of its own message. In the absence of this regulation, no functional protein is produced (in males). Sex-lethal also controls splicing of the *tra* gene, producing functional *Tra* protein in females (but not males). *Tra* is itself a splicing regulator. It acts on pre-mRNA from the *doublesex* gene. When the *dsx* mRNA is spliced in response to *Tra* protein, a version of Doublesex protein is produced (in females) with a stretch of 30 amino acids at its carboxy-terminal end that distinguishes it from the form of the protein produced in the absence of the *Tra* regulator (in males). The female form of *Dsx* activates genes required for female development and represses those for male development. The male form, which has a stretch of 150 amino acids at the carboxy-terminal end, represses genes that direct female development. *Sxl* protein acts as a splicing repressor by binding to the pyrimidine tract at the 3' splice site (see Fig. 14-3). The *Tra* protein, in contrast, acts as a splicing activator. It binds to an enhancer sequence in one of the exons of *dsx* RNA (see Fig. 14-11).

### An Alternative Splicing Switch Lies at the Heart of Pluripotency

No topic has generated more interest in the field of mammalian developmental biology or garnered more attention for potential breakthroughs in medicine than the use of embryonic stem cells to generate specialized cell types in the laboratory. Embryonic stem cells are undifferentiated cells found in the embryo that give rise to the tissues and cell types of the adult animal and hence are said to be pluripotent (see Appendix 1). Indeed, embryonic stem cells can be induced in the laboratory to differentiate into muscle, nerve, pancreatic, and other specialized cell types. It has recently become possible to trigger the dedifferentiation of somatic cells back into pluripotent cells (so-called induced pluripotent stem cells or **iPS cells**) by artificially enhancing the production of a relatively small number of key transcriptional factors known to be required for maintenance of pluripotency (Box 21-1). Examples of such transcription factors are OCT4 and NANOG. The production of OCT4 and NANOG (and other key regulatory proteins) is stimulated by a transcription factor known as FOXP1, a member of the Forkhead family of DNA-binding proteins. There

**FIGURE 14-26** An alternative splicing switch lies at the heart of pluripotency.

The transcription factor FOXP1 binds to specific sites on DNA via a domain known as a Winged helix domain. The specificity of the DNA site it recognizes can be changed by alternative splicing of its mRNA. Thus, when one particular exon is included (exon 18a), the resulting isoform (FOXP1) recognizes sites of one sequence, but when alternative splicing generates a form including exon 18b in place of 18a, the resulting isoform (FOXP1-ES) binds to a different DNA sequence and thus controls expression of a different set of genes.



are in fact two isoforms of the FoxP1 protein. One is the product of an mRNA spliced to include a particular exon, 18b, while the other splicing variant carries exon 18a instead (Fig. 14-26). The protein encoded by the exon 18b–carrying mRNA is called FOXP1-ES: this activates genes (OCT4, NANOG, etc.) that promote dedifferentiation, and thus stimulates iPS cell formation. In contrast, the exon 18a–containing form encodes FOXP1 itself, and this has the opposite effect: it fails to stimulate expression of OCT4 and NANOG and instead activates genes that promote differentiation.

How does the switch from exon 18a to exon 18b explain the switch from the expression of stem-cell-promoting genes to genes promoting differentiation? FOXP1 binds to DNA via a domain known as a Winged helix, which recognizes a particular DNA sequence. The switch from exon 18a to exon 18b substitutes 35 residues in the Winged helix region of the protein including four amino acids that are known to contact DNA. This substitution brings about a change in the DNA sequence specificity of FOXP1. Indeed, the replacement of two of these residues, an asparagine and a histidine with a glycine and a threonine, respectively, can be tied to a particular base-pair difference between the consensus sequence for FOXP1 and that for FOXP1-ES.

Thus, sitting at the top of a hierarchy of events controlling the transcriptional circuits for pluripotency and differentiation is a switch between alternative splice forms of the mRNA for FOXP1. This is reminiscent of alternative splicing of the *sex-lethal* gene, which, as we have seen, sets in motion a chain of events that determines the sex of flies. In the case of *sex-lethal*, it is known how the ratio of X chromosomes to autosomes dictates which splice variant is produced in males and which in females. In the case of FOXP1, however, the factors that determine which form is produced in embryonic stem cells and which in differentiating cells await discovery. Nevertheless, these examples from flies and mammals underscore the role of alternative splicing in a wide and growing variety of gene regulatory events.

We have considered the various ways in which splicing is performed in eukaryotic systems as well as the diversity of components involved. The loss of function in any of these components may lead to serious consequences, as we describe in Box 14-4, Defects in Pre-mRNA Splicing Cause Human Disease.

## EXON SHUFFLING

### Exons Are Shuffled by Recombination to Produce Genes Encoding New Proteins

As we have noted, all eukaryotes have introns, and yet these elements are rare—almost non-existent—in bacteria. There are two likely explanations for this situation.

First, in the so-called **introns early model**, introns existed in all organisms but were lost from bacteria. If introns originally did exist in bacteria, why might they subsequently have been lost? The argument is that these

#### MEDICAL CONNECTIONS

##### Box 14-4 Defects in Pre-mRNA Splicing Cause Human Disease

As discussed in the text, the vast majority of human genes contain introns. Indeed, the large majority of human genes contain multiple introns. It is therefore not surprising that many point mutations that cause disease in humans turn out to be nucleotide substitutions that impair pre-mRNA splicing. In fact, estimates indicate that at least 15% of all point mutations that cause human disease alter recognition sequences for splicing. A classic example is β-thalassemia. This human genetic disorder is characterized by a defect in the production of β-globin, a subunit of hemoglobin. One kind of β-thalassemia is caused by a mutation in the first intron of the β-globin gene that changes the sequence TTGGT to TTAGT. This mutation creates a sequence that resembles a normal 3' splice site (Py tract AG/G) (see Fig. 14-3). As a result, the splicing of β-globin pre-mRNA in afflicted individuals predominantly occurs at the mutationally created 3' splice site rather than at the normal site.

An inherited disease known as “familial isolated growth hormone deficiency type II” is caused by a defect in the splicing of the pre-mRNA for growth hormone, resulting in individuals who are short in stature. Frasier syndrome is a urogenital disorder that is attributed to a defect in pre-mRNA splicing for a gene known to be important for kidney and gonad development. Two additional examples are a kind of dementia that is due to a splicing defect in the mRNA for a cytoskeleton protein and a form of cystic fibrosis.

Yet other disorders are caused by mutations that impair the splicing machinery itself. One example is retinitis pigmentosa, which is characterized by progressive degeneration of the retina and eventually blindness. Mutations at many genes cause retinitis pigmentosa, and most of these genes have retina-specific functions. Some of these mutations, however, are in genes for components of the spliceosome. Because afflicted individuals have one normal copy of the gene as well as the mutant copy, the splicing protein is produced but is present in lower than normal amounts.

Why is the effect of a lower than normal level of a splicing component manifest in a specific tissue, the retina? One possible explanation stems from the fact that the photopigment rhodopsin of the retina undergoes a high level of turnover. Thus, the splicing machinery must meet the very high demand for opsin (the

protein component of rhodopsin) production to replace that lost from degradation. Hence, the retina might be more sensitive to a partial impairment of splicing than other tissues that do not have the burden of producing a specific protein at high levels.

Spinal muscular atrophy (SMA) is one of the most common genetic causes of mortality in children. The disease, which is characterized by the progressive loss of spinal neurons, results from a mutation in a gene for a ubiquitous component of the splicing machinery known as SMN (survival motor neuron), whose precise function is not well understood. As in the case of retinitis pigmentosa, we are left with the mystery of why the effect of the splicing defect is principally manifest in motor neurons.

The examples considered thus far are inherited disorders. But disease-causing mutations that impair splicing also arise somatically. An example comes from mutants of the gene for the cell cycle regulator p73. The p73 protein exists in multiple forms as a result of alternative splicing of its mRNA. Mutations that cause faulty alternative splicing of the p73 pre-mRNA have been implicated in a kind of cancer known as squamous cell carcinoma. It is likely that somatic mutations that impair splicing or cause faulty alternative splicing for many other pre-mRNAs also contribute to the etiology of cancer.

It is sometimes said that medicine is the greatest teacher of biology. Certainly, and as we have seen, this adage aptly applies to the field of pre-mRNA splicing, where the study of human genetic disorders has provided a wealth of insights into the sequences that govern splicing and the machinery that performs it. Indeed, the very discovery of snRNPs, the most fundamental components of the pre-mRNA splicing machinery, arose from studies of a form of the autoimmune disease lupus in which afflicted individuals produce antibodies against these ribonuclear protein particles. It seems likely that the continued study of human genetic disorders will lead to additional insights into the mechanisms of pre-mRNA maturation.

#### Correcting Splicing Defects as a Way of Treating Diseases

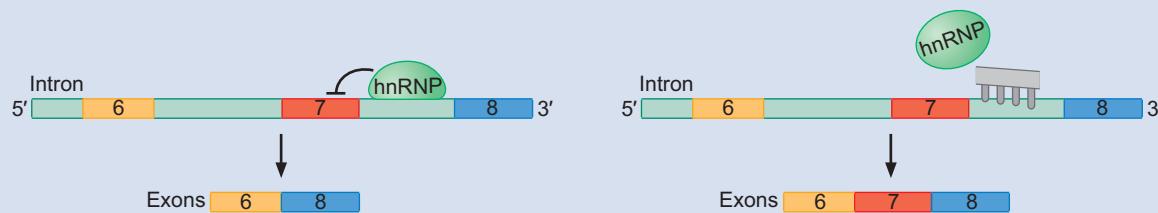
Attempts are underway to treat a number of these diseases by correcting their causative splicing defects. Here we outline the approach taken with one of them, SMA. As we described

**Box 14-4 (Continued)**

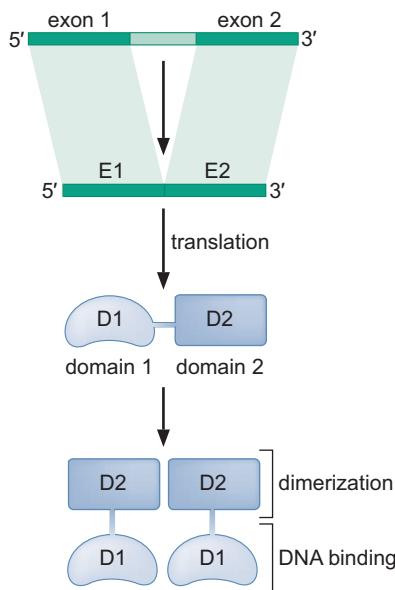
above, SMA is the result of mutations that eliminate the component of the splicing machinery encoded by the *SMN1* gene. Normally, the *SMN1* gene undergoes alternative splicing such that 90% of its mature mRNA includes a particular exon, exon 7, required to generate full-length protein. Humans have another gene called *SMN2* that encodes an identical protein. This protein could potentially replace the function of *SMN1*, but the *SMN2* transcripts are spliced such that only about 10% of the mRNA derived from this gene includes the required exon 7. It was thus decided that one way to compensate for the complete loss of *SMN1* might be to manipulate splicing of

*SMN2* such that it could be forced to produce mRNAs that include exon 7.

One approach is to direct antisense oligonucleotides against sequences within the primary *SMN2* transcript that regulate splicing of exon 7. As shown in the figure (Box 14-4 Fig. 1), the oligo was designed such that it recognizes a specific sequence within the transcript through Watson–Crick base pairing and, in so doing, excludes the binding of a splicing repressor, hnRNP. This strategy was shown to work in mice engineered to express the human *SMN2* gene and is now being pursued in human clinical trials.



**BOX 14-4 FIGURE 1** The *SMN2* splicing modulation. The single-stranded antisense oligonucleotide is designed to bind to the region of the RNA transcript where normally the splicing repressor hnRNP binds to block inclusion of exon 7. By binding there, the oligo displaces hnRNP and allows inclusion of exon 7 in the mature mRNA, and thus it facilitates production of the SMN protein. (Adapted, with permission, from Rigo F. et al. 2012. *J. Cell Biol.* **199**: 21–25, Fig. 2, p. 23. doi:10.1083/jcb.2012087. © Rockefeller University Press.)



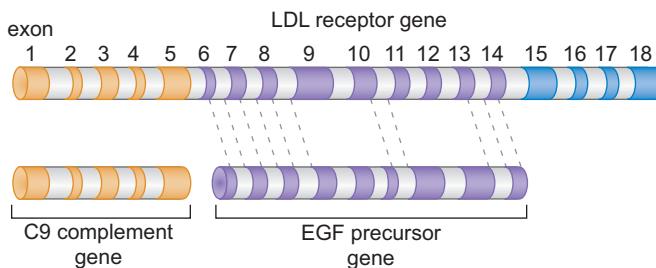
**FIGURE 14-27** Exons encode protein domains. In this example, the DNA-binding domain of a protein is encoded by one exon, whereas the dimerization domain of that same protein is encoded by a separate exon. Protein domains fold independently of the rest of the protein in which they are found and often perform a single function (as we discussed in Chapter 6). Thus, exons can often be exchanged between proteins productively.

“gene-rich” organisms (see Chapters 8 and 12) have streamlined their genomes in response to selective pressure to increase the rate of chromosome replication and cell division. (Recall also that among eukaryotes, the yeast, which are unicellular and rapidly growing, have far fewer introns than do complex multicellular organisms.)

In the alternative view, introns never existed in bacteria but, rather, arose later in evolution. According to this so-called **introns late model**, introns were inserted into genes that previously had no introns, perhaps by a transposon-like mechanism (see Chapter 12).

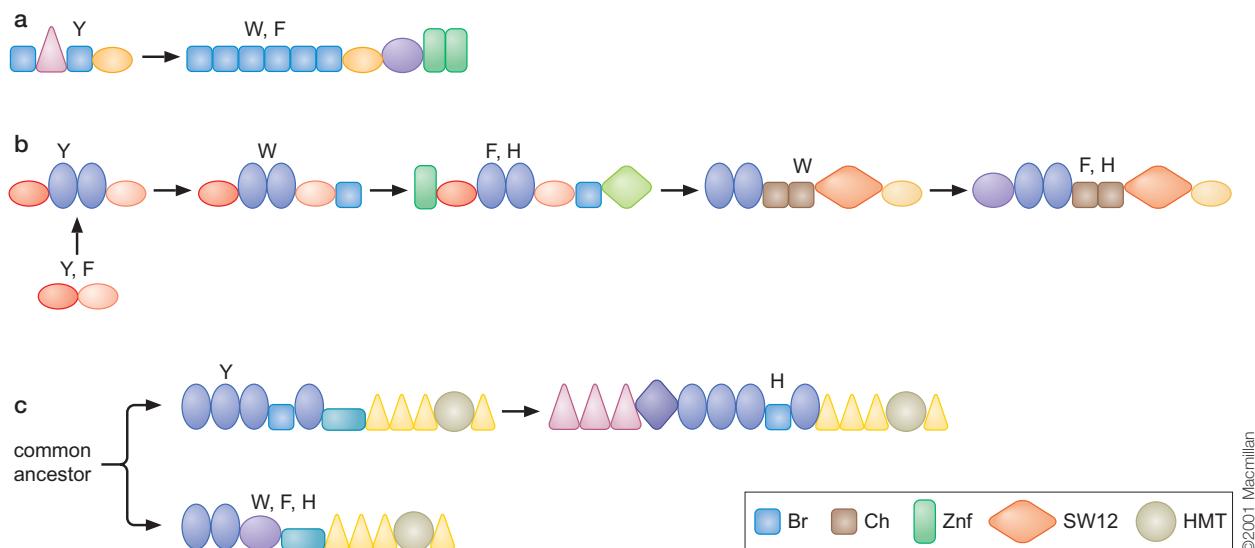
Irrespective of which explanation is true—and at this stage, it is impossible to decide the matter unambiguously—there is the second, perhaps more interesting, question: why have introns been retained in eukaryotes and, in particular, in the extensive form seen in multicellular eukaryotes? One clear advantage is that the presence of introns, and the need to remove them, allows for alternative splicing, which can generate multiple protein products from a single gene. But, on an even grander scale, there is possibly another advantage afforded these organisms: having the coding sequence of genes divided into several exons allows new genes to be created by reshuffling exons. Three observations strongly suggest that this process actually occurs.

- First, the borders between exons and introns within a given gene often coincide with the boundaries between domains (see Chapter 6) within the protein encoded by that gene. That is, it seems that each exon very often encodes an independently folding unit of protein (often corresponding to an independent function). For example, consider the DNA-binding protein depicted in Figure 14-27. Like most DNA-binding proteins, this one has two domains—the DNA recognition domain and the dimerization domain. As shown in the figure, these domains (D1 and D2) are encoded by separate exons (E1 and E2) within the gene.



- Second, many genes, and the proteins they encode, have apparently arisen during evolution in part via exon duplication and divergence. Proteins made up of repeating units (such as immunoglobulins) have probably arisen this way (see Chapter 12, Fig. 12-35). The presence of introns between each exon makes the duplication more likely.
- Third, related exons are sometimes found in otherwise unrelated genes. That is, there is evidence that exons really have been reused in genes encoding different proteins. As an example, consider the low-density lipoprotein (LDL) receptor gene (Fig. 14-28). This gene contains some exons that are clearly evolutionarily related to exons found in the gene encoding the epidermal growth factor (EGF) precursor. At the same time, it has other exons that are clearly related to exons from the C9 complement gene (Fig. 14-28). More extensive examples of exon accretion are apparent from the complete sequences of genomes. As shown in Figure 14-29, there are

**FIGURE 14-28 Genes made up of parts of other genes.** The LDL receptor (the plasma low-density lipoprotein receptor) gene contains a stretch of six exons closely related to six exons from the C9 complement gene and eight closely related to eight from the EGF (epidermal growth factor) precursor gene. Thus, the LDL receptor gene is made up of exons shuffled between other genes; although not shown here, these same parts appear in yet other genes as well. The introns are, in many cases, not positioned in exactly the same positions within the EGF precursor gene and the comparable region of the LDL receptor gene. When they are in the same place, this is indicated by dotted lines.



©2001 Macmillan

**FIGURE 14-29 Accumulation, loss, and reshuffling of domains during the evolution of a family of proteins.** Proposed routes whereby different related proteins might have evolved by gain and loss of specific domains are shown. Three examples are given; in each case, the proteins in question are chromatin-modifying enzymes (Chapter 8) from yeast (Y), worms (W), flies (F), and humans (H). Each protein is depicted by a series of differently colored and shaped domains, and above each protein is shown the organism(s) in which proteins are found containing the domain arrangement shown. Some arrangements are found in more than one organism, and in some cases, a given organism has more than one related arrangement of similar domains. A few of the domains—those whose functions we discussed in Chapters 8 and 19—are identified and are as follows: bromodomain (Br); chromodomain (Ch); a histone methyltransferase domain (HMT); an ATPase activity associated with chromatin-remodeling enzymes (SW12); and a zinc finger domain (Znf). (Adapted, with permission, from Lander et al. 2001. *Nature* **409**: 906, Fig. 42. © Macmillan.)

numerous examples of proteins made up of highly related domains used in various combinations, encoded by genes made up of shuffled exons.

As we have seen, exons tend to be rather short (~150 nucleotides or so), whereas introns vary in length and can be very long, indeed (up to several hundred kilobases). The size ratio ensures that, for the average gene in a higher eukaryote, recombination is more likely to occur within the introns than within the exons. Thus, exons are more likely to be reshuffled than disrupted. The mechanism of splicing—the use of the 5' and 3' splice sites—guarantees that almost all recombinant genes will be expressed, because the splice sites in different genes are largely interchangeable. In addition, alternative splicing can allow new exons to be tried without discarding the original gene product—that is, both the new and old products can be made initially.

## RNA EDITING

---

### RNA Editing Is Another Way of Altering the Sequence of an mRNA

**RNA editing**, like RNA splicing, can change the sequence of an RNA after it has been transcribed. Thus, the protein produced upon translation is different from that predicted from the gene sequence, but this example is perhaps even more dramatic than the case of splicing—instead of stretches of the mRNA being reassorted, during editing, individual bases are either inserted, deleted, or changed. That is, the coding information in the RNA is altered. There are two mechanisms that mediate editing: site-specific deamination of adenines or cytosines, and guide RNA–directed uridine insertion or deletion. We consider each in turn.

In one form of site-specific **deamination**, a specifically targeted cytosine residue within mRNA is converted into uridine by deamination. For a given mRNA species that undergoes editing, that process typically occurs only in certain tissues or cell types and in a regulated manner. Figure 14-30 shows the mammalian apolipoprotein B gene. This gene has several exons, within one of which is a particular CAA codon that is targeted for editing; it is the C within this codon that gets deaminated. That deamination, performed by the enzyme **cytidine deaminase**, converts the C to a U (Fig. 14-31). In this example, the deamination occurs in a tissue-specific manner: messages are edited in intestinal cells but not in liver cells.

The CAA codon, which is translated as glutamine in the unedited message in the liver, is thus converted to UAA—a stop codon—in the intestine. The result is that the full-length protein (of some 4500 amino acids) is produced in the liver, but a truncated polypeptide of only about 2100 amino acids is made in the intestine (see Fig. 14-30).

The two forms of apolipoprotein B are both involved in lipid metabolism. The longer form, found in the liver, is involved in the transport of endogenously synthesized cholesterol and triglycerides. The smaller version, found in the intestines, is involved in the transport of dietary lipids to various tissues.

Other examples of mRNA editing by enzymatic deamination include adenosine deamination. This reaction, performed by the enzyme **ADAR** (**adenosine deaminase acting on RNA**)—of which there are three in humans—produces inosine. Inosine can base-pair with cytosine, and thus this change can readily alter the sequence of the protein encoded by the mRNA. An ion channel expressed in mammalian brains is the target of this type of editing. A single edit in its mRNA elicits a single-amino-acid



**FIGURE 14-30** RNA editing by deamination. The RNA made from the human apolipoprotein gene is edited in a tissue-specific manner by deamination of a specific cytidine to generate a uridine. This event occurs in RNAs in the intestine, but not in those found in the liver. The result, as described in the text, is that a stop codon introduced into the intestinal mRNA generates a shorter protein than that produced in the liver. The figure is not drawn to scale: Thus the edited exon is exon 26, and the codon marked as filling it is in reality only a very short part of that exon.

change in the protein, which, in turn, alters the  $\text{Ca}^{2+}$  permeability of the channel. In the absence of this editing, brain development is seriously impaired.

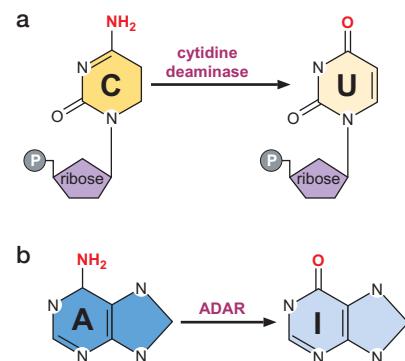
This type of editing—enzymatic deamination—seems to be quite rare, but important. In *Drosophila*, it has been estimated that there may be as few as 20 cytosines targeted for deamination, but all are in genes involved in neurotransmitter production or activity.

It is not entirely clear how the deaminase enzymes work so specifically: their active sites could act on any cytosine. Often the enzymes are part of complexes in which other components might influence the enzymes' specificity of action. In addition, in the case of the cytosine deaminase that works on apolipoprotein-B, the enzyme bears an RNA-binding domain that helps recognize the specific site for deamination by recognizing either a specific sequence or perhaps a particular secondary structure in the RNA.

Another role for deaminase enzymes in the cellular defenses against HIV infection is described in Box 14-5, Deaminases and HIV.

### Guide RNAs Direct the Insertion and Deletion of Uridines

A very different form of RNA editing is found in the RNA transcripts that encode proteins in the mitochondria of trypanosomes. In this case, multiple Us are inserted into specific regions of mRNAs after transcription (or, in other cases, Us may be deleted). These insertions can be so extensive that, in an extreme case, they amount to as many as half the nucleotides of the mature mRNA. The addition of Us to the message changes codons and reading frames, completely altering the “meaning” of the message. As an example, consider the trypanosome *coxII* gene. In a specific region of the mRNA of this gene, four Us are inserted between adjacent bases at three sites

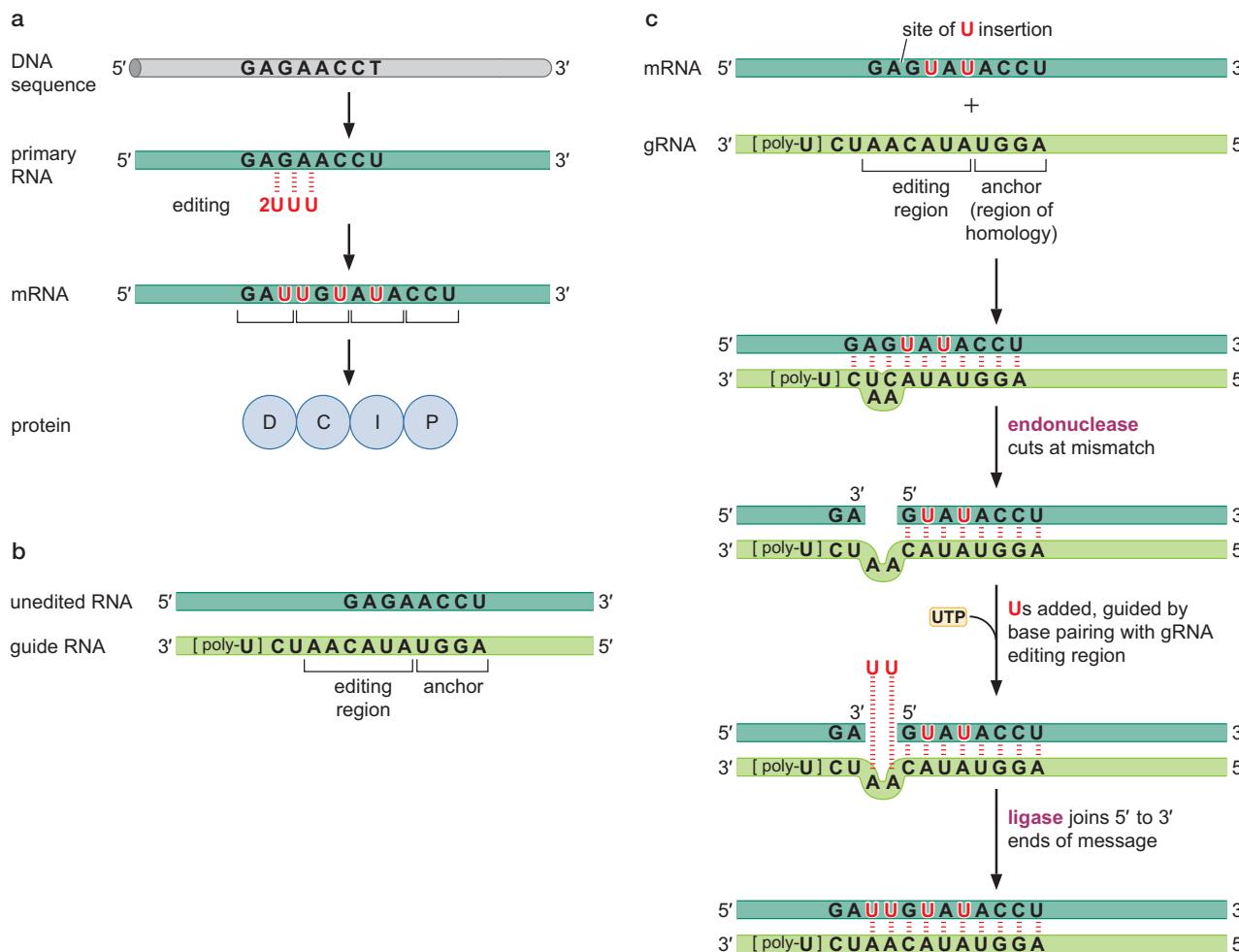


**FIGURE 14-31** The deamination of cytosine and adenine to produce uracil and inosine. (a) The amino group on the nucleotide ring is removed by the cytidine deaminase enzyme. (b) In the case of adenine deamination, the same chemical group is removed from the adenine by ADAR to generate inosine.

(two Us at one site and one U at each of two additional sites). These additions alter some codons and cause a “-1” change in the reading frame, a shift that is required to generate the *correct* open reading frame, as shown in Figure 14-32a.

How are these additional bases inserted? Us are inserted into the message by so-called **guide RNAs (gRNAs)**, as shown in Figure 14-32. These gRNAs range from 40 to 80 nucleotides in length and are encoded by genes distinct from those that encode the mRNAs on which they act. Each gRNA is divided into three regions. The first, at the 5' end, is called the anchor and directs the gRNA to the region of the mRNA it will edit; the second determines exactly where the Us will be inserted within the edited sequence; and the third, at the 3' end, is a poly-U stretch. We now look more closely at how the gRNAs direct editing.

The anchor region of the gRNA contains a sequence that can base-pair with a region of the message immediately beside (3' to) the region that will be edited (Fig. 14-32b). This is followed by the editing “instructions”: a stretch of gRNA complementary to the region in the message to be edited



**FIGURE 14-32** RNA editing by guide RNA-mediated U insertion. Shown is the editing of the trypanosome *coxII* gene RNA. (a) The positions of the four U nucleotides inserted into the pre-mRNA of the *coxII* gene. These generate the correct reading frame and coding information in the mRNA. (b) The sequence of the guide RNA that determines the U insertion pattern, and the sequence of the unedited stretch of mRNA. (c) The editing reaction itself.

## MEDICAL CONNECTIONS

### Box 14-5 Deaminases and HIV

Deamination of the human *apolipoprotein B* mRNA described in the text is undertaken by an enzyme called APOBEC1 (apolipoprotein B-editing enzyme, catalytic polypeptide-like 1). This is a member of a family of enzymes that directs deamination of cytidines in both RNA and DNA. Another member of the family—APOBEC3G (A3G)—is a potent inhibitor of infection by a range of retroviruses, including HIV.

As viral particles, retroviruses such as HIV carry RNA genomes. Upon infection, the RNA is converted to a cDNA copy by reverse transcription (see Chapter 12). It is the minus strand of the cDNA produced during reverse transcription that is attacked by the A3G enzyme. The enzyme deaminates Cs to produce Us in that DNA strand, leading to hypermutation at levels the virus cannot accommodate, or even to destruction of the damaged strand by DNA glycosylase and apurinic-apyrimidinic endonuclease (Chapter 10).

To counter this, wild-type HIV produces a protein called Vif (viral infectivity factor) that directs proteosomal degradation of the A3G enzyme, thereby excluding it from viral particles and protecting the virus in its next round of infection. Vif is required by the virus to grow in all of its biologically relevant target cells *in vivo*. Some cell lines used in laboratories to grow virus were found to support growth of HIV lacking the Vif function. Such cells were called **permissive**. Heterokaryons (cells made by fusing two other cell types) made from permissive and nonpermissive cells had the nonpermissive character. This revealed that the nonpermissive cells make a factor that countered viral replication. That factor was shown to be deaminase A3G, and permissive cells could be made nonpermissive simply by expressing A3G.

but containing additional As. The As are at positions in the gRNA opposite where Us will be inserted into the mRNA. At the 3' end of the gRNA is the poly-U region. The role of the nucleotides in this region is unclear, although it is proposed that they tether the gRNA to purine-rich sequences in the mRNA upstream of (5' to) the edited region.

As shown in Figure 14-32c, the gRNA and mRNA form an RNA–RNA duplex with looped-out single-stranded regions opposite where Us will be inserted. An endonuclease recognizes and cuts the mRNA opposite these loops. Editing involves the transfer of Us into the gap in the message. This process is catalyzed by the enzyme 3' terminal uridylyl transferase (TUTase).

After the addition of Us, the two halves of the mRNA are joined by an RNA ligase, and the “editing” region of the gRNA continues its action along the mRNA in a 3'-to-5' direction. A single gRNA can be responsible for inserting several Us at different sites (as is the case for the one shown in Fig. 14-32). Furthermore, in some cases, several different gRNAs work on different regions of the same message.

## mRNA TRANSPORT

### Once Processed, mRNA Is Packaged and Exported from the Nucleus into the Cytoplasm for Translation

Once it has been fully processed—capped, spliced, and polyadenylated—an mRNA is transported out of the nucleus and into the cytoplasm (Fig. 14-33), where it is translated to give its protein product (Chapter 15). Movement from the nucleus to the cytoplasm is not a passive process. Indeed, it must be carefully regulated: the fully processed mRNAs represent only a small proportion of the RNA found in the nucleus, and many of the other RNAs would be detrimental to the cell if exported. These include, for example, damaged or misprocessed RNAs and liberated introns (which, being, as they tend to be, so much larger than the exons, represent a larger population of RNA than do the mature mRNAs).

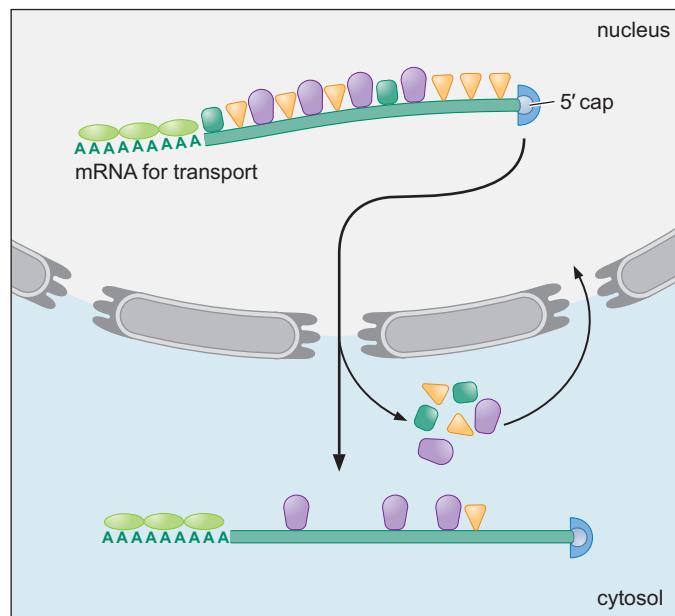
How are RNAs selected and transported? As we have emphasized in this and the previous chapter, from the moment an RNA molecule starts to be transcribed, it becomes associated with proteins of various sorts: initially proteins involved in capping, then splicing factors, and finally the proteins that mediate polyadenylation. Some of these proteins are replaced at various steps along the processing path, but others (including, e.g., some SR proteins—the serine–arginine-rich splicing regulators) (Fig. 14-11) are not; moreover, additional proteins join. As a result, a typical mature mRNA carries a collection of proteins that identifies it as being mRNA destined for transport. Other RNAs not only lack the particular signature collection required for transport, but have their own alternative sets of proteins that actively block export. Thus, for example, excised introns will often carry hnRNPs (repressors of splicing seen above), and these probably mark such an RNA for nuclear retention and destruction.

In addition to residual SR proteins, mature mRNAs carry another group of proteins that bind specifically to exon–exon junctions (which are only found in spliced species, of course). The mRNAs do also carry some hnRNPs, but fewer than are typically bound to introns, and also in a different context, of course. This emphasizes the fact that it is the set of proteins, not any individual kind of protein, that marks RNAs for either export or retention in the nucleus.

Export takes place through a special structure in the nuclear membrane called the **nuclear pore complex**. Small molecules—those under ~50 kDa—can pass through these pores unaided, but larger molecules and complexes, including mRNAs and their associated proteins, require active transport. (Other molecules—proteins made in the cytoplasm but with functions in the nucleus, for example—are transported in the other direction, from the cytoplasm into the nucleus, through these same pores.)

The mechanisms of nuclear transport are beyond the scope of this volume. Suffice it to say that some of the proteins associated with the RNA carry nuclear export signals that are recognized by export receptors that guide the RNA out through the pore. Once in the cytoplasm, the proteins are discarded and are then recognized for import back into the nucleus, where they associate with another mRNA and repeat the cycle (Fig. 14-33).

**FIGURE 14-33** Transport of mRNAs out of the nucleus. RNA export from the nucleus is an active process, and only certain (appropriate) RNAs are selected for transport. To be selected for transport, the RNA must have the correct collection of proteins bound to it. These will distinguish it from other RNAs, which must be retained in the nucleus or destroyed. Proteins that recognize exon:exon boundaries, for example, indicate that an mRNA that has been appropriately spliced, whereas proteins that bind introns indicate that an RNA that should be retained in the nucleus. Once in the cytosol, some proteins are shed and others are taken on in readiness for translation (Chapter 15).



Export requires energy, and this is supplied by hydrolysis of GTP by a GTPase protein called Ran. Like other GTPases, Ran exists in two conformations depending on whether complexed with GTP or GDP, and the transition from one state to the other drives movement into or out of the nucleus.

## SUMMARY

---

In almost all bacterial and phage genes, the open reading frame is a single stretch of codons with no break. But the coding sequence of many eukaryotic genes is split into stretches of codons interrupted by stretches of non-coding sequences.

The coding stretches in these split genes are called exons (for “expressed sequences”), and the non-coding stretches are called introns (for “intervening sequences”). Some non-coding regions are also included in mature mRNAs—5' and 3' untranslated regions of mRNA and entire non-coding RNAs such as microRNAs (Chapter 20). Such regions are therefore also classified as exons. The numbers and sizes of the introns and exons vary enormously from gene to gene. Thus, in yeast, only a relatively small proportion of genes have introns, and where they occur, they tend to be short and few in number (one or occasionally two per gene). In multicellular organisms such as humans, the number of genes containing introns is much larger, as is the number of introns per gene (up to 362 in an extreme case). The sizes of exons do vary but are often ~150 nucleotides; introns, on the other hand, vary from 61 bp to as much as 800 kb.

When a gene containing introns is transcribed, the RNA initially contains those introns. These are then removed to produce the mature mRNA. The process of intron removal is called splicing.

Many intron-containing genes give rise to a unique mRNA species. That is, in each case, all of the introns are removed from the original RNA, leaving an mRNA composed of all of the exons. But in other cases, splicing can produce several different mRNAs from the same gene by splicing the original RNA in different patterns. Thus, for example, some genes contain alternative versions of some of their exons, and only one of these variants ends up in a given mRNA. In other cases, a given exon might be removed (along with the introns) from some copies of the RNA—again producing alternative versions of mRNA from the same gene. We considered in detail one extreme example of alternative splicing—the *Dscam* gene of *Drosophila*. In this case, one of its exons comes in 48 variants, all of which are found in the pre-mRNA, but only one of which (different in any given case) is found in each mRNA.

Sequences found at the boundary between introns and exons allow the cell to identify introns for removal. These splicing sequences are almost exclusively within the introns (where there are no restrictions imposed by the need to encode amino acids, as there are in exons). These sequences are called the 3' and 5' splice sites, denoting their relative locations at one end of the intron or the other end. To splice out an intron also requires a sequence element, called the branch site, near the 3' end of the intron.

Intron removal proceeds via two transesterification reactions. In the first, an A in the branch site attacks a G in the 5' splice site. In the second, the liberated 5' exon attacks the 3' splice site. These reactions have two consequences. First and foremost, they fuse the two exons. Second, they release the intron in the form of a branched structure called a lariat.

Splicing of nuclear pre-mRNAs requires a large complex of proteins and RNAs called the spliceosome. This is made up of so-called snRNPs, of which there are five—U1, U2, U4, U5, and U6 snRNPs. Each of these comprises an RNA molecule, called the U1 to U6 snRNA, respectively, and a number of proteins, the majority of which are different in each case. The RNA components have a central role in recognizing introns and catalyzing their removal. The spliceosome is a very dynamic structure. That is, at different steps during the process of splicing, the spliceosome constitution alters—different subunits of the machine join and leave the complex, each performing a particular function.

The dynamic nature of the spliceosome is such that the order of events is not always exactly the same. Thus, some complexes can come and go in a somewhat different order than described in the canonical pathway of spliceosome assembly, and steps can occasionally reverse. This is likely a result of the kinetics that drives each step varying on different RNAs and under different circumstances. After catalyzing the splicing reaction, the spliceosome is disassembled, which is important for ensuring that the splicing reactions are not reversed and the spliced mRNA is readily released.

There are a few rare introns that can remove themselves from within RNA molecules by a process known as self-splicing. Although not strictly an enzymatic reaction, the RNA of the intron nevertheless mediates the chemistry of removal. These self-splicing introns come in two classes, one of which (group II) splices by the same chemical pathway as that mediated by the spliceosome. These introns probably represent the evolutionary origin of modern spliceosomal introns, and the two-step chemical pathway used by both reflects that evolutionary relationship (and perhaps explains why the spliceosome does not remove introns by a more direct single-step mechanism).

The splice sites are defined by rather short sequences with low levels of conservation. It thus represents a significant challenge for the splicing machinery to recognize and splice only at correct sites. There are various mechanisms by which the spliceosome enhances accuracy. First, it assembles on the sites soon after they have been synthesized. This ensures that they are selected before other downstream sites are available to compete. Second, there are other proteins—SR proteins—

that bind near legitimate splice sites and help recruit the splicing machinery to those sites. In this way, authentic sites effectively have a higher affinity for the machinery than do so-called pseudo-sites of similar sequence.

There are a large variety of SR proteins. Each binds RNA with one surface and stimulates binding of the splicing machinery with another. Some SR proteins regulate splicing. That is, a given SR protein may be found only in one cell type and mediate a particular splicing event only in that cell type. Other SR proteins are only active in the presence of specific physiological signals, and thus a given splicing event only occurs in response to that signal. In this way, SR proteins resemble transcriptional activators, as we shall see in subsequent chapters. In addition, analogous to transcriptional regulation, there are repressors of splicing that exclude splicing of specific introns under certain circumstances or in other cases interact with components of the spliceosome and disrupt assembly. We considered in detail the case of sex determination in *Drosophila*,

*sophila*, where a cascade of regulated, alternative splicing events determines whether the fly develops as a male or female.

Together with the other processing events dealt with in Chapter 13, splicing is required before mRNAs can be transported out of the nucleus through nuclear pores.

It is believed that a given exon typically encodes an independently folding (and functional) protein domain. Thus, such an exon can readily function in combination with other different exons. This suggests that it has been relatively easy, through evolution, to generate new proteins by shuffling existing exons between genes.

RNA editing is another mechanism that allows an RNA to be changed after transcription so as to encode a different protein from that encoded by the gene. Two mechanisms for editing are enzymatic modification of bases and the insertion or deletion of multiple U nucleotides within the message.

## BIBLIOGRAPHY

### Books

- Atkins J.F., Gesteland R.F., and Cech T.R. 2011. *RNA worlds: From life's origins to diversity in gene regulation*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

### Mechanisms of Splicing and the Spliceosome

- Hoskins A.A. and Moore M.J. 2012. The spliceosome: A flexible, reversible macromolecular machine. *Trends Biochem. Sci.* **27**: 179–188.
- Newman A.J. and Nagai K. 2010. Structural studies of the spliceosome: Blind men and an elephant. *Curr. Opin. Struct. Biol.* **20**: 82–89.
- Rino J. and Carmo-Fonseca M. 2009. The spliceosome: A self-organized macromolecular machine in the nucleus? *Trends Cell Biol.* **19**: 375–384.
- Semlow D.R. and Staley J.P. 2012. Staying on message: Ensuring fidelity in pre-mRNA splicing. *Trends Biochem. Sci.* **37**: 263–273.
- Wachtel C. and Manley J.L. 2009. Splicing of mRNA precursors: The role of RNAs and proteins in catalysis. *Mol. Biosyst.* **5**: 311–316.
- Wahl M.C., Will C.L., and Lührmann R. 2009. The spliceosome: Design principles of a dynamic RNP machine. *Cell* **136**: 701–718.

### Self-Splicing

- Lambowitz A.M. and Zimmerly S. 2011. Group II introns: Mobile ribozymes that invade DNA. *Cold Spring Harb. Perspect. Biol.* **3**: a003616.
- Michel F., Costa M., and Westhof E. 2009. The ribozyme core of group II introns: A structure in want of partners. *Trends Biochem. Sci.* **34**: 189–199.
- Toor N., Keating K.S., and Pyle A.M. 2009. Structural insights into RNA splicing. *Curr. Opin. Struct. Biol.* **19**: 260–266.

### Alternative Splicing and Regulation

- Heyd F. and Lynch K.W. 2011. Degrade, move, regroup: Signaling control of splicing proteins. *Trends Biochem. Sci.* **36**: 397–404.
- Kalsotra A. and Cooper T.A. 2011. Functional consequences of developmentally regulated alternative splicing. *Nat. Rev. Genet.* **12**: 715–729.
- Li Q., Lee J.A., and Black D.L. 2007. Neuronal regulation of alternative pre-mRNA splicing. *Nat. Rev. Neurosci.* **8**: 819–831.

Maniatis T. and Tasic B. 2002. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418**: 236–243.

McManus C.J. and Gravelley B.R. 2011. RNA structure and the mechanisms of alternative splicing. *Curr. Opin. Genet. Dev.* **21**: 373–379.

Muñoz M.J., de la Mata M., and Kornblith A.R. 2010. The carboxy terminal domain of RNA polymerase II and alternative splicing. *Trends Biochem. Sci.* **35**: 497–504.

Nilsen T.W. and Gravelley B.R. 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**: 457–463.

Pawlicki J.M. and Steitz J.A. 2010. Nuclear networking fashions pre-messenger RNA and primary microRNA transcripts for function. *Trends Cell Biol.* **20**: 52–61.

### mRNA Transport

Dreyfuss G., Kim V.N., and Kataoka N. 2002. Messenger-RNA-binding proteins and the messages they carry. *Nat. Rev. Mol. Cell Biol.* **3**: 195–205.

Stewart M. 2010. Nuclear export of mRNA. *Trends Biochem. Sci.* **35**: 609–617.

### Evolution

Irimia M., Penny D., and Roy S.W. 2007. Coevolution of genomic intron number and splice sites. *Trends Genet.* **23**: 321–325.

Koonin E.V. 2009. Intron-dominated genomes of early ancestors of eukaryotes. *J. Hered.* **100**: 618–623.

Roy S.W. and Gilbert W. 2006. The evolution of spliceosomal introns: Patterns, puzzles and progress. *Nat. Rev. Genet.* **7**: 211–221.

### RNA Editing

Aphasizhev R. and Aphasizheva I. 2011. Uridine insertion/deletion editing in trypanosomes: A playground for RNA-guided information transfer. *Wiley Interdiscip. Rev. RNA* **2**: 669–685.

Blanc V and Davidson N.O. 2010. APOBEC-1-mediated RNA editing. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **2**: 594–602.

Chiu Y.L. and Greene W.C. 2008. The APOBEC3 cytidine deaminases: An innate defensive network opposing exogenous retroviruses and endogenous retroelements. *Annu. Rev. Immunol.* **26**: 317–353.

- Nishikura K. 2010. Functions and regulation of RNA editing by ADAR deaminases. *Annu. Rev. Biochem.* **79**: 321–349.
- Paro S., Li X., O'Connell M.A., and Keegan L.P. 2012. Regulation and functions of ADAR in *Drosophila*. *Curr. Top. Microbiol. Immunol.* **353**: 221–236.
- Rosenthal J.J. and Seeburg P.H. 2012. A-to-I RNA editing: Effects on proteins key to neural excitability. *Neuron* **74**: 432–439.
- Stuart K.D., Schnaufer A., Ernst N.L., and Panigrahi A.K. 2004. Complex management: RNA editing in trypanosomes. *Trends Biochem. Sci.* **30**: 97–105.

### Splicing and Disease

- David C.J. and Manley J.L. 2010. Alternative pre-mRNA splicing regulation in cancer: Pathways and programs unhinged. *Genes Dev.* **24**: 2343–2364.
- Kole R., Krainer A.R., and Altman S. 2012. RNA therapeutics: Beyond RNA interference and antisense oligonucleotides. *Nat. Rev. Drug Discov.* **11**: 125–140.
- Padgett R.A. 2012. New connections between splicing and human disease. *Trends Genet.* **28**: 147–154.

## QUESTIONS

### MasteringBiology®

For instructor-assigned tutorials and problems, go to *MasteringBiology*.

For answers to even-numbered questions, see Appendix 2: Answers.

**Question 1.** Suggest why the average number of introns per gene for the yeast *Saccharomyces cerevisiae* is much lower than the average number of introns per gene for *Homo sapiens*.

**Question 2.** Are the 5' splice site and 3' splice site labeled 5' and 3' with respect to the ends of the intron or the ends of the exons? In addition to the 5' and 3' splice sites, what other sequence is required in splicing? Where is this sequence located?

**Question 3.** Starting with an unspliced, pre-RNA, describe the intermediates produced after the first step in the splicing reaction.

**Question 4.** Given that the two key reactions for splicing could proceed in the forward or reverse direction, what prevents the splicing reactions from proceeding backward in vivo?

**Question 5.** Explain the basis of interaction between small nuclear ribonuclear proteins (snRNPs) and the pre-mRNA splicing substrate.

**Question 6.** If the 5' splice site sequence changed from 5'-GUAAGU-3' to 5'-GUAUGU-3', predict the effect of the sequence change on U1 binding and U6 snRNP binding in an in vitro protein–RNA binding assay.

**Question 7.** Compare and contrast the mechanisms of self-splicing by group I and group II introns.

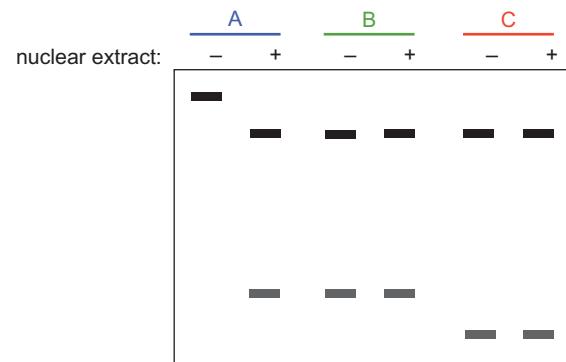
**Question 8.** Describe the product of a splicing reaction in which the spliceosome recognizes a “pseudo” 3' splice site within the intron instead of the actual 3' splice site. The “pseudo” site is slightly upstream of the actual 3' splice site. How could this alter the protein product after translation?

**Question 9.** Explain how steric hindrance can lead to mutually exclusive splicing.

**Question 10.** How does nonsense-mediated decay contribute to determining the final alternatively spliced products available for translation?

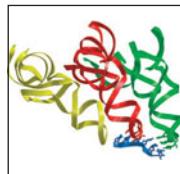
**Question 11.** Explain why the cell can use the mechanism of cytidine deamination of mRNA but not cytosine deamination of DNA for cell-specific expression of a specific gene.

**Question 12.** In a biochemical experiment, you compare the products from splicing reactions carried out in vitro using three different substrates. In each case the substrate is a construct containing a single intron surrounded by two exons, and in all cases the construct is the same overall size. But in one case, the intron is a group I intron, in another a group II intron, and in the third an intron removed by the spliceosome. Each construct is labeled in a manner that allows it to be detected after gel electrophoresis, and each is tested in two reactions—one, conditions that support self-splicing, and two, in the presence of nuclear extract as well. Match the intron type with the appropriate results (A, B, or C) in the gel shown below. Note that, for simplification, only the final products of the splicing reaction are seen, but before degradation of the introns.



*This page intentionally left blank*

CHAPTER 15



## Translation

THE CENTRAL QUESTION ADDRESSED IN THIS CHAPTER, as well as the next chapter, is how genetic information contained within the order of nucleotides in messenger RNA (mRNA) is interpreted to generate the linear sequences of amino acids in proteins. This process is known as **translation**. Of the events we have discussed, translation is among the most highly conserved across all organisms and among the most energetically costly for the cell. In rapidly growing bacterial cells, up to 80% of the cell's energy and 50% of the cell's dry weight are dedicated to protein synthesis. Indeed, the synthesis of a single protein requires the coordinated action of well over 100 proteins and RNAs. Consistent with the more complex nature of the translation process, we have divided our discussion into two chapters. In this first chapter, we describe the events that allow decoding of the mRNA, and in Chapter 16, we describe the nature of the genetic code and its recognition by transfer RNAs (tRNAs).

Translation is a much more formidable challenge in information transfer than the transcription of DNA into RNA. Unlike the complementarity between the DNA template and the ribonucleotides of the mRNA, the side chains of amino acids have little or no specific affinity for the purine and pyrimidine bases found in RNA. For example, the hydrophobic side chains of the amino acids alanine, valine, leucine, and isoleucine cannot form hydrogen bonds with the amino and keto groups of the nucleotide bases. Likewise, it is hard to imagine that several different combinations of three bases of RNA could form surfaces with unique affinities for the aromatic amino acids phenylalanine, tyrosine, and tryptophan. Thus, it seemed unlikely that direct interactions between the mRNA template and the amino acids could be responsible for the specific and accurate ordering of amino acids in a polypeptide.

With these considerations in mind, in 1955 Francis H. Crick proposed that before their incorporation into polypeptides, amino acids must attach to a special adaptor molecule that is capable of directly interacting with and recognizing the three-nucleotide-long coding units of the mRNA. Crick imagined that the adaptor would be an RNA molecule because it would need to recognize the code by Watson–Crick base-pairing rules. Just two years later, Paul C. Zamecnik and Mahlon B. Hoagland showed that before their incorporation into proteins, amino acids are attached to a class of RNA molecules (representing 15% of all cellular RNA). These RNAs are called transfer RNAs (tRNAs) because their attached amino acid is subsequently *transferred* to the growing polypeptide chain.

The machinery responsible for translating the language of mRNAs into the language of proteins is composed of four primary components: **mRNAs**,

### O U T L I N E

- Messenger RNA, 510
  - Transfer RNA, 513
  - Attachment of Amino Acids to tRNA, 515
  - The Ribosome, 519
  - Initiation of Translation, 528
  - Translation Elongation, 535
  - Termination of Translation, 544
  - Regulation of Translation, 549
  - Translation-Dependent Regulation of mRNA and Protein Stability, 563
- Visit Web Content for Structural Tutorials and Interactive Animations

**tRNAs, aminoacyl-tRNA synthetases, and the ribosome.** Together, these components accomplish the extraordinary task of translating a code written in a four-base alphabet into a second code written in the language of the 20 amino acids. The mRNA provides the information that must be interpreted by the translation machinery and is the template for translation. The protein-coding region of the mRNA consists of an ordered series of three-nucleotide-long units called **codons** that specify the order of amino acids. The tRNAs provide the physical interface between the amino acids being added to the growing polypeptide chain and the codons in the mRNA. Enzymes called aminoacyl-tRNA synthetases couple amino acids to specific tRNAs that recognize the appropriate codon(s). The final major player in translation is the ribosome, a remarkable, multimegadalton machine composed of both RNA and protein. The ribosome coordinates the correct recognition of the mRNA by each tRNA and catalyzes peptide-bond formation between the growing polypeptide chain and the amino acid attached to the selected tRNA.

We shall first consider the key attributes of each of these four components. We then describe how these components work together to accomplish translation. Recent progress in elucidating the structure of the components of the translational machinery make this an exciting area—one that is rich in mechanistic insights. Among the questions we will ask are: What is the organization of nucleotide sequence information in mRNA? What is the structure of tRNAs, and how do aminoacyl-tRNA synthetases recognize and attach the correct amino acids to each tRNA? Finally, how does the ribosome orchestrate the decoding of nucleotide sequence information and the addition of amino acids to the growing polypeptide chain?

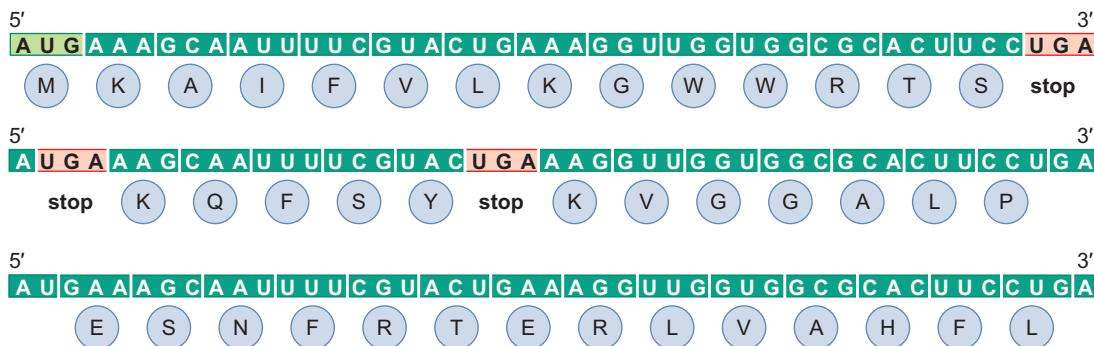
## MESSENGER RNA

---

### Polypeptide Chains Are Specified by Open Reading Frames

The translation machinery decodes only a portion of each mRNA. As we saw in Chapter 2 and will consider in detail in Chapter 16, the information for protein synthesis is in the form of three-nucleotide codons, which each specifies one amino acid. The protein-coding region(s) of each mRNA is composed of a contiguous, non-overlapping string of codons called an **open reading frame** (commonly known as an **ORF**). Each ORF specifies a single protein and starts and ends at internal sites within the mRNA. That is, the ends of an ORF are distinct from the ends of the mRNA.

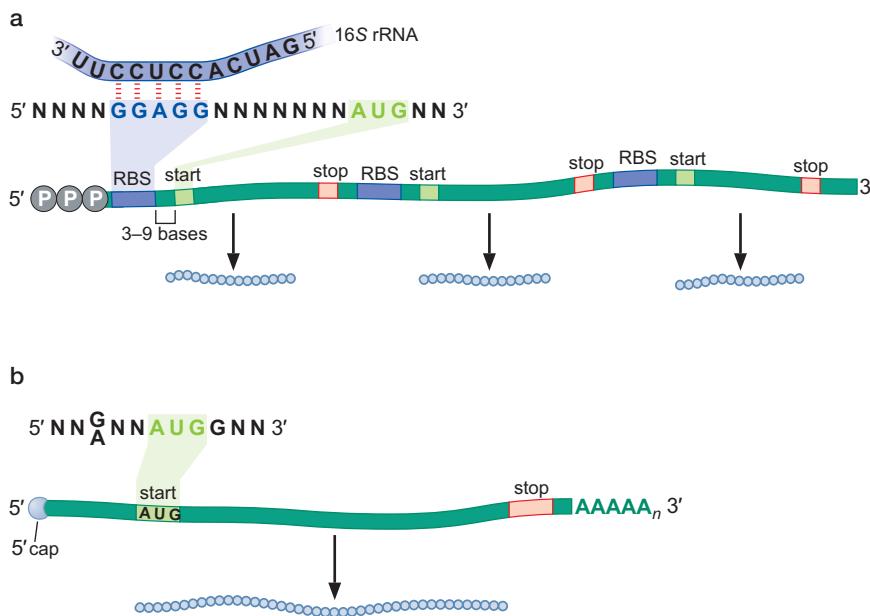
Translation starts at the 5' end of the ORF and proceeds one codon at a time to the 3' end. The first and last codons of an ORF are known as the **start** and **stop codons**. In bacteria, the start codon is usually 5'-AUG-3', but 5'-GUG-3' and sometimes even 5'-UUG-3' are also used. Eukaryotic cells always use 5'-AUG-3' as the start codon. The start codon has two important functions. First, it specifies the first amino acid to be incorporated into the growing polypeptide chain. Second, it defines the reading frame for all subsequent codons. Because each codon is immediately adjacent to (but not overlapping with) the next codon, and because codons are three nucleotides long, any stretch of mRNA could be translated in three different reading frames (Fig. 15-1). Once translation starts, however, the reading frame is determined. Thus, by setting the location of the first codon, the start codon determines the location of all following codons.



**FIGURE 15-1** Three possible reading frames of the *Escherichia coli* *trp* leader sequence. Start codons are shaded in green, and stop codons are shaded in red. The amino acid sequence encoded by each reading frame is indicated in the single-letter code below each codon.

Stop codons, of which there are three (5'-UAG-3', 5'-UGA-3', and 5'-UAA-3'), define the end of the ORF and signal termination of polypeptide synthesis. We can now fully appreciate the origin of the term *open reading frame*. It is a contiguous stretch of codons “read” in a particular frame (as set by the first codon) that is “open” to translation because it lacks a stop codon (i.e., until the last codon in the ORF).

mRNAs contain at least one ORF. The number of ORFs per mRNA is different between eukaryotes and prokaryotes. Eukaryotic mRNAs almost always contain a single ORF. In contrast, prokaryotic mRNAs frequently contain two or more ORFs and hence can encode multiple polypeptide chains. mRNAs containing multiple ORFs are known as **polycistronic mRNAs**, and those encoding a single ORF are known as **monocistronic mRNAs**. As we learned in Chapter 13, the polycistronic mRNAs found in bacteria often encode proteins that perform related functions, such as different steps in the biosynthesis of an amino acid or nucleotide. The structures of typical prokaryotic and eukaryotic mRNAs are shown in Figure 15-2.



**FIGURE 15-2** Structure of messenger RNA. (a) A polycistronic prokaryotic message with three ORFs. Each ribosome-binding site is indicated by a purple box labeled RBS. (b) A monocistronic eukaryotic message. The 5' cap is indicated by a “ball” at the end of the mRNA.

### Prokaryotic mRNAs Have a Ribosome-Binding Site That Recruits the Translational Machinery

For translation to occur, the ribosome must be recruited to the mRNA. To facilitate binding by a ribosome, many prokaryotic ORFs contain a short sequence upstream (on the 5' side) of the start codon called the **ribosome-binding site (RBS)**. This element is also referred to as a **Shine–Dalgarno sequence** after the scientists who discovered it by comparing the sequences of multiple mRNAs. The RBS, typically located 3–9 bp on the 5' side of the start codon, is complementary to a sequence located near the 3' end of one of the ribosomal RNA components, the 16S ribosomal RNA (rRNA) (see Fig. 15-2a). The RBS base-pairs with this RNA, thereby aligning the ribosome with the beginning of the ORF. The core of this region of the 16S rRNA has the sequence 5'-CCUCCU-3'. Not surprisingly, prokaryotic RBS are most often a subset of the sequence 5'-AGGAGG-3'. The extent of complementarity and the spacing between the RBS and the start codon has a strong influence on how actively a particular ORF is translated: high complementarity and proper spacing promote active translation, whereas limited complementarity and/or poor spacing generally support lower levels of translation.

Some prokaryotic ORFs lack a strong RBS but are nonetheless actively translated. These ORFs are not the first ORF in an mRNA but instead are located just after another ORF in a polycistronic message (not all prokaryotic mRNAs are polycistronic). In these cases, the start codon of the downstream ORF often overlaps the 3' end of the upstream ORF (most often as the sequence 5'-AUGA-3', which contains a start and a stop codon). Thus, a ribosome that has just completed translating the upstream ORF is positioned to begin translating from the start codon for the downstream ORF. This arrangement circumvents the need for an RBS to recruit the ribosome. This phenomenon of linked translation between overlapping ORFs is known as **translational coupling**. It is important to note that in this situation translation of the downstream ORF requires translation of the upstream ORF. Indeed, with two translationally coupled genes, a mutation that leads to a premature stop codon in the upstream ORF also prevents translation of the downstream ORF.

### Eukaryotic mRNAs Are Modified at Their 5' and 3' Ends to Facilitate Translation

Unlike their prokaryotic counterparts, eukaryotic mRNAs recruit ribosomes using a specific chemical modification called the **5' cap**, which is located at the extreme 5' end of the mRNA (see Chapter 13, Fig. 13-25). The 5' cap is a methylated guanine nucleotide that is joined to the 5' end of the mRNA via an unusual 5'-to-5' linkage. Created in three steps (see Chapter 13, Fig. 13-24), the guanine nucleotide of the 5' cap is connected to the 5' end of the mRNA through three phosphate groups. The resulting 5' cap is required to recruit the ribosome to the mRNA. Once bound to the mRNA, the ribosome moves in a 5' → 3' direction until it encounters a 5'-AUG-3' start codon, a process called **scanning**.

Two other features of eukaryotic mRNAs stimulate translation. One feature is the presence, in some mRNAs, of a purine three bases upstream of the start codon and a guanine immediately downstream (5'-G/ANNAUGG-3'). This sequence was originally identified by Marilyn Kozak and is referred to as the Kozak sequence. Many eukaryotic mRNAs lack these bases, but their presence increases the efficiency of translation. In contrast to the situation in prokaryotes, these bases are thought to interact with the initiator tRNA, not with an RNA component of the ribosome. A second feature

that contributes to efficient translation is the presence of a poly-A tail at the extreme 3' end of the mRNA. As we saw in Chapter 13, this tail is added enzymatically by the enzyme poly-A polymerase. Despite its location at the 3' end of the mRNA, the poly-A tail enhances the level of translation of the mRNA by enhancing the recruitment of key translation initiation factors. Importantly, in addition to their roles in translation, these 5'- and 3'-end modifications also protect eukaryotic mRNAs from rapid degradation (as we shall discuss in the section Translation-Dependent Regulation of mRNA and Protein Stability).

## TRANSFER RNA

### tRNAs Are Adaptors between Codons and Amino Acids

The heart of protein synthesis is the “translation” of nucleotide sequence information (in the form of codons) into amino acids. This is accomplished by tRNA molecules, which act as adaptors between codons and the amino acids they specify. There are many types of tRNA molecules, but each is attached to a specific amino acid, and each recognizes a particular codon, or codons, in the mRNA (most tRNAs recognize more than one codon, as we shall discuss in Chapter 16). tRNA molecules are between 75 and 95 ribonucleotides in length. Although the exact sequence varies, all tRNAs have certain features in common. First, all tRNAs end at the 3' terminus with the sequence 5'-CCA-3' (see Box 15-1, CCA-Adding Enzymes: Synthesizing RNA without a Template). Consistent with this absolute conservation, the 3' end of this sequence (and of the tRNA) is the site that is attached to the cognate amino acid.

A second striking aspect of tRNAs is the presence of several unusual bases in their primary structure. These unusual features are created post-

#### ► ADVANCED CONCEPTS

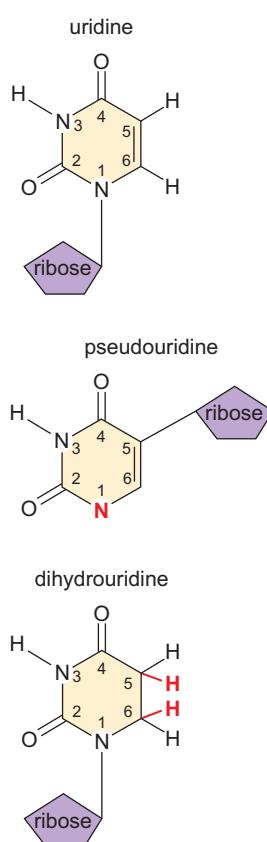
##### Box 15-1 CCA-Adding Enzymes: Synthesizing RNA without a Template

As we have described, the 5'-CCA-3' end is universally conserved for all tRNAs and is absolutely required for protein synthesis. Oddly, when the genes that encode tRNAs were cloned, it was found that many do not encode the CCA end. Instead, these genes end three nucleotides short of the 3' ends found in the mature tRNA. In fact, the genes encoding many bacterial tRNAs and almost all eukaryotic tRNAs lack this final three-base sequence. How then do these tRNAs acquire their CCA ends? The answer is provided by a specialized RNA polymerase called a **CCA-adding enzyme**. As its name indicates, this enzyme adds the terminal CCA to tRNAs that initially lack this sequence. Surprisingly, there is no nucleic acid component to this enzyme; that is, CCA-adding enzymes add a specific sequence to the end of the tRNA without an RNA or DNA template.

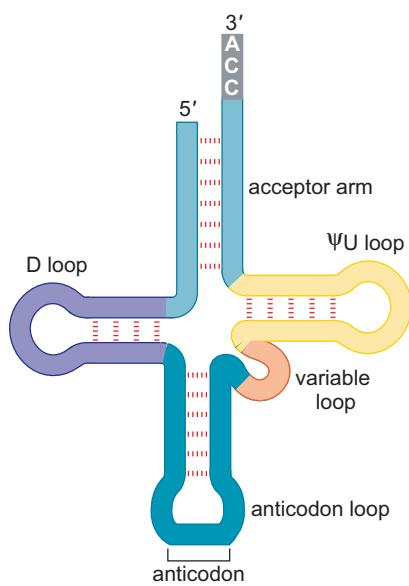
How do CCA-adding enzymes add a specific sequence without a template? A series of three-dimensional (3D) structures of these enzymes has begun to reveal the solution. First, like other RNA and DNA polymerases, CCA-adding enzymes have only one active site that uses a similar two-metal ion mechanism of catalysis (see Chapter 9, Fig. 9-6). Within this active site, an amino acid and a phosphate from the terminal

tRNA nucleotide form hydrogen bonds with A and C bases but not with G or U. This specificity can be understood by observing the pattern of hydrogen-bond donors and acceptors on each of the bases (see Chapter 6, Fig. 6-14). The patterns of A and C are overlapping, but G and U (or T) have opposite and complementary patterns. Indeed, it is this complementarity that is responsible for the specificity of base pairing within the double-stranded DNA helix. This hydrogen-bonding pattern explains the specificity of the enzyme for C and A.

Specificity for the addition of C versus A is controlled by changes in the active site as each base is added. Unlike other polymerases, the tRNA template does not change its position as each additional nucleotide is added. Instead, the template tRNA is held firmly in place, and each added nucleotide alters the structure of the active site. The result of these changes is that the active site is specific for C when a CCA-less tRNA binds but is altered to be specific for A after two C residues have been added. Once the A residue is added, the active site is no longer accessible to additional bases, and the tRNA with its newly added CCA end is released.



**FIGURE 15-3** A subset of modified nucleosides found in tRNA. Uridine and two uridine-related nucleotides are shown.



**FIGURE 15-4** Cloverleaf representation of the secondary structure of tRNA. In this representation of a tRNA, the base pairings between different parts of the tRNA are indicated by the dotted red lines.

transcriptionally by enzymatic modification of normal bases in the polynucleotide chain. For example, **pseudouridine** ( $\Psi$ U) is derived from uridine by an isomerization in which the site of attachment of the uracil base to the ribose is switched from the nitrogen at ring position 1 to the carbon at ring position 5 (Fig. 15-3). Likewise, **dihydrouridine** (D) is derived from uridine by enzymatic reduction of the double bond between the carbons at positions 5 and 6. Other unusual bases found in tRNA include hypoxanthine, thymine, and methylguanine. These modified bases are not essential for tRNA function, but cells lacking these modified bases show reduced rates of growth. This observation suggests that the modified bases lead to improved tRNA function. For example, as we shall see in Chapter 16, hypoxanthine plays an important role in the process of codon recognition by certain tRNAs.

### tRNAs Share a Common Secondary Structure That Resembles a Cloverleaf

As we saw in Chapter 5, RNA molecules typically contain regions of self-complementarity that enable them to form limited stretches of double helix that are held together by base pairing. Other regions of RNA molecules have no complement and hence are single-stranded. tRNA molecules show a characteristic and highly conserved pattern of single-stranded and double-stranded regions (secondary structure) that can be illustrated as a cloverleaf (Fig. 15-4). The principal features of the tRNA cloverleaf are an acceptor stem, three stem-loops (referred to as the  $\Psi$ U loop, the D loop, and the anticodon loop), and a fourth variable loop. Descriptions of each of these features follows.

- The **acceptor stem**, so-named because it is the site of attachment of the amino acid, is formed by pairing between the 5' and 3' ends of the tRNA molecule. The 5'-CCA-3' sequence at the extreme 3' end of the molecule is a single-strand region that protrudes from this double-strand stem.
- The  **$\Psi$ U loop** is so-named because of the characteristic presence of the unusual base  $\Psi$ U in the loop. The modified base is often found within the sequence 5'-T $\Psi$ UCG-3'.
- The **D loop** takes its name from the characteristic presence of dihydrouridines in the loop.
- The **anticodon loop**, as its name implies, contains the anticodon, a three-nucleotide-long sequence that is responsible for recognizing the codon by base pairing with the mRNA. The anticodon is always bracketed on the 3' end by a purine and on its 5' end by uracil.
- The **variable loop** sits between the anticodon loop and the  $\Psi$ U loop and, as its name implies, varies in size from 3 to 21 bases.

### tRNAs Have an L-Shaped Three-Dimensional Structure

The cloverleaf reveals regions of self-complementarity within tRNAs. What is the actual three-dimensional (3D) configuration of this adaptor molecule? X-ray crystallography reveals an L-shaped tertiary structure in which the terminus of the acceptor stem is at one end of the molecule and the anticodon loop is  $\sim 70$  Å away at the other end (Fig. 15-5c). To understand the relationship of this L-shaped structure to the cloverleaf, consider the following: the acceptor stem and the stem of the  $\Psi$ U loop form an extended helix in the final tRNA structure (Fig. 15-5b). Similarly, the anticodon stem and the stem of the

D loop form a second extended helix. These two extended helices align at a right angle to each other, with the D loop and the  $\Psi$ U loop coming together.

Three kinds of interactions stabilize this L-shaped structure. First, the formation of the two extended regions of base pairing results in base-stacking interactions similar to those seen in double-stranded DNA. Second, hydrogen bonds are formed between bases in different helical regions that are brought near each other in 3D space by the tertiary structure. These base–base interactions are generally unconventional (non-Watson–Crick) bonding. Finally, there are interactions between the bases and the sugar–phosphate backbone.

## ATTACHMENT OF AMINO ACIDS TO tRNA

### tRNAs Are Charged by the Attachment of an Amino Acid to the 3'-Terminal Adenosine Nucleotide via a High-Energy Acyl Linkage

tRNA molecules to which an amino acid is attached are said to be charged, and tRNAs that lack an amino acid are said to be **uncharged**. Charging requires an acyl linkage between the carboxyl group of the amino acid and the 2'- or 3'-hydroxyl group (see later discussion) of the adenosine nucleotide that protrudes from the acceptor stem at the 3' end of the tRNA. This acyl linkage is a high-energy bond because its hydrolysis results in a large change in free energy. This is significant for protein synthesis: the energy released when this acyl bond is broken is coupled to the formation of the peptide bonds that link amino acids to each other in polypeptide chains.

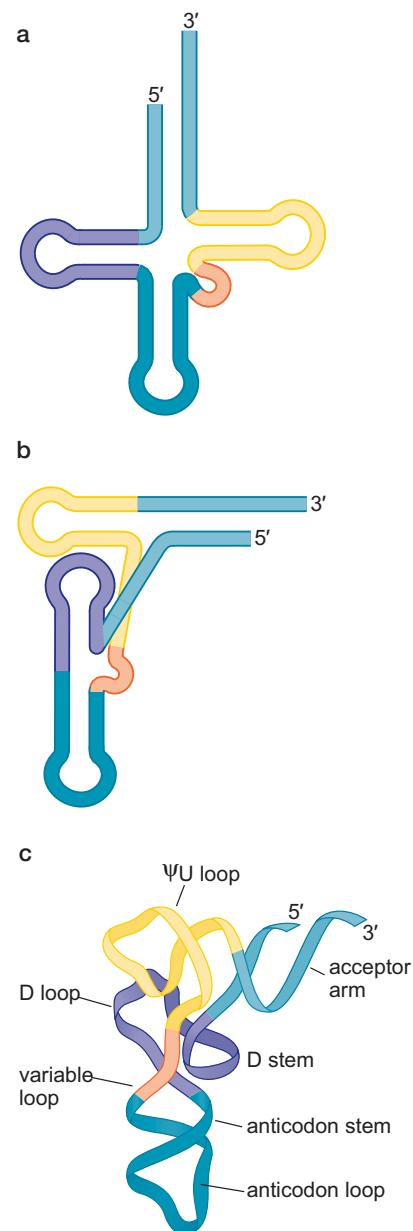
### Aminoacyl-tRNA Synthetases Charge tRNAs in Two Steps

All aminoacyl-tRNA synthetases attach an amino acid to a tRNA in two enzymatic steps (Fig. 15-6). Step one is **adenylylation** in which the amino acid reacts with ATP to become adenylylated with the concomitant release of pyrophosphate. Adenylylation refers to transfer of AMP, as opposed to adenylation, which would indicate the transfer of adenine. As we have seen in the case of polynucleotide synthesis (see Chapter 9), the principal driving force for the adenylylation reaction is the subsequent hydrolysis of pyrophosphate by pyrophosphatase. As a result of adenylylation, the amino acid is attached to adenylic acid via a high-energy ester bond in which the carbonyl group of the amino acid is joined to the phosphoryl group of AMP. Step two is **tRNA charging** in which the adenylylated amino acid, which remains tightly bound to the synthetase, reacts with tRNA. This reaction results in the transfer of the amino acid to the 3' end of the tRNA via the 2'- or 3'-hydroxyl and the release of AMP.

There are two classes of tRNA synthetases (Table 15-1). Class I enzymes attach the amino acid to the 2'-OH of the tRNA and are generally monomeric. Class II enzymes attach the amino acid to the 3'-OH of the tRNA and are typically dimeric or tetrameric. Although the initial coupling between the tRNA and the amino acid is different, once released from the synthetase, the amino acid rapidly equilibrates between attachment at the 3'-OH and the 2'-OH.

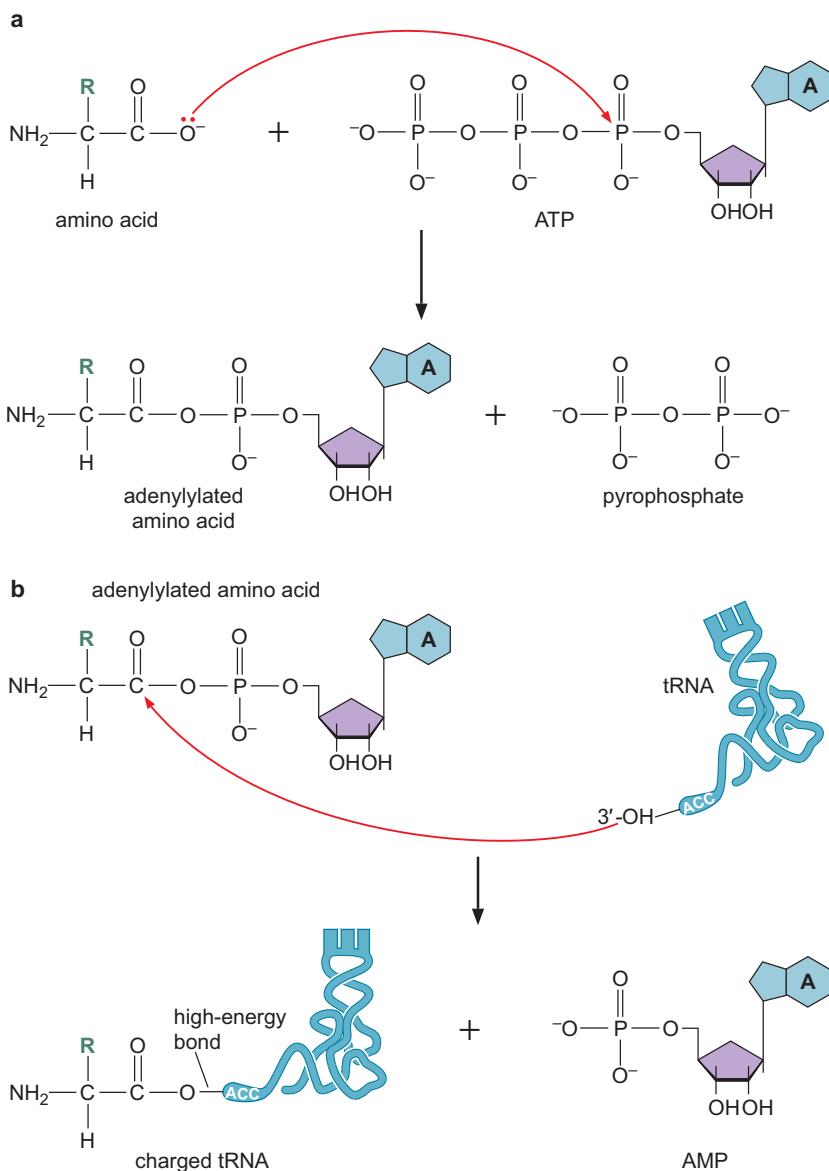
### Each Aminoacyl-tRNA Synthetase Attaches a Single Amino Acid to One or More tRNAs

Each of the 20 amino acids is attached to the appropriate tRNA by a single, dedicated tRNA synthetase. Because most amino acids are specified by more than one codon (see Chapter 16), it is not uncommon for one syn-



**FIGURE 15-5** Conversion between the cloverleaf and the actual 3D structure of a tRNA. (a) Cloverleaf representation. (b) L-shaped representation showing the location of the base-paired regions of the final folded tRNA. (c) Ribbon representation of the actual folded structure of a tRNA. Note that although this diagram illustrates how the actual tRNA structure is related to the cloverleaf representation, a tRNA does not attain its final structure by first base pairing and then folding into an L shape.

**FIGURE 15-6** The two steps of aminoacyl-tRNA charging. (a) Adenylylation of amino acid. (b) Transfer of the adenylylated amino acid to tRNA. The process shown is for a class II tRNA synthetase (which attaches the amino acid to the 3'-OH).



**TABLE 15-1** Classes of Aminoacyl-tRNA Synthetases

Class II	Quaternary Structure	Class I	Quaternary Structure
Gly	( $\alpha_2\beta_2$ )	Glu	( $\alpha$ )
Ala	( $\alpha_4$ )	Gln	( $\alpha$ )
Pro	( $\alpha_2$ )	Arg	( $\alpha$ )
Ser	( $\alpha_2$ )	Cys	( $\alpha_2$ )
Thr	( $\alpha_2$ )	Met	( $\alpha_2$ )
His	( $\alpha_2$ )	Val	( $\alpha$ )
Asp	( $\alpha_2$ )	Ile	( $\alpha$ )
Asn	( $\alpha_2$ )	Leu	( $\alpha$ )
Lys	( $\alpha_2$ )	Tyr	( $\alpha$ )
Phe	( $\alpha_2\beta_2$ )	Trp	( $\alpha$ )

Adapted, with permission, from Delarue M. 1995. *Curr. Opin. Struct. Biol.* 5: 48–55, Table 1. © Elsevier.

Class I enzymes are generally monomeric, whereas class II enzymes are dimeric or tetrameric, with residues from two subunits contributing to the binding site for a single tRNA.  $\alpha$  and  $\beta$  refer to subunits of the tRNA synthetases, and the subscripts indicate their stoichiometry.

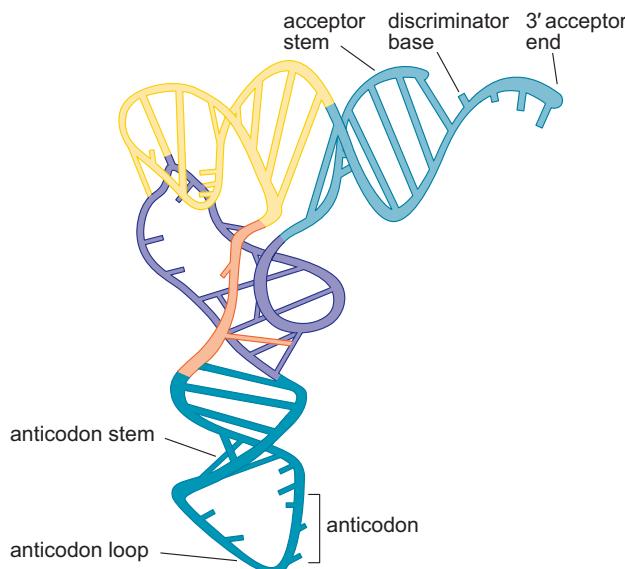
thetase to recognize and charge more than one tRNA (known as iso-accepting tRNAs). Nevertheless, the same tRNA synthetase is responsible for charging all tRNAs for a particular amino acid. Thus, one and only one tRNA synthetase attaches each amino acid to all of the appropriate tRNAs.

Most organisms have 20 different tRNA synthetases, but this is not always the case. For example, some bacteria lack a synthetase for charging the tRNA for glutamine ( $tRNA^{Gln}$ ) with its cognate amino acid. Instead, a single species of aminoacyl-tRNA synthetase charges  $tRNA^{Gln}$  as well as  $tRNA^{Glu}$  with glutamate. A second enzyme then converts (by amination) the glutamate moiety of the charged  $tRNA^{Gln}$  molecules to glutamine. That is,  $Glu-tRNA^{Gln}$  is aminated to  $Gln-tRNA^{Gln}$  (the prefix identifies the attached amino acid, and the superscript identifies the type of codon the tRNA recognizes). The presence of this second enzyme removes the need for a glutamine tRNA synthetase. Nevertheless, an aminoacyl-tRNA synthetase can never attach more than one kind of amino acid to a given tRNA.

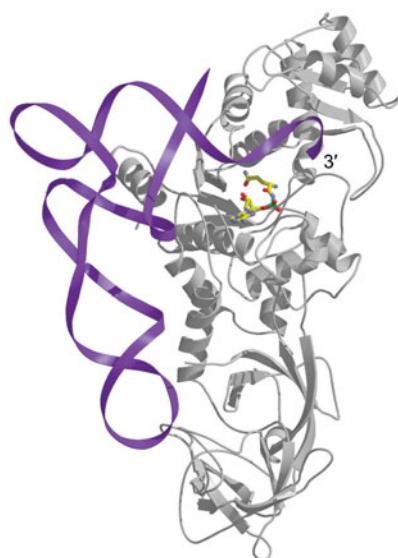
### tRNA Synthetases Recognize Unique Structural Features of Cognate tRNAs

As we can see from the above considerations, aminoacyl-tRNA synthetases face two important challenges: they must recognize the correct set of tRNAs for a particular amino acid, and they must charge all of these iso-accepting tRNAs with the correct amino acid. Both processes must be performed with high fidelity.

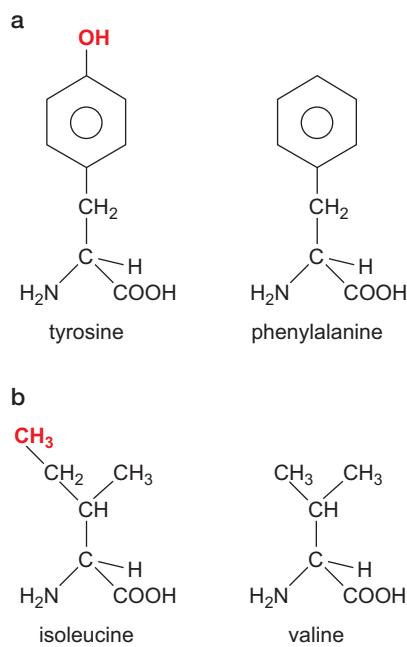
Let us first consider the specificity of tRNA recognition: what features of the tRNA molecule enable a synthetase to discriminate the correct set of iso-accepting tRNAs from the tRNAs for the other 19 amino acids? Genetic, biochemical, and X-ray crystallographic evidence indicates that the specificity determinants are clustered at two distant sites on the molecule: the acceptor stem and the anticodon loop (Fig. 15-7). The acceptor stem is an especially important determinant for the specificity of tRNA synthetase recognition. In some cases, changing a single base in the acceptor stem (known as the



**FIGURE 15-7** Structure of tRNA: elements required for aminoacyl synthetase recognition.



**FIGURE 15-8** Cocrystal structure of glutaminyl aminoacyl-tRNA synthetase with tRNA<sup>Gln</sup>. Enzyme (gray); tRNA<sup>Gln</sup> (purple). The yellow, red, and green molecule is glutaminyl-AMP. Note the proximity of this molecule to the 3' end of the tRNA and the points of contact between the tRNA and the synthetase. (Rath V.L. et al. 1998. *Structure* 6: 439–449.) Image prepared with MolScript, BobScript, and Raster3D.



**FIGURE 15-9** Distinguishing features of similar amino acids.

**discriminator base**) is sufficient to convert the recognition specificity of a tRNA from one synthetase to another. Nonetheless, the anticodon loop frequently contributes to discrimination as well. The synthetase for glutamine, for example, makes numerous contacts both in the acceptor stem and across the anticodon loop, including the anticodon itself (Fig. 15-8).

One might expect that the anticodon would always be used for recognition by tRNA synthetases because it is the ultimate defining feature of a tRNA—the anticodon dictates the amino acid that the tRNA is responsible for incorporating into the growing polypeptide chain. However, because each amino acid is usually specified by more than one codon, recognition of the anticodon cannot be used in many cases. For example, the amino acid serine is specified by six codons, including 5'-AGC-3' and 5'-UCA-3', which are completely different from one another. Hence, the tRNAs for serine necessarily have a variety of different anticodons, which could not be easily recognized by a single tRNA synthetase. Therefore, to recognize its tRNAs, the synthetase for serine must rely on determinants that lie outside of the anticodon.

### Aminoacyl-tRNA Formation Is Very Accurate

The challenge faced by aminoacyl-tRNA synthetases in selecting the correct amino acid is perhaps even more daunting than the challenge the enzyme faces in recognizing the appropriate tRNA (Fig. 15-9). The reason for this is the relatively small size of amino acids and, in some cases, their similarity. Despite this challenge, the frequency of mischarging is very low; typically, less than 1 in 1000 tRNAs is charged with the incorrect amino acid. In certain cases, it is easy to understand how this high accuracy is achieved. For example, the amino acids cysteine and tryptophan differ substantially in size, shape, and chemical groups. Even in the case of the similar-looking amino acids tyrosine and phenylalanine (see Fig. 15-9a), the opportunity for forming a strong and energetically favorable hydrogen bond with the hydroxyl moiety of the former but not the latter allows the synthetase for tyrosine (tyrosyl-tRNA synthetase) to discriminate effectively against phenylalanine.

It is more challenging to understand the case of isoleucine and valine, which differ by only a single methylene group (see Fig. 15-9b). Valyl-tRNA synthetase can sterically exclude isoleucine from its catalytic pocket because isoleucine is larger than valine. In contrast, valine should slip easily into the catalytic pocket of the isoleucyl-tRNA synthetase. Although both amino acids will fit into the isoleucyl-tRNA synthetase amino acid–binding site, interactions with the extra methylene group on isoleucine will provide an extra  $-2$  to  $-3$  kcal/mol of free energy (see Chapter 3, Table 3-1). As we described in Chapter 3, even this relatively small difference in free energy will make binding to isoleucine  $\sim$ 100-fold more likely than binding to valine if the two amino acids are present at equal concentrations. Thus, valine would be attached to isoleucine tRNAs  $\sim$ 1% of the time; however, this is an unacceptably high rate of error. As we have discussed, the actual frequency of misincorporation is  $<0.1\%$ . How is this additional level of fidelity achieved?

### Some Aminoacyl-tRNA Synthetases Use an Editing Pocket to Charge tRNAs with High Accuracy

One common mechanism to increase the fidelity of an aminoacyl-tRNA synthetase is to proofread the products of the charging reaction as we have seen for DNA polymerases in Chapter 9. For example, in addition to its catalytic

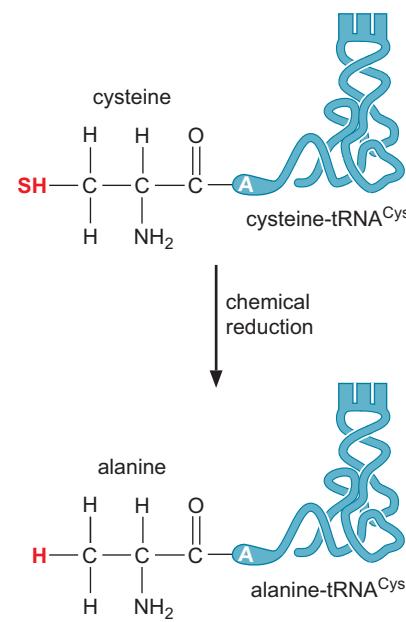
pocket (for adenyllylation), isoleucyl-tRNA synthetase has a nearby editing pocket (a deep cleft in the enzyme) that allows it to proofread the product of the adenyllylation reaction. AMP-valine (as well as adenyllylates of other small amino acids, such as alanine) can fit into this editing pocket, where it is hydrolyzed and released as free valine and AMP. In contrast, AMP-isoleucine is too large to enter the editing pocket and is therefore not subject to hydrolysis. As a consequence, isoleucyl-tRNA synthetase discriminates against valine twice: in the initial binding and adenyllylation of the amino acid (discriminating by a factor of  $\sim 100$ ), and then in the editing of the adenyllylated amino acid (again discriminating by a factor of  $\sim 100$ ), for an overall selectivity of  $\sim 10,000$ -fold (i.e., an error rate of  $\sim 0.01\%$ ).

### The Ribosome Is Unable to Discriminate between Correctly and Incorrectly Charged tRNAs

The reason that so much responsibility falls on aminoacyl-tRNA synthetases to couple the proper amino acid with its cognate tRNA is that the ribosome cannot distinguish between correctly and incorrectly charged tRNAs. In other words, the ribosome “blindly” accepts any charged tRNA that shows a proper codon–anticodon interaction, whether or not the tRNA is charged with the correct amino acid.

This conclusion is supported by two kinds of experiments: one genetic and the other biochemical. The genetic experiment involves the isolation of a mutant tRNA that carries a nucleotide substitution in the anticodon. Recall that tRNA synthetases frequently do not rely on interaction with the anticodon to recognize cognate tRNAs. Hence, a subset of tRNAs can be mutated in their anticodons but still be charged with their usual cognate amino acids. As a consequence of the anticodon mutation, however, the mutant tRNA delivers its amino acid to the wrong codon. In other words, the ribosome and the auxiliary proteins that work in conjunction with the ribosome (which we shall discuss shortly) primarily check that the charged tRNA makes a proper codon–anticodon interaction with the mRNA. The ribosome and these proteins do little to prevent an incorrectly charged tRNA from adding an inappropriate amino acid to the growing polypeptide.

A classic biochemical experiment nicely illustrates the point that the ribosome recognizes tRNA and not the amino acid that it is carrying. Consider the charged tRNA cysteine-tRNA<sup>Cys</sup> (remember that the prefix identifies the amino acid and the superscript identifies the nature of the tRNA). The cysteine attached to cysteine-tRNA<sup>Cys</sup> can be converted to an alanine by chemical reduction to give alanine-tRNA<sup>Cys</sup> (Fig. 15-10). When added to a cell-free protein-synthesizing system, alanine-tRNA<sup>Cys</sup> introduces alanines at codons that specify insertion of cysteine. Thus, the translation machinery relies on the high fidelity of the aminoacyl-tRNA synthetases to ensure the accurate decoding of each mRNA (see Box 15-2, Selenocysteine).



**FIGURE 15-10** Chemical reduction of cysteine-tRNA<sup>Cys</sup> to alanine-tRNA<sup>Cys</sup>.

## THE RIBOSOME

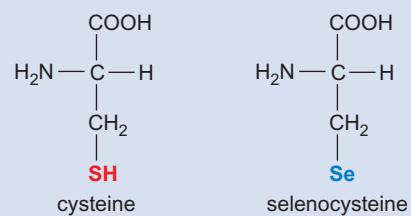
The **ribosome** is the macromolecular machine that directs the synthesis of proteins. Consistent with the additional challenges of translating a nucleic acid code into an amino acid code, the ribosome is larger and more complex than the minimal machinery required for DNA or RNA synthesis. Indeed, single polypeptides can perform DNA or RNA synthesis (although DNA replication and transcription are more frequently mediated by larger multi-

## ► ADVANCED CONCEPTS

**Box 15-2 Selenocysteine**

Certain proteins, such as the enzymes glutathione peroxidase and formate dehydrogenase, contain an unusual amino acid called selenocysteine, which is part of the catalytic center of the enzymes. Selenocysteine contains the trace element selenium in place of the sulfur atom of cysteine (Box 15-2 Fig. 1). Interestingly, selenocysteine is not incorporated into proteins by chemical modification after translation (as is true for certain other unusual amino acids, such as hydroxyproline, which is found in collagen). Instead, selenocysteine is generated enzymatically from serine carried on a special tRNA that is charged by serine-tRNA synthetase. This altered tRNA is used to incorporate selenocysteine directly into enzymes such as glutathione peroxidase as they are synthesized. A dedicated (EF-Tu-like; see below) translation elongation factor delivers selenocysteinyl-tRNA to the ribosome at a codon (UGA) that would normally be recognized as a stop codon. Incorporation of

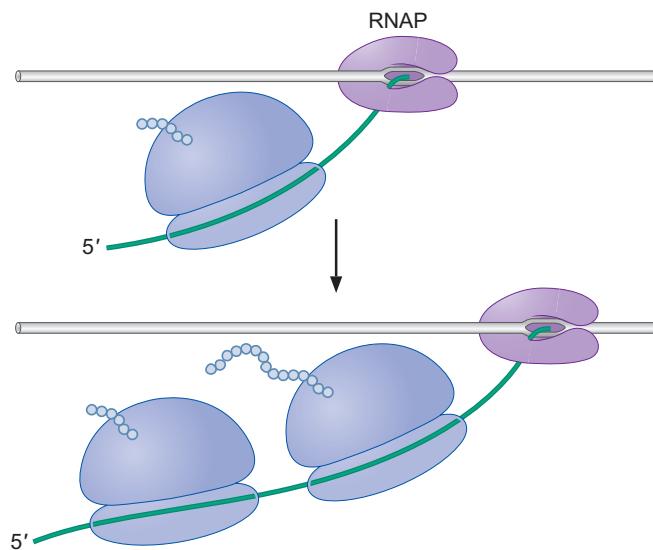
selenocysteine at UGA codons requires the presence of a special sequence element elsewhere in the mRNA. Thus, selenocysteine can be thought of as a twenty-first amino acid that is incorporated into proteins by a modification of the standard translation machinery of the cell.



**BOX 15-2 FIGURE 1** The structures of cysteine and selenocysteine.

subunit complexes). In contrast, the machinery for polymerizing amino acids is composed of at least three RNA molecules and more than 50 different proteins, with an overall molecular mass of  $>2.5$  MDa. Compared with the speed of DNA replication—200–1000 nucleotides per second—translation takes place at a rate of only two to 20 amino acids per second.

In prokaryotes, the transcription machinery and the translation machinery are located in the same compartment. Thus, the ribosome can commence translation of the mRNA as it emerges from the RNA polymerase. This situation allows the ribosome to proceed in tandem with the RNA polymerase as it elongates the transcript (Fig. 15-11). Recall that the 5' end of an RNA is synthesized first, and thus the ribosome, which begins translation at the 5' end of the mRNA, can start translating a nascent transcript as soon as it emerges from the RNA polymerase. Interestingly, there are several instances



**FIGURE 15-11** Prokaryotic RNA polymerase and ribosomes at work on the same mRNA.

in which the coupling of transcription and translation is exploited during the regulation of gene expression, as we shall see in Chapter 18.

Although slow relative to DNA synthesis in prokaryotes, the ribosome is capable of keeping up with the transcription machinery. The typical prokaryotic rate of translation of 20 amino acids per second corresponds to the translation of 60 nucleotides (20 codons) of mRNA per second. This is similar to the rate of 50–100 nucleotides per second synthesized by RNA polymerase.

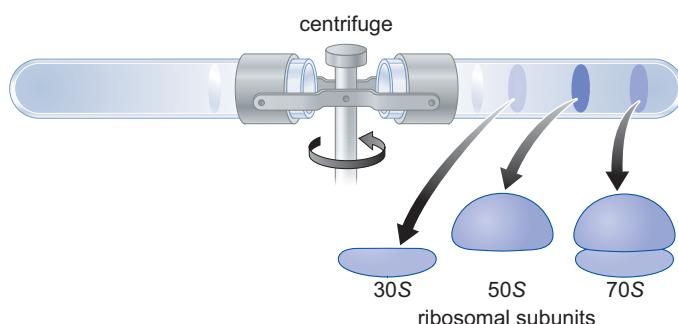
In contrast to the situation in prokaryotes, translation in eukaryotes is completely separate from transcription. These events occur in separate compartments of the cell: transcription occurs in the nucleus, whereas translation occurs in the cytoplasm. Perhaps because of the lack of coupling to transcription, eukaryotic translation proceeds at the more leisurely speed of two to four amino acids per second.

### The Ribosome Is Composed of a Large and a Small Subunit

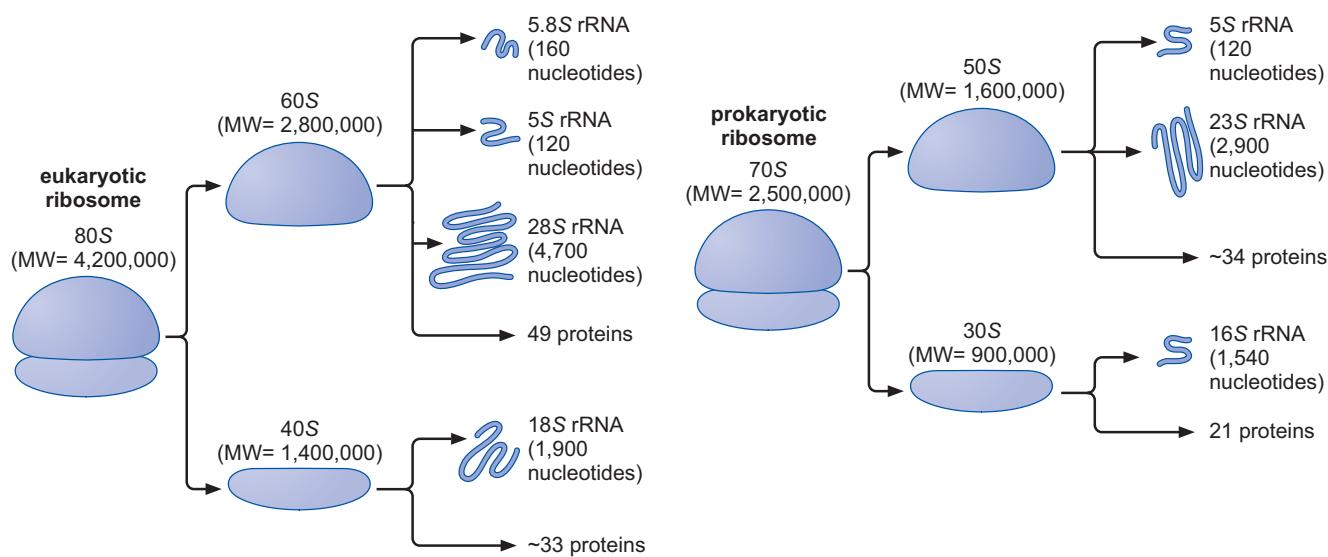
The ribosome is composed of two subassemblies of RNA and protein known as the large and small subunits. The large subunit contains the **peptidyl transferase center**, which is responsible for the formation of peptide bonds. The small subunit contains the **decoding center** in which charged tRNAs read or “decode” the codon units of the mRNA.

By convention, the large and small subunits are named according to the velocity of their sedimentation when subjected to a centrifugal force (Fig. 15-12). The unit used to measure sedimentation velocity is the **Svedberg** (*S*; the larger the *S* value the faster the sedimentation velocity and the larger the molecule), which is named after the Nobel Laureate and inventor of the ultracentrifuge, Theodor Svedberg. In bacteria, the large subunit has a sedimentation velocity of 50 Svedberg units and is accordingly known as the **50S** subunit, whereas the small subunit is called the **30S** subunit. The intact prokaryotic ribosome is referred to as the **70S** ribosome. Note that 70S is less than the sum of 50S and 30S! The explanation for this apparent discrepancy is that sedimentation velocity is determined by both shape and size and hence is not an exact measure of mass. The eukaryotic ribosome is somewhat larger, composed of **60S** and **40S** subunits, which together form an **80S** ribosome.

The large and small subunits are each composed of one or more RNAs (known as ribosomal RNAs or rRNAs), and many ribosomal proteins (Fig. 15-13). Svedberg units are once again used to distinguish among the rRNAs. Thus, in bacteria, the 50S subunit contains a **5S** rRNA and a **23S** rRNA, whereas the 30S subunit contains a single **16S** rRNA. Although there are far more ribosomal proteins than rRNAs in each subunit, more than two-thirds of the mass of the prokaryotic ribosome is RNA. This is true because



**FIGURE 15-12** Sedimentation by ultracentrifugation separates the bacterial ribosome subunits from the full ribosome.



**FIGURE 15-13** Composition of the prokaryotic and eukaryotic ribosomes. The rRNA and protein composition of the different subunits are indicated. The length of the rRNA and the number of ribosomal proteins are indicated for each subunit.

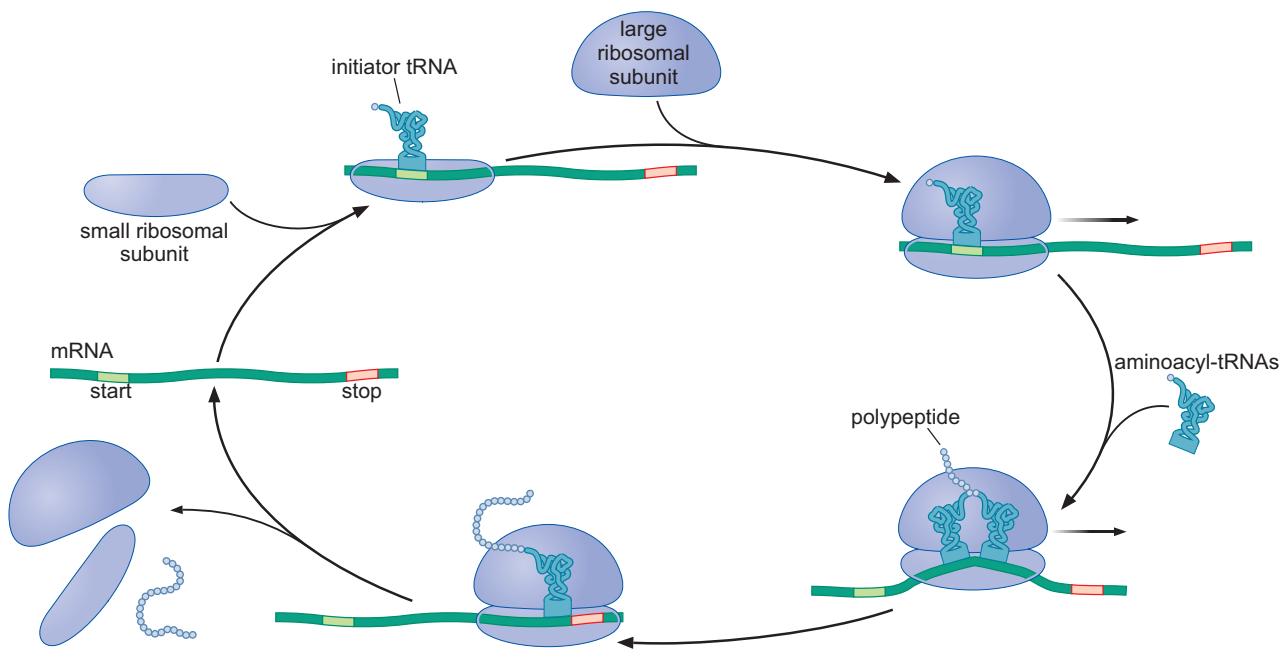
the ribosomal proteins are small (the average molecular mass of a ribosomal protein in the bacterial small subunit is  $\sim 15$  kDa). In contrast, the 16S and 23S rRNAs are large. Recall that, on average, a single nucleotide has a molecular mass of 330 Da; therefore, on its own, the 2900-nucleotide-long 23S rRNA has a molecular mass of almost 1000 kDa.

### The Large and Small Subunits Undergo Association and Dissociation during Each Cycle of Translation

Each time a protein is synthesized, the translation components undergo a specific series of events in which the small and large subunits of the ribosome associate with each other and the mRNA, translate the target mRNA, and then dissociate after completing synthesis of the protein. This sequence of association and dissociation is known as the **ribosome cycle** (Fig. 15-14; see also Interactive Animation 15-1). Briefly, translation begins with the binding of the mRNA and an initiating tRNA to a free, small subunit of the ribosome. The small subunit–mRNA–initiator–tRNA complex then recruits a large subunit to create an intact ribosome with the mRNA sandwiched between the two subunits. Protein synthesis is initiated in the next step, commencing at the start codon at the 5' end of the message and progressing toward the 3' end of the mRNA. As the ribosome translocates from codon to codon, one charged tRNA after another is slotted into the decoding and peptidyl transferase centers of the ribosome. When the elongating ribosome encounters a stop codon, the now completed polypeptide chain is released, and the ribosome dissociates from the mRNA as separate large and small subunits. The separated subunits are now available to bind to a new mRNA molecule and repeat the cycle of protein synthesis.

Although a ribosome can synthesize only one polypeptide at a time, each mRNA can be translated simultaneously by multiple ribosomes (for simplicity, let us assume that the message we are considering is monocistronic). An mRNA bearing multiple ribosomes is known as a **polyribosome** or a **polysome** (Fig. 15-15). A single ribosome contacts  $\sim 30$  nucleotides of mRNA, but the large size of the ribosome only allows a density of one ribosome for every





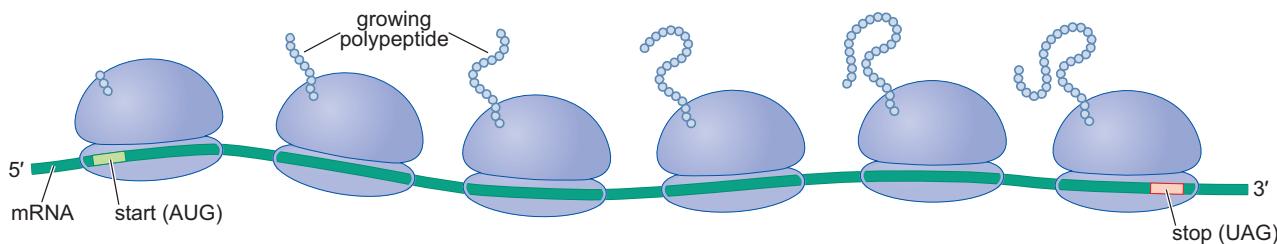
**FIGURE 15-14** Overview of the events of translation: the ribosome cycle.

80 nucleotides of mRNA. Still, even a small ORF of 1000 bases (which would encode a protein of  $\sim 35$  kDa) can bind more than 10 ribosomes and therefore direct the simultaneous synthesis of multiple polypeptides.

The ability of multiple ribosomes to function on a single mRNA explains the relatively limited abundance of mRNA in the cell (typically 1%–5% of total RNA). If an mRNA could be translated by only one ribosome at a time, then as few as 10% of the ribosomes would be engaged in protein synthesis in a typical cell. Instead, the association of multiple ribosomes with each mRNA ensures that the majority of the ribosomes are engaged in translation at any given time.

### New Amino Acids Are Attached to the Carboxyl Terminus of the Growing Polypeptide Chain

As we know, both polynucleotide and polypeptide chains have intrinsic polarities. Thus, for each of these molecules, we can ask which end of the chain is synthesized first. We learned in Chapters 9 and 13 that DNA and RNA are synthesized by adding each new nucleotide triphosphate to the 3' end of the growing polynucleotide chain (often referred to as synthesis in the 5'  $\rightarrow$  3' direction).



**FIGURE 15-15** A polyribosome.

What is the order of synthesis of a growing polypeptide chain? This was first determined in a classic experiment performed by Howard Dintzis that is described in Chapter 2. This experiment found that each new amino acid must be added to the carboxyl terminus of the growing polypeptide chain (often referred to as synthesis in the amino- to carboxy-terminal direction). As described in the next section, this directionality is a direct result of the chemistry of protein synthesis.

### Peptide Bonds Are Formed by Transfer of the Growing Polypeptide Chain from One tRNA to Another

The ribosome catalyzes a single chemical reaction—the formation of a peptide bond. This reaction occurs between the amino acid residue at the carboxyl-terminal end of the growing polypeptide and the incoming amino acid to be added to the chain. Both the growing chain and the incoming amino acid are attached to tRNAs; as a result, during peptide-bond formation, the growing polypeptide is continuously attached to a tRNA.

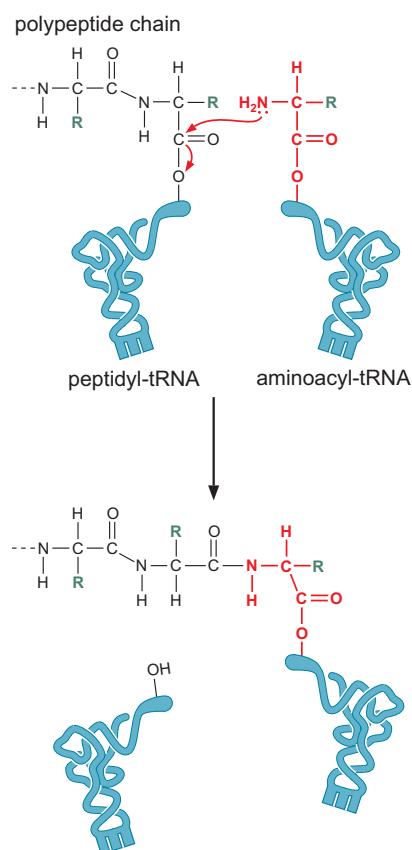
The actual substrates for each round of amino acid addition are two charged species of tRNAs—an aminoacyl-tRNA and a **peptidyl-tRNA**. As we discussed above in this chapter (see the section on Attachment of Amino Acids to tRNAs), the aminoacyl-tRNA is attached at its 3' end to the carboxyl group of the amino acid. The peptidyl-tRNA is attached in exactly the same manner (at its 3' end) to the carboxyl terminus of the growing polypeptide chain. The bond between the aminoacyl-tRNA and the amino acid is *not* broken during the formation of the next peptide bond. Instead, the bond between the peptidyl-tRNA and the growing polypeptide chain is broken as the growing chain is attached to the amino group of the amino acid attached to the aminoacyl-tRNA to form a new peptide bond.

To catalyze peptide-bond formation, the 3' ends of these two tRNAs are brought into close proximity by the ribosome. The resulting tRNA positioning allows the amino group of the amino acid attached to aminoacyl-tRNA to attack the carbonyl group of the most carboxyl-terminal amino acid attached to the peptidyl-tRNA. The result of this nucleophilic attack is the formation of a new peptide bond between the amino acids attached to the tRNAs and the release of the polypeptide chain from the peptidyl tRNA (Fig. 15-16). There are two consequences of this method of polypeptide synthesis. First, this mechanism of peptide-bond formation requires that the amino terminus of the protein be synthesized before the carboxyl terminus. Second, the growing polypeptide chain is transferred from the peptidyl-tRNA to the aminoacyl-tRNA. For this reason, the reaction to form a new peptide bond is called the **peptidyl transferase reaction**.

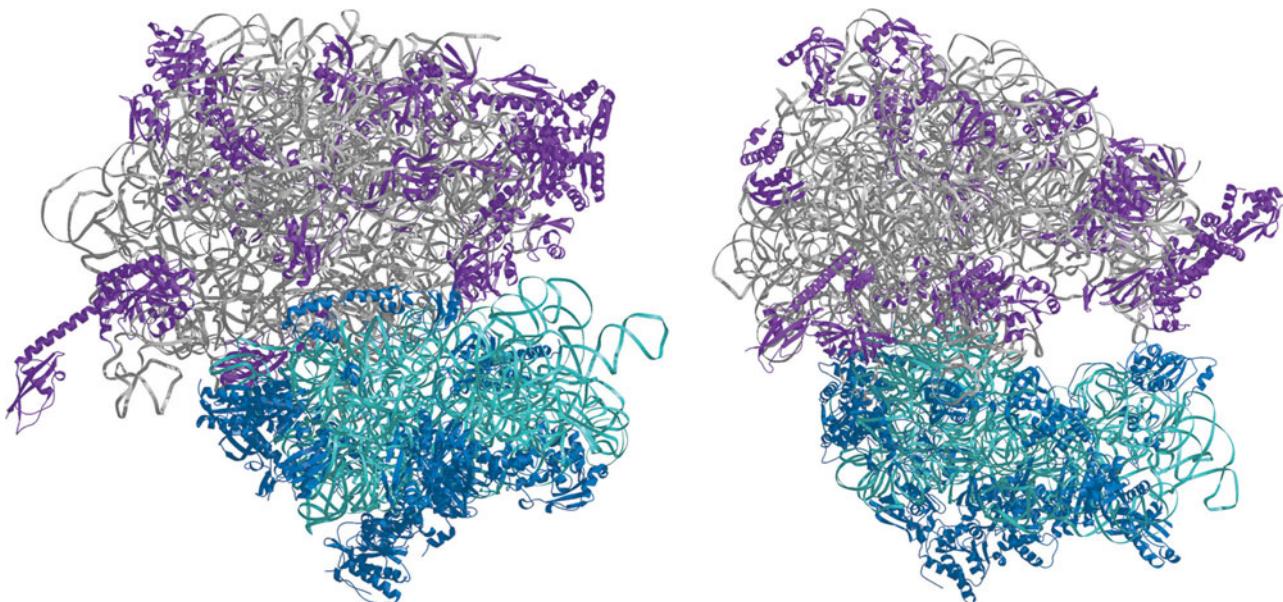
Interestingly, peptide-bond formation takes place without the simultaneous hydrolysis of a nucleoside triphosphate. This is because peptide-bond formation is driven by breaking the high-energy acyl bond that joins the growing polypeptide chain to the tRNA. Recall that this bond was created during the tRNA synthetase–catalyzed reaction that is responsible for charging tRNA. The charging reaction involves the hydrolysis of a molecule of ATP. Thus, the energy for peptide-bond formation originates from the molecule of ATP that was hydrolyzed during the tRNA charging reaction (Fig. 15-6).

### Ribosomal RNAs Are Both Structural and Catalytic Determinants of the Ribosome

Although the ribosome and its basic functions were discovered more than 40 years ago, the determination of numerous high-resolution, 3D structures of



**FIGURE 15-16** The peptidyl transferase reaction.



**FIGURE 15-17** Two views of the ribosome. The 50S subunit is above the 30S subunit in both views. The cavity between the 50S and 30S subunits in the right-hand image represents the site of tRNA association (see Fig. 15-19b). The RNA component of the 50S subunit is shown in gray; the protein component of the 50S subunit is shown in purple; the RNA component of the 30S subunit is shown in light blue; the protein component of the 30S subunit is shown in dark blue. (Yusupov M.M. et al. 2001. *Science* 292: 883–896.) Images prepared with MolScript, BobScript, and Raster3D.

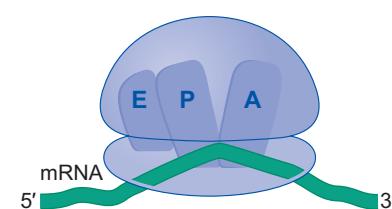
the ribosome has vastly increased our understanding of the workings of this molecular machine (Fig. 15-17). Perhaps the most important outcome of these studies is the finding that rRNAs are much more than structural components of the ribosome. Rather, they are directly responsible for the key functions of the ribosome. The most obvious example of this is the demonstration that the peptidyl transferase center is composed almost entirely of RNA, as discussed in detail later. RNA also plays a central role in the function of the small subunit of the ribosome. The anticodon loops of the charged tRNAs and the codons of the mRNA contact the 16S rRNA, not the ribosomal proteins of the small subunit.

A further indication of the importance of RNA in the structure and function of the ribosome is that most ribosomal proteins are on the periphery of the ribosome, not in its interior (Fig. 15-17; see also Structural Tutorial 15-1). The core functional domains of the ribosome (the peptidyl transferase center and the decoding center) are composed either entirely or mostly from RNA. Portions of some ribosomal proteins do reach into the core of the subunits, where their function seems to be to stabilize the tightly packed rRNAs by shielding the negative charges of their sugar–phosphate backbones. Indeed, it is likely that the contemporary ribosome evolved from a primitive protein-synthesizing machine that was composed entirely of RNA and that the ribosomal proteins were added to enhance the function of this primordial RNA machine.

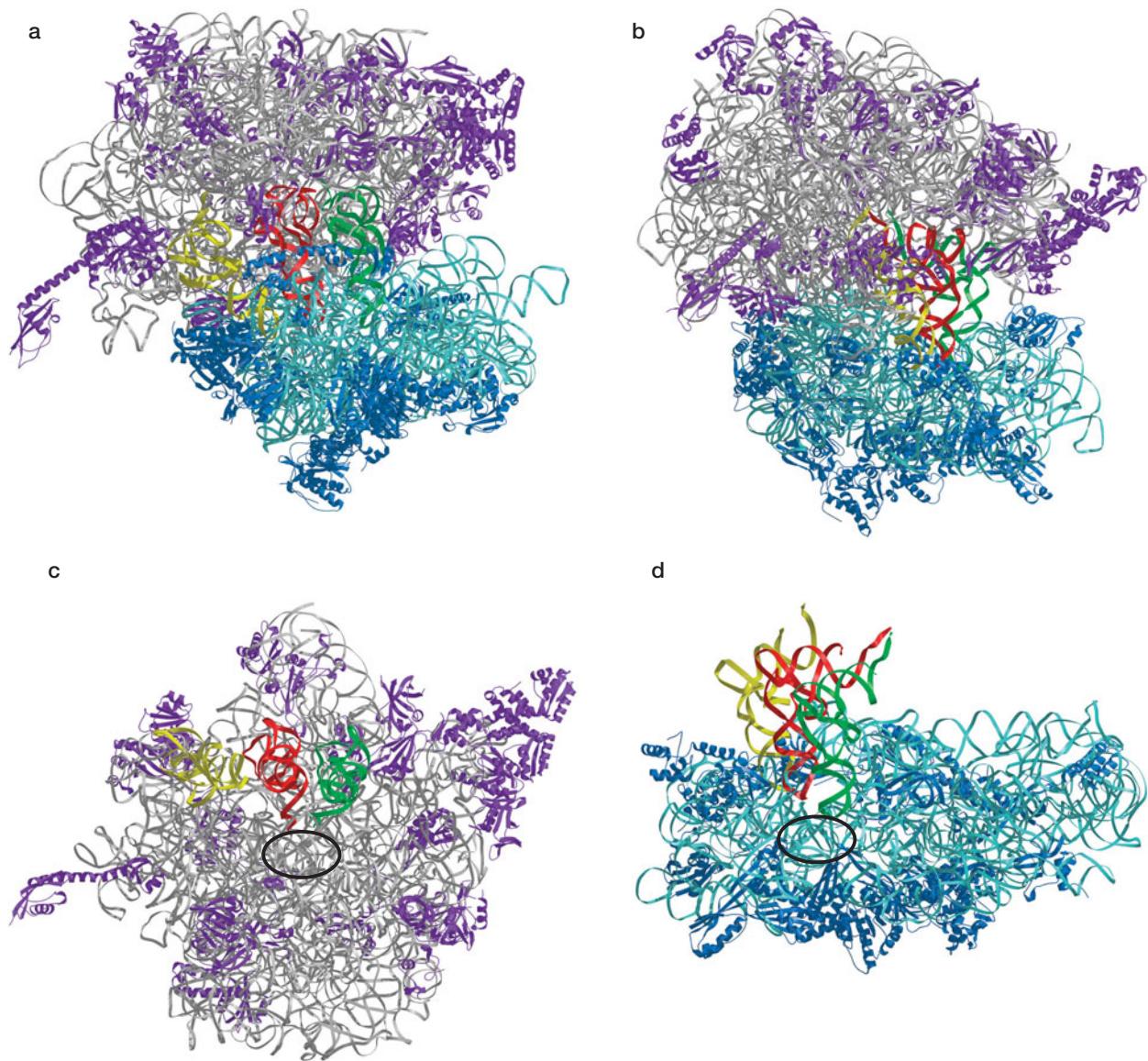
### The Ribosome Has Three Binding Sites for tRNA

To perform the peptidyl transferase reaction, the ribosome must be able to bind at least two tRNAs simultaneously. In fact, the ribosome contains three tRNA-binding sites, called the A-, P-, and E-sites (Figs. 15-18 and 15-19). The

WEB  
STRUCTURAL  
TUTORIAL



**FIGURE 15-18** The ribosome has three tRNA-binding sites. The schematic illustration of the ribosome shows the three binding sites (E, P, and A) that each spans the two subunits.



**FIGURE 15-19** Views of the 3D structure of the ribosome including three bound tRNAs. The E-, P-, and A-site tRNAs are shown in yellow, red, and green, respectively. The colors representing the RNA and protein components of the small and large subunits are the same as those in Figure 15-17. (a,b) Two views of the ribosome bound to the three tRNAs in the E-, P-, and A-sites. Note that the left (a) and right (b) views shown here correspond to those views of the ribosome shown in Figure 15-17. (c) The isolated 50S subunit with tRNAs as seen in the full ribosome. (This view is as if you were looking up at the large subunit from the small subunit.) The peptidyl transferase center is circled. (d) The isolated 30S subunit with tRNAs as seen in the full ribosome. The decoding center is circled. (Yusupov M.M. et al. 2001. *Science* **292**: 883–896.) Images prepared with MolScript, BobScript, and Raster3D.

**A-site** is the binding site for the aminoacylated-tRNA, the **P-site** is the binding site for the peptidyl-tRNA, and the **E-site** is the binding site for the tRNA that is released after the growing polypeptide chain has been transferred to the aminoacyl-tRNA (E is for “exiting”).

Each tRNA-binding site is formed at the interface between the large and the small subunits of the ribosome (Fig. 15-19a,b). In this way, the bound tRNAs can span the distance between the peptidyl transferase center in

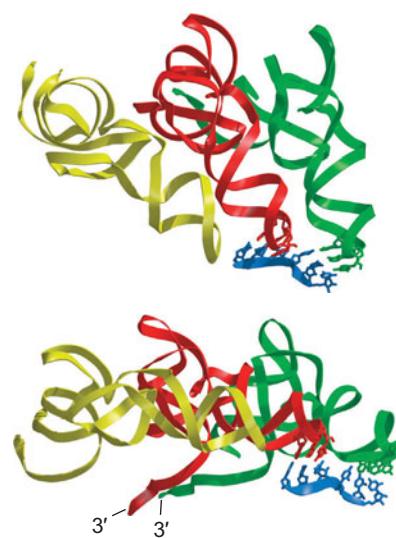
the large subunit (Fig. 15-19c) and the decoding center in the small subunit (Fig. 15-19d). The 3' ends of the tRNAs that are coupled to the amino acid or to the growing peptide chain are adjacent to the large subunit. The anticodon loops of the bound tRNAs are located adjacent to the small subunit.

### Channels through the Ribosome Allow the mRNA and Growing Polypeptide to Enter and/or Exit the Ribosome

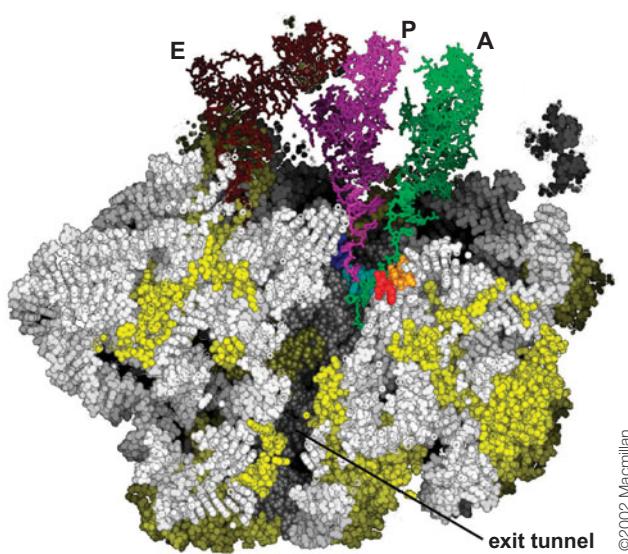
Both the decoding center and the peptidyl transferase center are buried within the intact ribosome. Yet, mRNA must be threaded through the decoding center during translation, and the nascent polypeptide chain must escape from the peptidyl transferase center. How do these polymers enter (in the case of mRNA) and exit the ribosome? The answer is provided by the structure of the ribosome, which reveals “tunnels” in and out of the ribosome.

The mRNA enters and exits the decoding center through two narrow channels in the small subunit. The entry channel is only wide enough for unpaired RNA to pass through. This feature ensures that the mRNA is in a single-stranded form as it enters the decoding center by removing any intramolecular base-pairing interactions that may have formed in the mRNA. In between the two channels is a region that is accessible to tRNAs and where adjacent codons can bind to the aminoacyl-tRNA and peptidyl-tRNA in the A- and P-sites, respectively. Interestingly, there is a pronounced kink in the mRNA between the two codons that facilitates maintenance of the correct reading frame (Fig. 15-20). This kink places the vacant A-site codon in a distinctive position that ensures that the incoming aminoacyl-tRNA does not have access to bases immediately adjacent to the codon.

A second channel through the large subunit provides an exit path for the newly synthesized polypeptide chain (Fig. 15-21). As with the mRNA channel, the size of the peptide exit channel limits the conformation of the growing polypeptide chain. In this case, a polypeptide can form an  $\alpha$  helix within the channel, but other secondary structures (such as  $\beta$  sheets) and tertiary interactions can form only after the polypeptide exits the large ribosomal subunit. For this reason, the final 3D structure of a newly synthesized protein is not attained until after it is released from the ribosome.

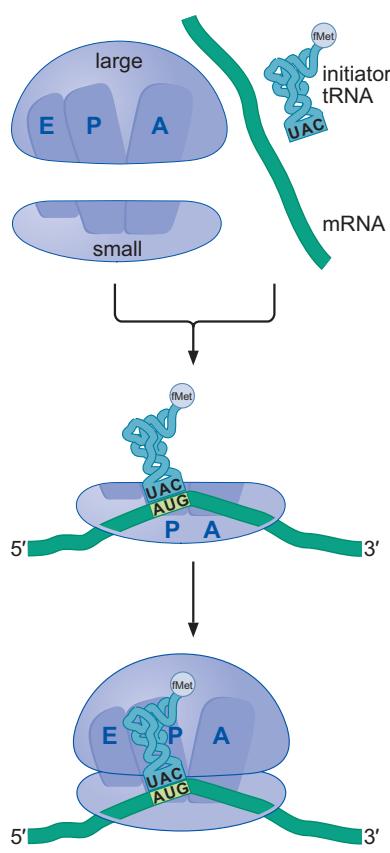


**FIGURE 15-20** The interaction between the A-site and P-site tRNAs and the mRNA within the ribosome. Two views of the structure of the mRNA and tRNAs are shown as they are found in the ribosome. For clarity, the ribosome is not shown. The E-, P-, and A-site tRNAs are shown in yellow, red, and green, respectively, and the mRNA is shown in blue. Only the bases involved in the codon–anticodon interaction are shown. The strong kink in the mRNA clearly distinguishes between the A-site and P-site codons. The close proximity of the 3' ends of the A-site and P-site tRNAs can be seen in the lower image. (Yusupov M.M. et al. 2001. *Science* **292**: 883–896.) Image prepared with Mol-Script, BobScript, and Raster3D.



©2002 Macmillan

**FIGURE 15-21** The polypeptide exit tunnel. In this image, the 50S subunit is cut in half to reveal the polypeptide exit tunnel. The rRNA is white; the ribosomal proteins are yellow. The three bound tRNAs are colored as follows: E-site (brown), P-site (purple), and A-site (green). The red and gold parts of the rRNA adjacent to the A-site tRNA are components of the peptidyl transferase center. (Courtesy of T. Martin Schmeing and Thomas Steitz; adapted, with permission, from Schmeing T.M. et al. 2002. *Nat. Struct. Biol.* **9**: 225–230. © Macmillan.)



**FIGURE 15-22** An overview of the events of translation initiation.

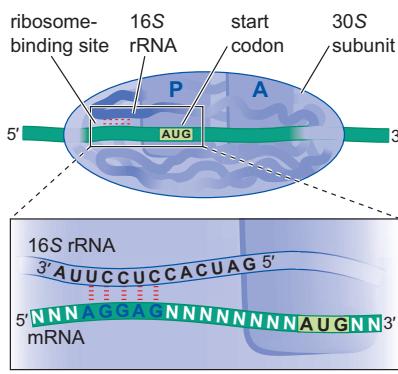
Now that we have described the four primary components of the translation process, the remainder of the chapter will focus on the individual stages of translation. Our description will proceed in order through the three stages of translation: initiation of the synthesis of a new polypeptide chain, elongation of the growing polypeptide, and termination of polypeptide synthesis. As we shall see, there are important similarities and differences between prokaryotes and eukaryotes in the strategies they use to perform these events. We shall consider the nature of the translation machinery from both kinds of cells in each of the following sections. As we have seen for DNA and RNA synthesis, although the ribosome is the center of activity, auxiliary factors play critical roles in each of the steps of translation and are required for protein synthesis to occur in a rapid and accurate fashion.

## INITIATION OF TRANSLATION

For translation to be successfully initiated, three events must occur (Fig. 15-22): the ribosome must be recruited to the mRNA; a charged tRNA must be placed into the P-site of the ribosome; and the ribosome must be precisely positioned over the start codon. The correct positioning of the ribosome over the start codon is critical because this establishes the reading frame for the translation of the mRNA. Even a 1-base shift in the location of the ribosome would result in the synthesis of a completely unrelated polypeptide (see the discussion of mRNA above and in Chapter 16). The dissimilar structures of prokaryotic and eukaryotic mRNAs result in distinctly different means of accomplishing these events. We start by addressing the initiation events in prokaryotes and then discuss the differences observed in eukaryotic cells.

### Prokaryotic mRNAs Are Initially Recruited to the Small Subunit by Base Pairing to rRNA

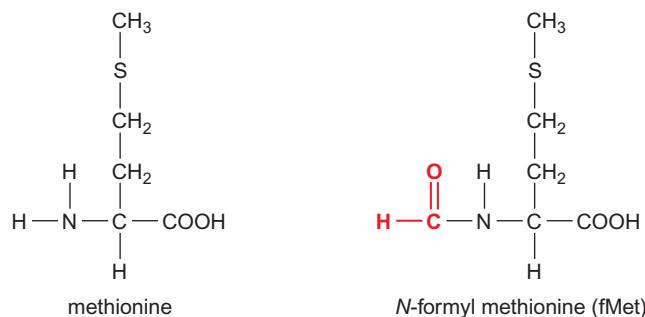
The assembly of the ribosome on an mRNA occurs one subunit at a time. The small subunit associates with the mRNA first. As described during our discussion of mRNA structure (see Fig. 15-2), in prokaryotes, the association of the small subunit with the mRNA is mediated by base-pairing interactions between the RBS and the 16S rRNA (Fig. 15-23). For ideally positioned RBSs, the small subunit is positioned on the mRNA such that the start codon will be in the P-site when the large subunit joins the complex. The large subunit joins its partner only at the very end of the initiation process, just before the formation of the first peptide bond. Thus, many of the key events of translation initiation occur in the absence of the full ribosome.



**FIGURE 15-23** The 16S rRNA interacts with the RBS to position the AUG in the P-site. This illustration shows an mRNA with the ideal separation between the RBS and the initiating AUG. This spacing places the AUG in the region of the P-site. Many mRNAs have non-ideal spacings leading to a reduced rate of translation initiation. Other mRNAs lack an RBS completely and recruit the ribosome by distinct mechanisms.

### A Specialized tRNA Charged with a Modified Methionine Binds Directly to the Prokaryotic Small Subunit

Translation initiation is the only time a tRNA binds to the P-site without previously occupying the A-site. This event requires a special tRNA known as the **initiator tRNA**, which base-pairs with the start codon—usually AUG or GUG. AUG and GUG have a different meaning when they occur within an ORF, where they are read by tRNAs for methionine ( $tRNA^{Met}$ ) and valine ( $tRNA^{Val}$ ), respectively (see Chapter 16). Although the initiator tRNA is first charged with a methionine, a formyl group is rapidly added to the methionine amino group by a separate enzyme (Met-tRNA transformylase). Thus,



**FIGURE 15-24** Methionine and *N*-formyl methionine.

rather than valine or methionine, the initiator tRNA is coupled to ***N*-formyl methionine** (Fig. 15-24). The charged initiator tRNA is referred to as **fMet-tRNA<sub>i</sub><sup>fMet</sup>**.

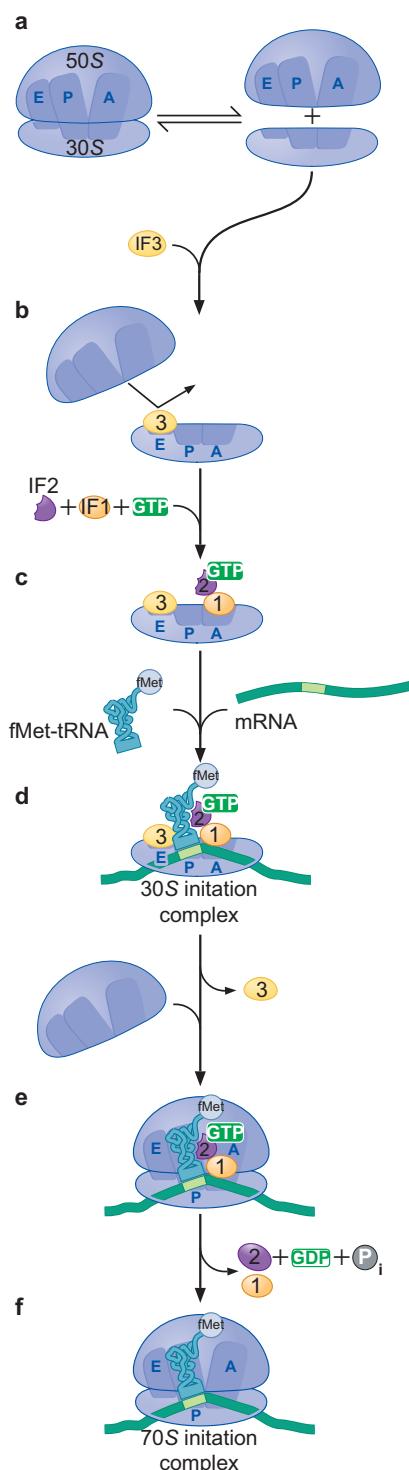
Because *N*-formyl methionine is the first amino acid to be incorporated into a polypeptide chain, one might think that all prokaryotic proteins have a formyl group at their amino termini. This is not the case, however, because an enzyme known as a **deformylase** removes the formyl group from the amino terminus during or after the synthesis of the polypeptide chain. In fact, many mature prokaryotic proteins do not even start with a methionine; aminopeptidases often remove the amino-terminal methionine as well as one or two additional amino acids.

### Three Initiation Factors Direct the Assembly of an Initiation Complex That Contains mRNA and the Initiator tRNA

The initiation of prokaryotic translation commences with the small subunit and is catalyzed by three **translation initiation factors** called **IF1**, **IF2**, and **IF3**. Each factor facilitates a key step in the initiation process.

- **IF1** prevents tRNAs from binding to the portion of the small subunit that will become part of the A-site.
- **IF2** is a GTPase (a protein that binds and hydrolyzes GTP) that interacts with three key components of the initiation machinery: the small subunit, IF1, and the charged initiator tRNA (fMet-tRNA<sub>i</sub><sup>fMet</sup>). By interacting with these components, IF2 facilitates the association of fMet-tRNA<sub>i</sub><sup>fMet</sup> with the small subunit and prevents other charged tRNAs from associating with the small subunit.
- **IF3** binds to the small subunit and blocks it from reassociating with a large subunit. Because initiation requires a free small subunit, the binding of IF3 is critical for a new cycle of translation. IF3 becomes associated with the small subunit at the end of a previous round of translation when it helps to dissociate the 70S ribosome into its large and small subunits.

Each of the initiation factors binds at, or near, one of the three tRNA-binding sites on the small subunit. Consistent with its role in blocking the binding of charged tRNAs to the A-site, IF1 binds directly to the portion of the small subunit that will become the A-site. IF2 binds to IF1 and reaches over the A-site into the P-site to contact the fMet-tRNA<sub>i</sub><sup>fMet</sup>. Finally, IF3 occupies the part of the small subunit that will become the E-site. Thus, of the three potential tRNA-binding sites on the small subunit, only the P-site is capable of binding a tRNA in the presence of the initiation factors.



**FIGURE 15-25** A summary of translation initiation in prokaryotes.

With all three initiation factors bound, the small subunit is prepared to bind to the mRNA and the initiator tRNA (Fig. 15-25). These two RNAs can bind in either order and independently of each other. As discussed above, binding to the mRNA involves base pairing between the RBS and the 16S rRNA in the small subunit. Meanwhile, binding fMet-tRNA<sub>i</sub><sup>fMet</sup> to the small subunit is facilitated by its interactions with IF2 bound to GTP and (once the mRNA is bound) base pairing between the anticodon and the start codon of the mRNA. Similarly, base pairing between the fMet-tRNA<sub>i</sub><sup>fMet</sup> and the mRNA serves to position the start codon in the P-site.

The last step of initiation involves the association of the large subunit to create the **70S initiation complex**. When the start codon and fMet-tRNA<sub>i</sub><sup>fMet</sup> base-pair, the small subunit undergoes a change in conformation. This altered conformation results in the release of IF3. In the absence of IF3, the large subunit is free to bind to the small subunit with its cargo of IF1, IF2, mRNA, and fMet-tRNA<sub>i</sub><sup>fMet</sup>. In particular, IF2 acts as an initial docking site of the large subunit, and this interaction subsequently stimulates the GTPase activity of IF2-GTP. IF2 bound to GDP has reduced affinity for the ribosome and the initiator tRNA, leading to the release of IF2-GDP as well as IF1 from the ribosome. Thus, the net result of initiation is the formation of an intact (70S) ribosome assembled at the start site of the mRNA with fMet-tRNA<sub>i</sub><sup>fMet</sup> in the P-site and an empty A-site. The ribosome-mRNA complex is now poised to accept a charged tRNA into the A-site and commence polypeptide synthesis.

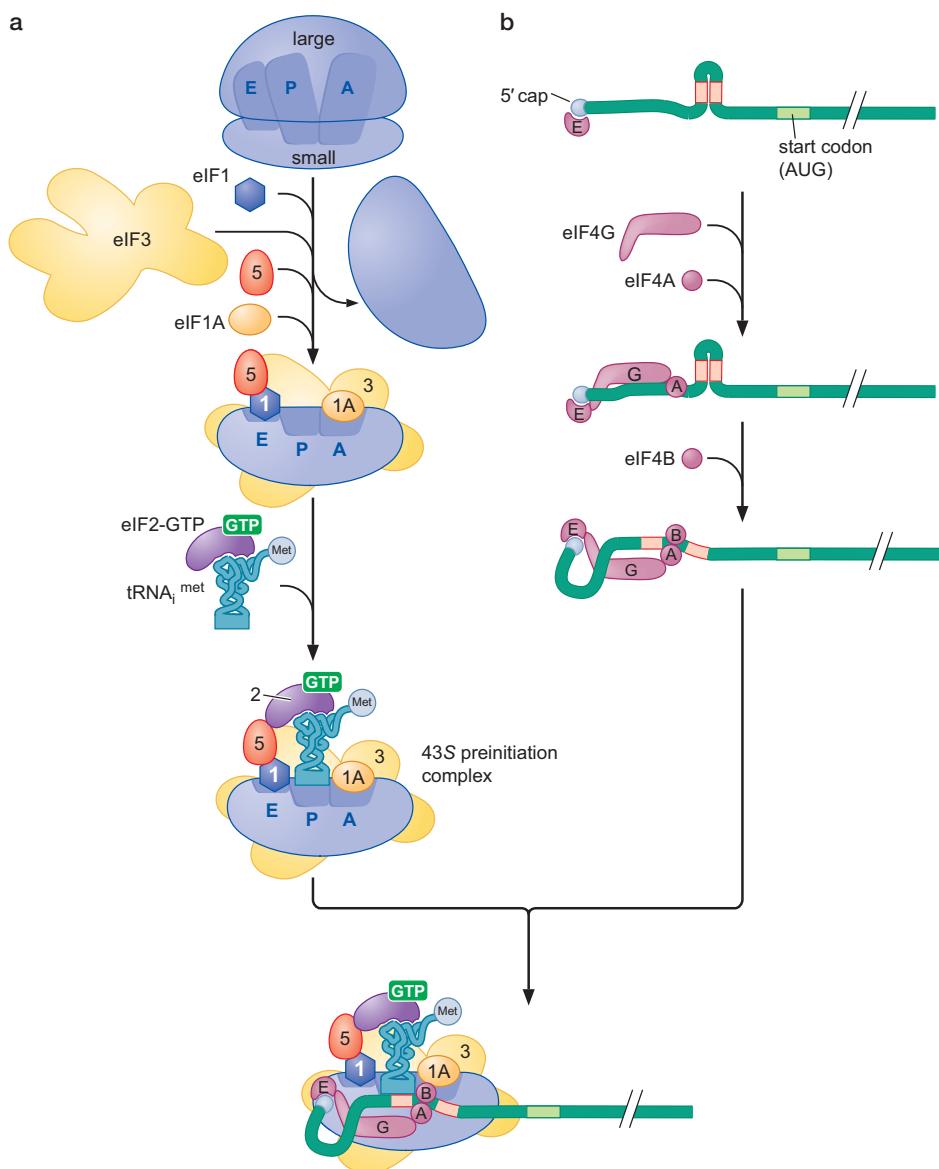
### Eukaryotic Ribosomes Are Recruited to the mRNA by the 5' Cap

Initiation of translation in eukaryotes is similar to prokaryotic initiation in many ways. Both use a start codon and a dedicated initiator tRNA, and both use initiation factors to form a complex with the small ribosomal subunit that assembles on the mRNA before addition of the large subunit. Nevertheless, eukaryotes use a fundamentally distinct method to recognize the mRNA and the start codon, which has important consequences for eukaryotic translation.

In eukaryotes, the small subunit is already associated with an initiator tRNA when it is recruited to the capped 5' end of the mRNA. It then “scans” along the mRNA in a 5' → 3' direction until it reaches the first 5'-AUG-3' (see the discussion of the Kozak sequence in the preceding section on mRNA), which it recognizes as the start codon. Thus, in most instances (see Box 15-3, uORFs and IRESs: Exceptions That Prove the Rule), only the first AUG can be used as the start site of translation in eukaryotic cells. Note that this method of initiation is consistent with the fact that the vast majority of eukaryotic RNAs encode a single polypeptide (monocistronic); recognition of an internal start codon is generally neither required nor possible.

As we have seen for other molecular processes (such as promoter recognition during transcription), eukaryotic cells require more auxiliary proteins to drive the initiation process than do prokaryotes. The events of initiation can be broken down into four steps. First, in contrast to the situation in prokaryotes, in eukaryotic cells, binding of the initiator tRNA to the small subunit *always* precedes association with the mRNA (Fig. 15-26a). Second, a separate set of auxiliary factors mediates the recognition of the mRNA. Third, the small ribosomal subunit bound to the initiator tRNA scans the mRNA for the first AUG sequence. Finally, the large subunit of the ribosome is recruited after the initiator tRNA base-pairs with the start codon. We now describe these events in detail.

As the eukaryotic ribosome completes a cycle of translation, it dissociates into free large and small subunits, and four initiation factors—eIF1,



**FIGURE 15-26** Assembly of the eukaryotic small ribosomal subunit and initiator tRNA onto the mRNA.

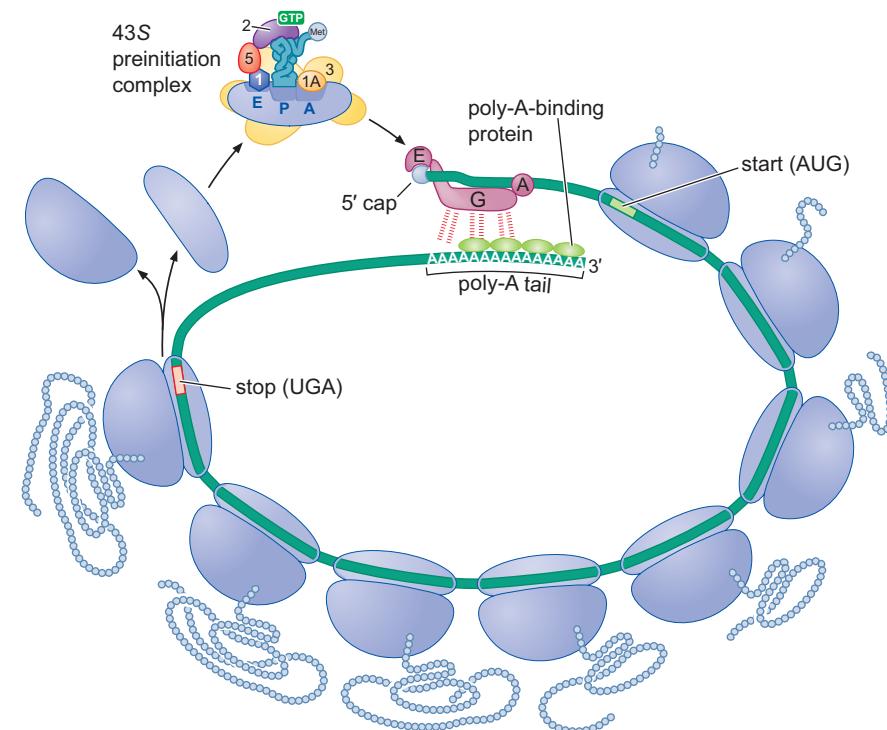
eIF1A, eIF3, and eIF5—bind to the small subunit. Together, eIF1, eIF1A, and eIF5 act in an analogous manner to the prokaryotic initiation factors IF3 and IF1 to prevent both large subunit binding and tRNA binding to the A-site. The initiator tRNA is escorted to the small subunit by the three-subunit GTP-binding protein eIF2. Like IF2, eIF2 will bind the initiator tRNA only in the GTP-bound state. The complex between the initiator tRNA and EIF2 is called the **ternary complex** (TC). For eukaryotes, the initiator tRNA (which, like its bacterial counterpart, is distinct from the tRNA<sub>i</sub> used after initiation) is charged with methionine, *not* N-formyl methionine, and is referred to as Met-tRNA<sub>i</sub><sup>Met</sup>. eIF2 positions the Met-tRNA<sub>i</sub><sup>Met</sup> in the P-site of the initiation factor–bound small subunit, resulting in the formation of the **43S preinitiation complex (43S PIC)**. It is noteworthy that eIF3 is almost as large as the entire 40S subunit but primarily binds the side of the small subunit near the RNA entry and exit sites. Nevertheless, eIF3 interacts with every member of the 43S PIC including the

initiator tRNA and, thus, facilitates many of the interactions involved in 43S PIC assembly.

In a separate series of reactions, the mRNA is prepared for recognition by the small subunit. This process begins with recognition of the 5' cap by the cap-binding protein eIF4E. A series of additional initiation factors is then recruited. eIF4G binds to both eIF4E and the mRNA, whereas eIF4A binds eIF4G and the mRNA (see Fig. 15-26b). The association of eIF4G with eIF4E is particularly important—the overall level of translation in the cell is controlled at this step by a family of proteins that compete with eIF4G binding called eIF4E-binding proteins (see Regulation of Translation later in this chapter). This complex is joined by eIF4B, which activates the RNA helicase activity of eIF4A. The helicase unwinds any secondary structures (such as hairpins) that may have formed at the end of the mRNA. Removal of secondary structures is critical because the 5' end of the mRNA must be unstructured to bind to the small subunit. Finally, interactions between the eIF4G bound to the unstructured mRNA and the initiation factors (particularly eIF3) bound to the small subunit recruit the 43S preinitiation complex to the mRNA to form the **48S preinitiation complex**.

### Translation Initiation Factors Hold Eukaryotic mRNAs in Circles

The presence of a poly-A tail contributes to the efficiency of eukaryotic translation. In addition to binding to the 5' end of eukaryotic mRNAs, the initiation factors that prepare the mRNA are also associated with the 3' end of the mRNA (Fig. 15-27). These interactions are primarily mediated by eIF4G, which binds both directly to the 3' end of the mRNA and to the **poly-A-binding protein** that coats the poly-A tail. These interactions result



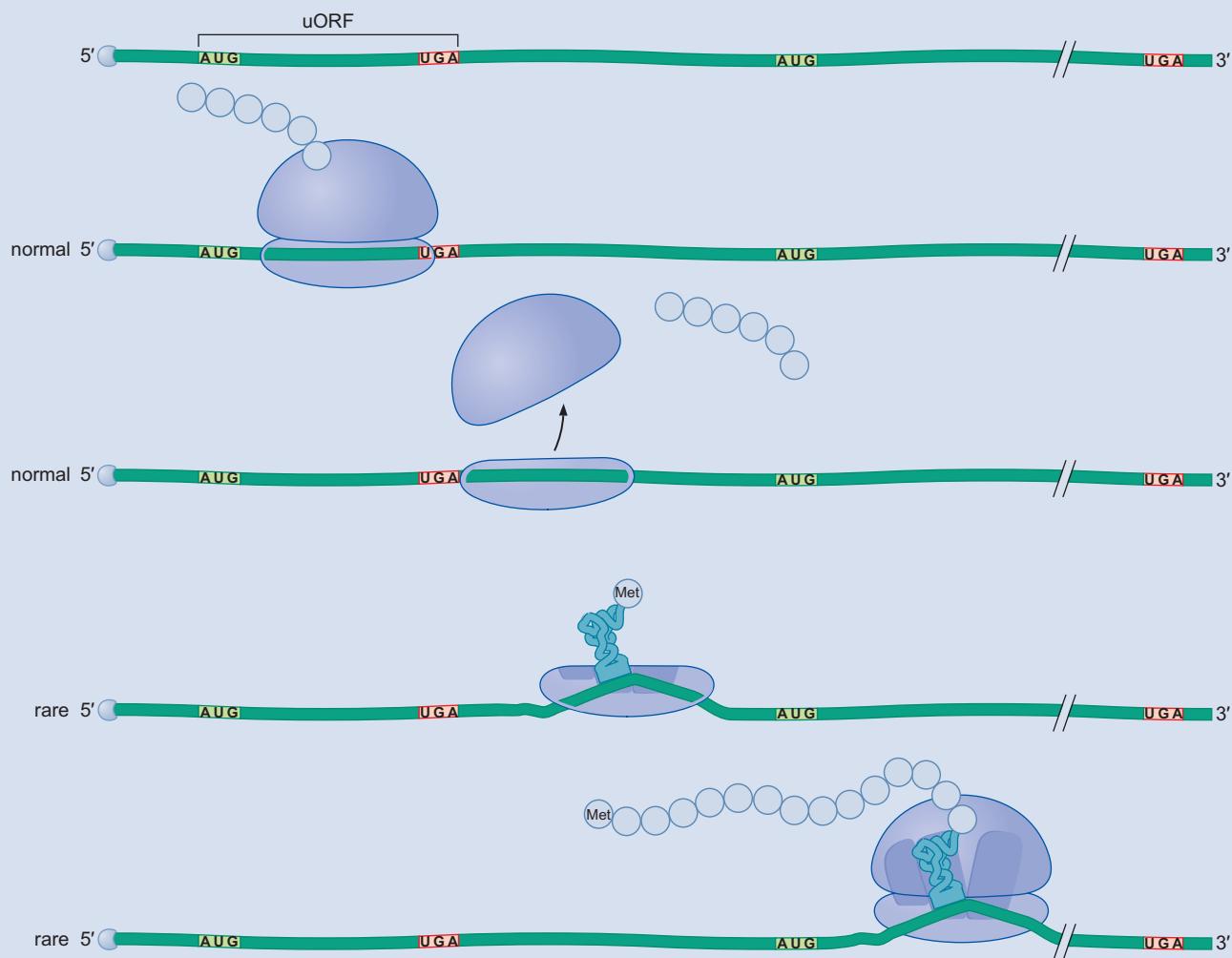
**FIGURE 15-27** A model for the circularization of eukaryotic mRNA. Circularization is mediated by interactions between eIF4G, the poly-A-binding protein, and the poly-A tail.

**Box 15-3 uORFs and IRESs: Exceptions That Prove the Rule**

Not all eukaryotic polypeptides are encoded by an ORF that starts with the AUG that is most proximal to the 5' terminus. In some cases, the first AUG is in a poor sequence context, resulting in its frequent bypass. In other cases, short, upstream ORFs (uORFs, typically encoding peptides less than 10 amino acids long) are translated, but a subset (typically <50%) of the 40S subunits is retained on the mRNA after termination of uORF translation. The short length of such uORFs allows interactions between initiation factors (e.g., eIF4G and eIF3) that tether the 40S subunit to the mRNA to be retained after termination (Box 15-3 Fig. 1). The retained 40S subunit continues scanning for the next AUG but can only identify an AUG after binding to a new ternary complex (TC; eIF2-initiator tRNA) because the anticodon of the initiator tRNA is required to detect an AUG (see Box 15-3 Fig. 1). In general, uORFs reduce but do not eliminate translation of the primary ORF. We discuss a specific example of how uORFs can be used to regulate translation later in this chapter.

A more extreme example of initiating translation at sites downstream from the most 5'-proximal AUG are internal ribosome entry sites (IRESs). IRESs are RNA sequences that function like prokaryotic RBSs. They recruit the small subunit to bind and initiate even in the absence of a 5' cap (Box 15-3 Fig. 2). IRESs are often encoded in viral mRNAs that lack a 5'-cap end and have a need to exploit the sequences of their genome maximally. By using an IRES, a viral mRNA can encode more than one protein, reducing the need for extended transcriptional regulatory sequences for each protein-coding sequence. More importantly, by bypassing the requirement for one or more initiation factors, an IRES can continue functioning in the absence of the factor. This distinct requirement can be exploited to allow only a small subset of IRES-containing mRNAs to be translated in a cell that is lacking an initiation factor. For example, during apoptosis (or programmed cell death), newly activated proteases destroy eIF4E, but a subset of proteins needed for apoptosis continues to be translated because of the presence of IRES sequences in their 5'-untranslated regions (UTRs) that bypass the need for eIF4E.

Different IRES sequences work by different mechanisms. At least one viral IRES directly binds to eIF4G, bypassing the

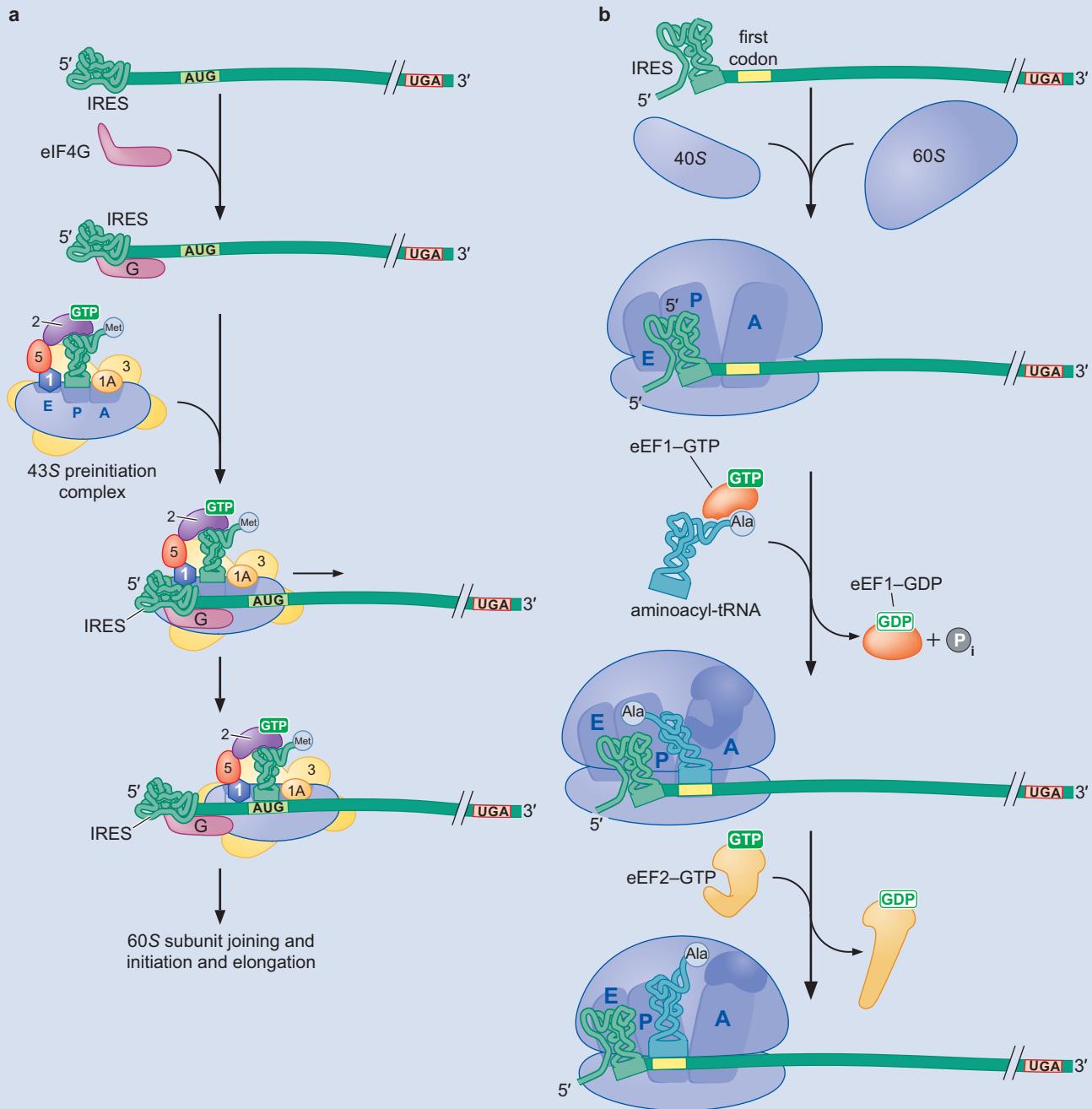


**BOX 15-3 FIGURE 1** uORFs regulate translation of downstream ORFs. In some cases, after a ribosome translates a uORF, the small subunit remains on the mRNA and resumes scanning for a second AUG. It can only identify a second AUG when it binds a new initiator rRNA.

**Box 15-3** (Continued)

normal requirement of the 5'-cap-binding protein eIF4E to recruit eIF4G (Box 15-3 Fig. 2a). The most extreme type of IRES is exemplified by the cricket paralysis virus mRNA. The 5'-UTR of this mRNA forms a complex RNA structure that mimics a tRNA bound to a mRNA in the PE hybrid state and binds directly to the P-site of the 40S subunit. In this way, the mRNA bypasses the need for all initiation factors and an initiator

tRNA (Box 15-3 Fig. 2b). Importantly, the RNA sequence downstream from this structure is placed in the A-site of the decoding center, allowing it to act as the first codon. The existence of IRESs that require no initiation factors has led to the hypothesis that early in evolution all mRNAs had such IRESs and that the initiation factors evolved later to make translation more efficient and versatile.



**BOX 15-3 FIGURE 2** IRESs bypass normal requirements for initiation of translation. Viruses frequently encode IRES sequences that fold into RNA structures that bypass the need for one or more eukaryotic translation factors. (a) The poliovirus IRES bypasses the requirement for the 5' cap by directly binding to eIF4G. (b) The Cricket paralysis virus encodes an mRNA with an elaborate IRES that folds to resemble a tRNA that binds directly to the P-site of the 40S and 60S ribosomal subunits. The downstream mRNA is placed into the A-site and encodes for an Ala as the first amino acid.

in the mRNA being held in a circular configuration with the 5' and 3' ends of the molecule in close proximity. Consistent with the poly-A tail contributing to efficient translation of mRNA, these interactions enhance several steps of initiation including eIF4E binding to the mRNA cap and large subunit recruitment. Importantly, the interactions of eIF4G and the poly-A-binding protein with the mRNA are maintained through multiple rounds of translation. In addition, this mRNA confirmation has the added benefit of locating recently terminated ribosomes near the AUG, presumably enhancing reinitiation.

### The Start Codon Is Found by Scanning Downstream from the 5' End of the mRNA

Once assembled at the 5' end of the mRNA, the small subunit and its associated factors move along the mRNA in a 5'→3' direction in an ATP-dependent process that is stimulated by the eIF4A/B-associated RNA helicase (Fig. 15-28). During this movement, the small subunit “scans” the mRNA for the first start codon. The start codon is recognized through base pairing between the anticodon of the initiator tRNA and the start codon. The importance of this interaction in identifying the start codon shows why it is critical that the initiator tRNA bind to the small subunit *before* it binds to the mRNA. Correct base pairing changes the conformation of the 48S complex, leading to release of eIF1 and a change in conformation of eIF5. Both of these events stimulate eIF2 to hydrolyze its associated GTP. In its GDP-bound state, eIF2 no longer binds the initiator tRNA and is released from the small subunit along with eIF5.

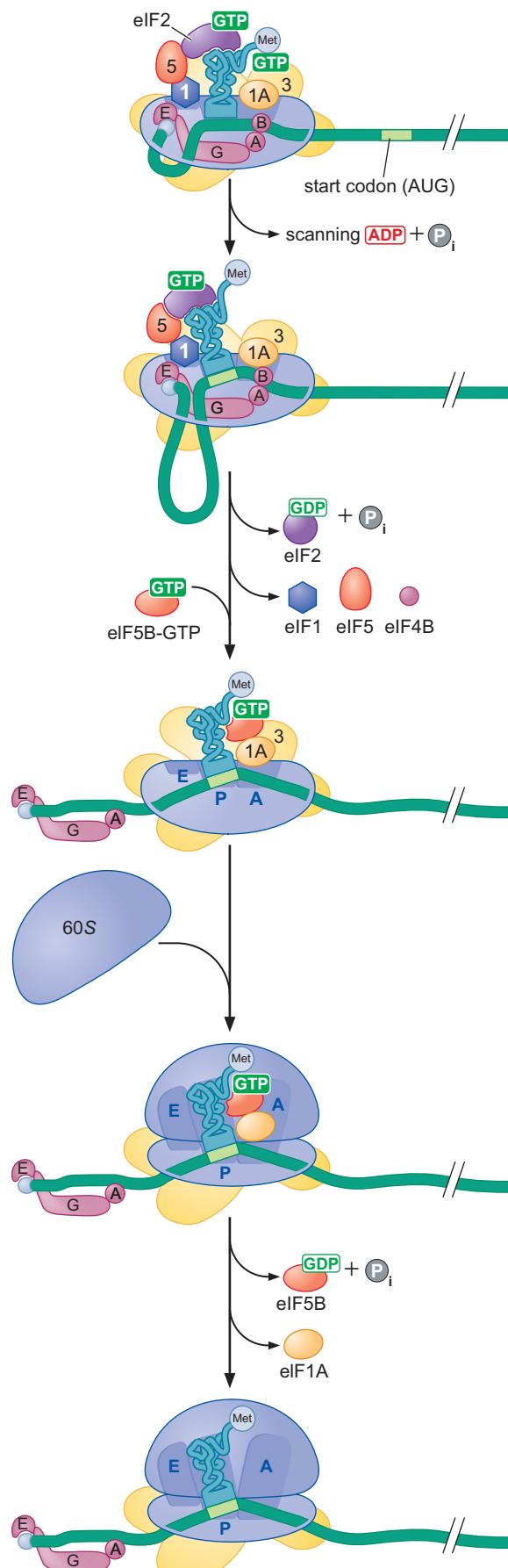
Loss of eIF2 allows the binding of a second GTP-regulated, initiator tRNA–binding protein called eIF5B. Upon binding the initiator tRNA, eIF5B · GTP stimulates the association of the 60S subunit with the correctly positioned 40S subunit. This association is possible because the factors that previously prevented this association (eIF1 and eIF5) have been released. As in the prokaryotic situation, binding of the large subunit leads to the release of the remaining initiation factors by stimulating GTP hydrolysis by eIF5B. As a result of these events, the Met-tRNA<sub>i</sub><sup>Met</sup> is placed in the P-site of the resulting **80S initiation complex** and the ribosome is now ready to accept a charged tRNA into its A-site and form the first peptide bond.

Although initiation in eukaryotic cells involves many more auxiliary factors, there are clear analogs of the bacterial initiation factors. Both IF1 and eIF1A bind to the A-site throughout the initiation process to prevent interaction of tRNAs with this region prematurely. The function of IF2 is split between eIF2 and eIF5B: eIF2 mediates initiator tRNA recruitment and detects pairing with the initiating AUG, and eIF5B performs the IF2 functions involved in large subunit recruitment. And like IF2, both eIF2 and eIF5B are regulated by the nucleotide they are bound to. Finally, IF3 and eIF1 both bind to the P-site of the small subunit and are both released upon base pairing of the initiator tRNA with the AUG.

## TRANSLATION ELONGATION

---

Once the ribosome is assembled with the charged initiator tRNA in the P-site, polypeptide synthesis can begin. There are three key events that must occur for the correct addition of each amino acid (Fig. 15-29). First, the correct aminoacyl-tRNA is loaded into the A-site of the ribosome as dictated by the A-site codon. Second, a peptide bond is formed between the aminoacyl-tRNA



**FIGURE 15-28** Identification of the initiating AUG by the 48S PIC and large subunit joining during eukaryotic translation initiation. See the text for a complete description.

in the A-site and the peptide chain that is attached to the peptidyl-tRNA in the P-site. This peptidyl transferase reaction, as we have seen, results in the transfer of the growing polypeptide from the tRNA in the P-site to the amino acid moiety of the charged tRNA in the A-site. Third, the resulting peptidyl-tRNA in the A-site and its associated codon must be **translocated** to the P-site so that the ribosome is poised for another cycle of codon recognition and peptide-bond formation. As with the original positioning of the mRNA, this shift must occur precisely to maintain the correct reading frame of the message. Two auxiliary proteins known as **elongation factors** control these events. Both of these factors use the energy of GTP binding and hydrolysis to enhance the rate and accuracy of ribosome function.

Unlike the initiation of translation, the mechanism of elongation is highly conserved between prokaryotic and eukaryotic cells. We limit our discussion to translation elongation in prokaryotes, which is understood in the greatest detail, but the events that occur in eukaryotic cells are similar to those in prokaryotes, both in the factors involved and in their mechanism of action.

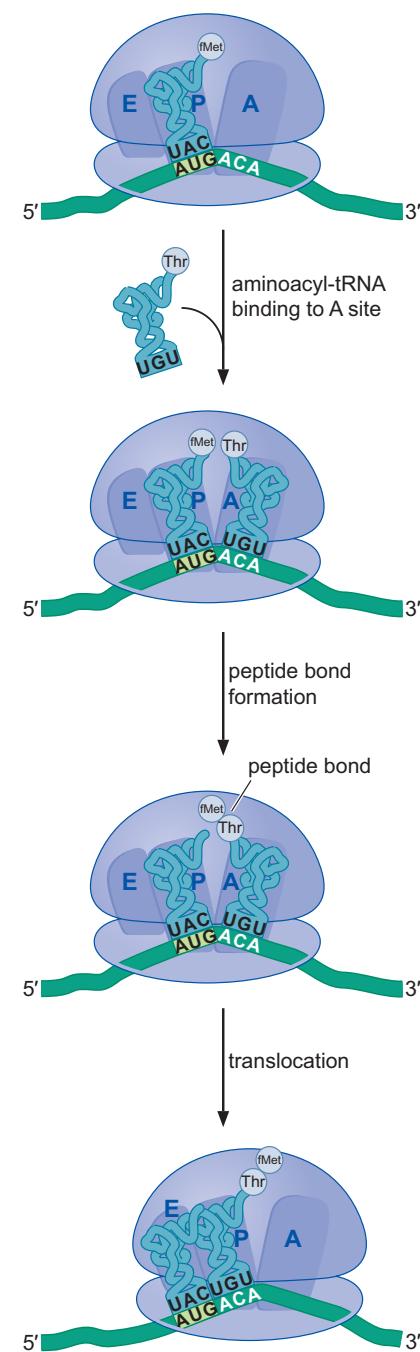
### Aminoacyl-tRNAs Are Delivered to the A-Site by Elongation Factor EF-Tu

Aminoacyl-tRNAs do not bind to the ribosome on their own. Instead, they are “escorted” to the ribosome by the elongation factor **EF-Tu** (Fig. 15-30). Once a tRNA is aminocylated, EF-Tu binds to the tRNA’s 3’ end, masking the coupled amino acid. This interaction prevents the bound aminoacyl-tRNA from participating in peptide-bond formation until it is released from EF-Tu.

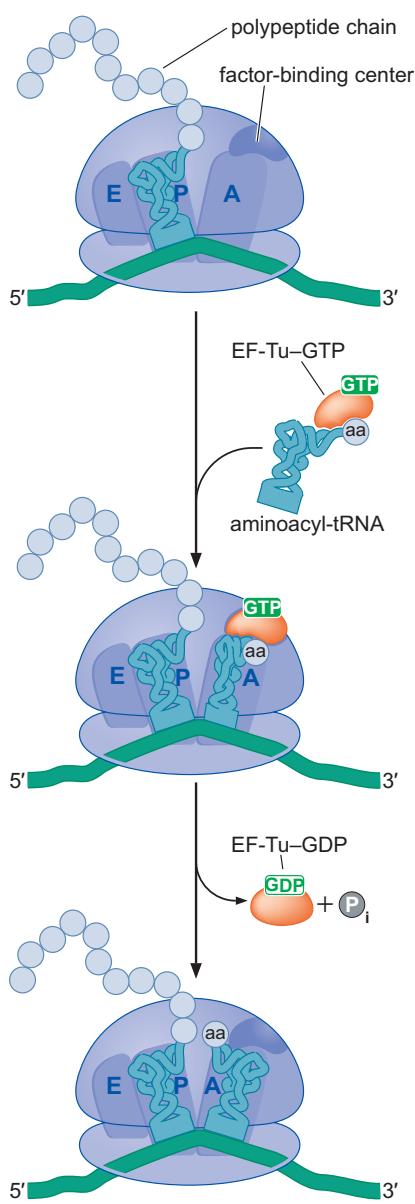
Like the initiation factor IF2, the elongation factor EF-Tu binds and hydrolyzes GTP, and the type of guanine nucleotide bound governs its function. EF-Tu can only bind to an aminoacyl-tRNA when it is associated with GTP. EF-Tu bound to GDP, or lacking any bound nucleotide, shows little affinity for aminoacyl-tRNAs. Thus, when EF-Tu hydrolyzes its bound GTP, any associated aminoacyl-tRNA is released. On its own, EF-Tu bound to an aminoacyl-tRNA does not hydrolyze GTP at a significant rate. Instead, the EF-Tu GTPase is activated when it associates with the same domain on the large subunit of the ribosome that activates the IF2 GTPase when the large subunit joins the initiation complex. This domain is known as the **factor-binding center**. EF-Tu only interacts with the factor-binding center after the tRNA enters the A-site *and* a correct codon–anticodon match is made. At this point, EF-Tu hydrolyzes its bound GTP and is released from the ribosome (Fig. 15-30). As we shall discuss later, control of GTP hydrolysis by EF-Tu is critical to the specificity of translation.

### The Ribosome Uses Multiple Mechanisms to Select against Incorrect Aminoacyl-tRNAs

The error rate of translation is between  $10^{-3}$  and  $10^{-4}$ . That is, no more than one in every 1000 amino acids incorporated into protein is incorrect. The ultimate basis for the selection of the correct aminoacyl-tRNA is the base pairing between the charged tRNA and the codon displayed in the A-site of the ribosome. Despite this, the energy difference between a correctly formed codon–anticodon pair and that of a near match cannot account for this level of accuracy. In many instances, only one of the three possible base pairs in the anticodon–codon interaction is mismatched, yet the ribosome rarely allows such mismatched aminoacyl-tRNAs to continue in the



**FIGURE 15-29** Summary of the steps of translation elongation.



**FIGURE 15-30** EF-Tu escorts aminoacyl-tRNA to the A-site of the ribosome. Charged tRNAs are bound to EF-Tu-GTP as they first interact with the A-site of the ribosome. When the correct codon–anticodon interaction occurs, EF-Tu interacts with the factor-binding center, hydrolyzes its bound GTP, and is released from the tRNA and the ribosome. After EF-Tu release, the tRNA rotates into the peptidyl transferase center of the ribosome (called accommodation).

translation process. At least three different mechanisms contribute to this specificity (see Fig. 15-31). In each case, these mechanisms select *against* incorrect codon–anticodon pairings.

One mechanism that contributes to the fidelity of codon recognition involves two adjacent adenine residues in the 16S rRNA component located within the A-site of the small subunit. These bases form hydrogen bonds with the minor groove of each correct base pair formed between the anticodon and the first two bases of the codon in the A-site (Fig. 15-31a). Recall (see Chapter 6, Fig. 6-14) that the hydrogen-bonding properties of a Watson–Crick G:C and A:U base pair are very similar in the minor groove. Thus, the adjacent A residues in the 16S rRNA do not discriminate between G:C or A:U base pairs and recognize either as correct. In contrast, non-Watson–Crick base pairs or mismatched bases form a minor groove that cannot be recognized by these bases, resulting in significantly reduced affinity for incorrect tRNAs. The net result of these interactions is that correctly paired tRNAs show a much lower rate of dissociation from the ribosome than do incorrectly paired tRNAs.

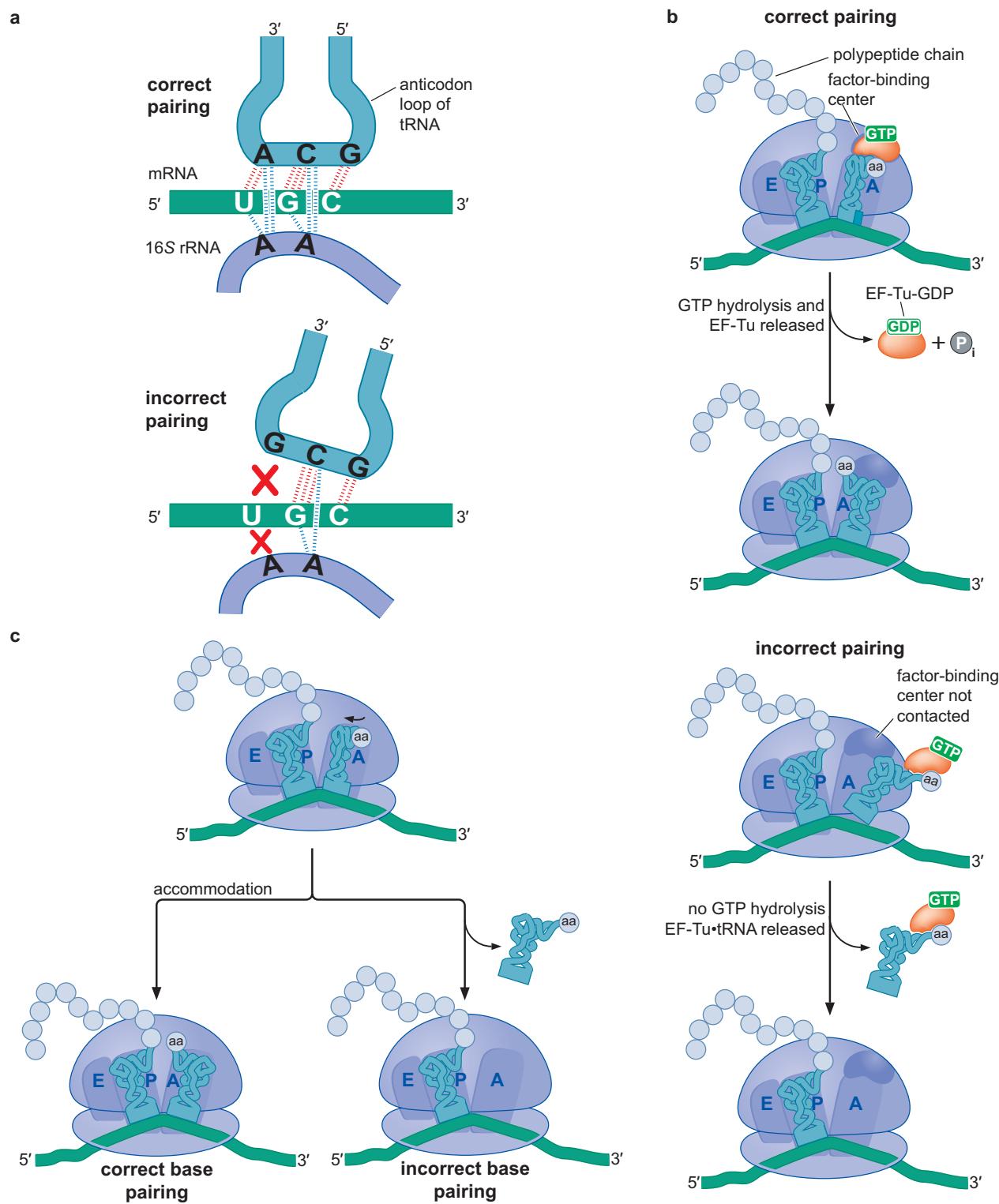
A second mechanism that helps to ensure correct codon–anticodon pairing involves the GTPase activity of EF-Tu (see Fig. 15-31b). As described above, release of EF-Tu from the tRNA requires GTP hydrolysis, which is highly sensitive to correct codon–anticodon base pairing. Even a single mismatch in the codon–anticodon base pairing alters the position of EF-Tu, reducing its ability to interact with the factor-binding center. This, in turn, leads to a dramatic reduction in EF-Tu GTPase activity. This mechanism is an example of kinetic selectivity and is related to the mechanisms used to ensure correct base pairing during DNA synthesis (see Chapter 9). In both cases, formation of correct base-pairing interactions dramatically enhances the rate of a critical biochemical step. For the DNA polymerase, this step was the formation of the phosphodiester bond. In this case, it is the hydrolysis of GTP by EF-Tu.

A third mechanism that ensures pairing accuracy is a form of proofreading that occurs after EF-Tu is released. When the charged tRNA is first introduced into the A-site in a complex with EF-Tu-GTP, its 3' end is distant from the site of peptide-bond formation. To participate successfully in the peptidyl transferase reaction, the tRNA must rotate into the peptidyl transferase center of the large subunit in a process called **accommodation** (Fig. 15-31c). During accommodation, the 3' end of the aminoacylated tRNA moves almost 70 Å. Incorrectly paired tRNAs frequently dissociate from the ribosome during accommodation. It is hypothesized that the rotation of the tRNA places a strain on the codon–anticodon interaction and that only a correctly paired anticodon can sustain this strain. Thus, mispaired tRNAs are more likely to dissociate from the ribosome before participating in the peptidyl transferase reaction.

In summary, in addition to the codon–anticodon interactions, the ribosome exploits minor groove interactions and two phases of proofreading to ensure that a correct aminoacyl-tRNA binds in the A-site. Each of these three additional selectivity mechanisms inhibits retention of aminoacyl-tRNAs that do not form correct codon–anticodon interactions.

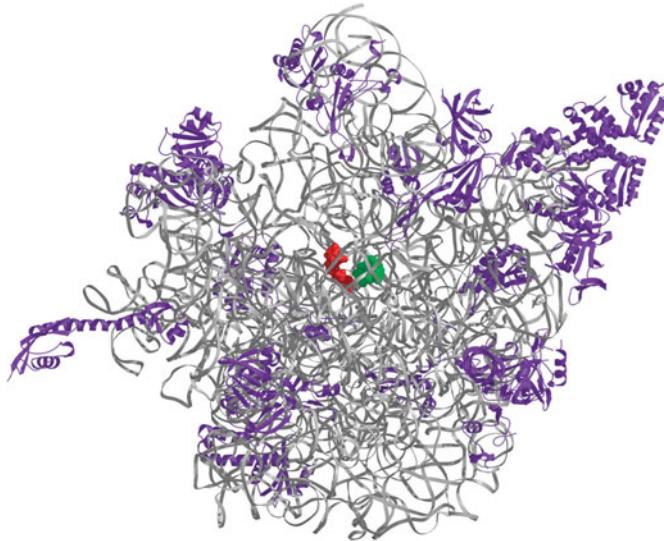
### The Ribosome Is a Ribozyme

Once the correctly charged tRNA has been placed in the A-site and has rotated into the peptidyl transferase center, peptide-bond formation takes place. This reaction is catalyzed by RNA, specifically the 23S rRNA component of the large subunit. Early evidence for this came from experiments in which it was shown that a large subunit that had been largely stripped of its proteins was still able to direct peptide bond formation. In support of



**FIGURE 15-31** Three mechanisms to ensure correct pairing between the tRNA and the mRNA. (a) Additional hydrogen bonds are formed between two adenine residues of the 16S rRNA and the minor groove of the anticodon–codon pair only when the first two bases of the anticodon–codon pair form correct Watson–Crick base pairs. (b) Correct codon–anticodon base pairing facilitates EF-Tu bound to the aminoacyl-tRNA to interact with the factor-binding center inducing GTP hydrolysis and EF-Tu release. (c) Only correctly base-paired aminoacyl-tRNAs remain associated with the ribosome as they rotate into the correct position for peptide-bond formation. This rotation is referred to as tRNA accommodation.

**FIGURE 15-32** RNA surrounds the peptidyl transferase center of the large ribosomal subunit. The 3D structure of the bacterial 50S subunit shows the RNAs (gray) and the ribosomal proteins (purple). The 3' ends of the A-site and P-site tRNAs that are immediately adjacent to the peptidyl transferase center are shown in green and red, respectively. (Yusupov M.M. et al. 2001. *Science* **292**: 883–896.) Image prepared with MolScript, BobScript, and Raster3D.

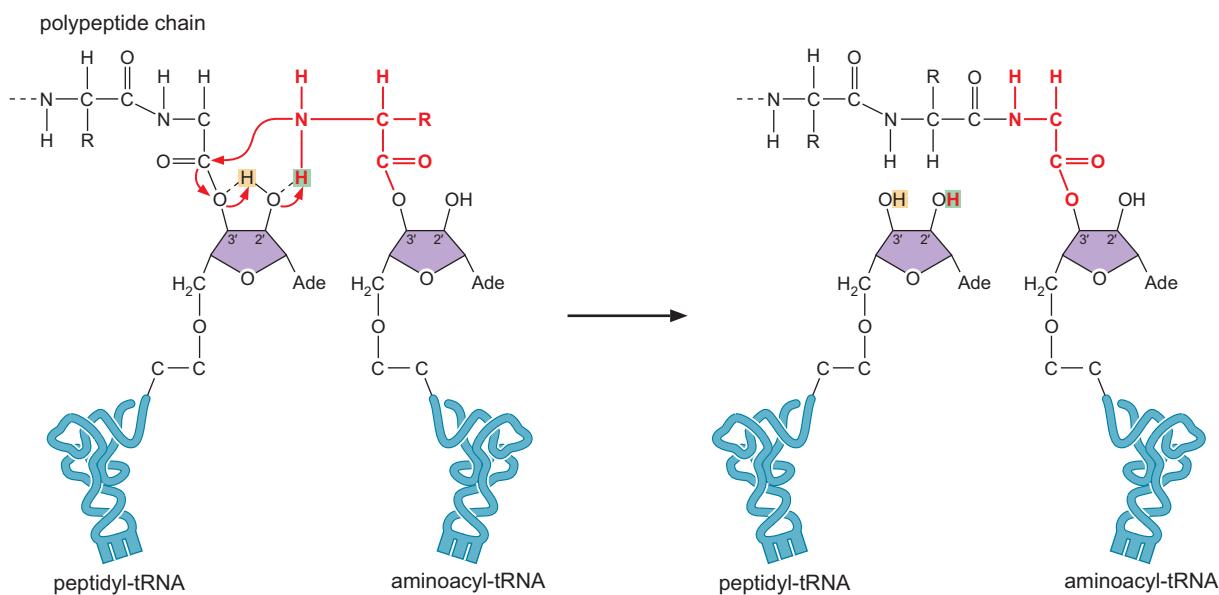


In this view, structural studies of the large subunit of the ribosome from one prokaryotic species showed that there was no amino acid within 18 Å of the active site (Fig. 15-32).

The 3D structure of the complete *Escherichia coli* ribosome with bound mRNA and tRNAs revealed that the very amino terminus of one protein (L27) does reach into the active site. This finding suggested a role for this protein in catalysis. To test this possibility, the nine amino acids at the L27 amino terminus that were in close proximity to the active site were eliminated by mutation. The resulting cells produced ribosomes with reduced but detectable peptidyl transferase activity, clearly indicating that this region of the L27 protein contributes to peptidyl transferase activity. The mutant ribosomes, however, still synthesized proteins at 30%–50% of wild-type levels and cells containing them continued to grow and divide. The ribosome promotes a 10<sup>7</sup>-fold increase in the rate of peptide-bond formation relative to the rate observed with substrates (aminoacyl-tRNAs) alone in solution. Clearly, the vast majority of this increase is retained, even without the presence of L27 in the active site. Thus, although this protein facilitates peptide-bond formation, it is not essential for peptide transferase activity. Like other ribosomal proteins, the most likely role for L27 is to correctly position one or more of the RNA components of the active site. More importantly, because this protein is the only one close enough to act catalytically, the rRNA component of the ribosome must be primarily responsible for catalyzing peptide-bond formation.

How then does the 23S rRNA catalyze peptide-bond formation? The exact mechanism remains to be determined, but some answers to this question are beginning to emerge. First, base pairing between the 23S rRNA and the CCA ends of the tRNAs in the A- and P-sites positions the α-amino group of the aminoacyl-tRNA to attack the carbonyl group of the growing polypeptide attached to the peptidyl-tRNA. These interactions are also likely to stabilize the aminoacyl-tRNA after accommodation. This type of catalytic mechanism is called entropic catalysis. That is, the enzyme works by bringing the substrates together in a manner that stimulates catalysis.

Because close proximity of substrates is rarely sufficient to generate high levels of catalysis, it is likely that other elements of the rRNA contribute to catalysis. Indeed, alterations that eliminate the 2'-OH of a highly conserved residue in the 23S rRNA (A2451 in the *E. coli* 23S rRNA) reduce the rates of catalysis by at least 10-fold. Recent studies have implicated a second



**FIGURE 15-33** Proposed role for the 2'-OH of the P-site tRNA in peptide-bond formation.

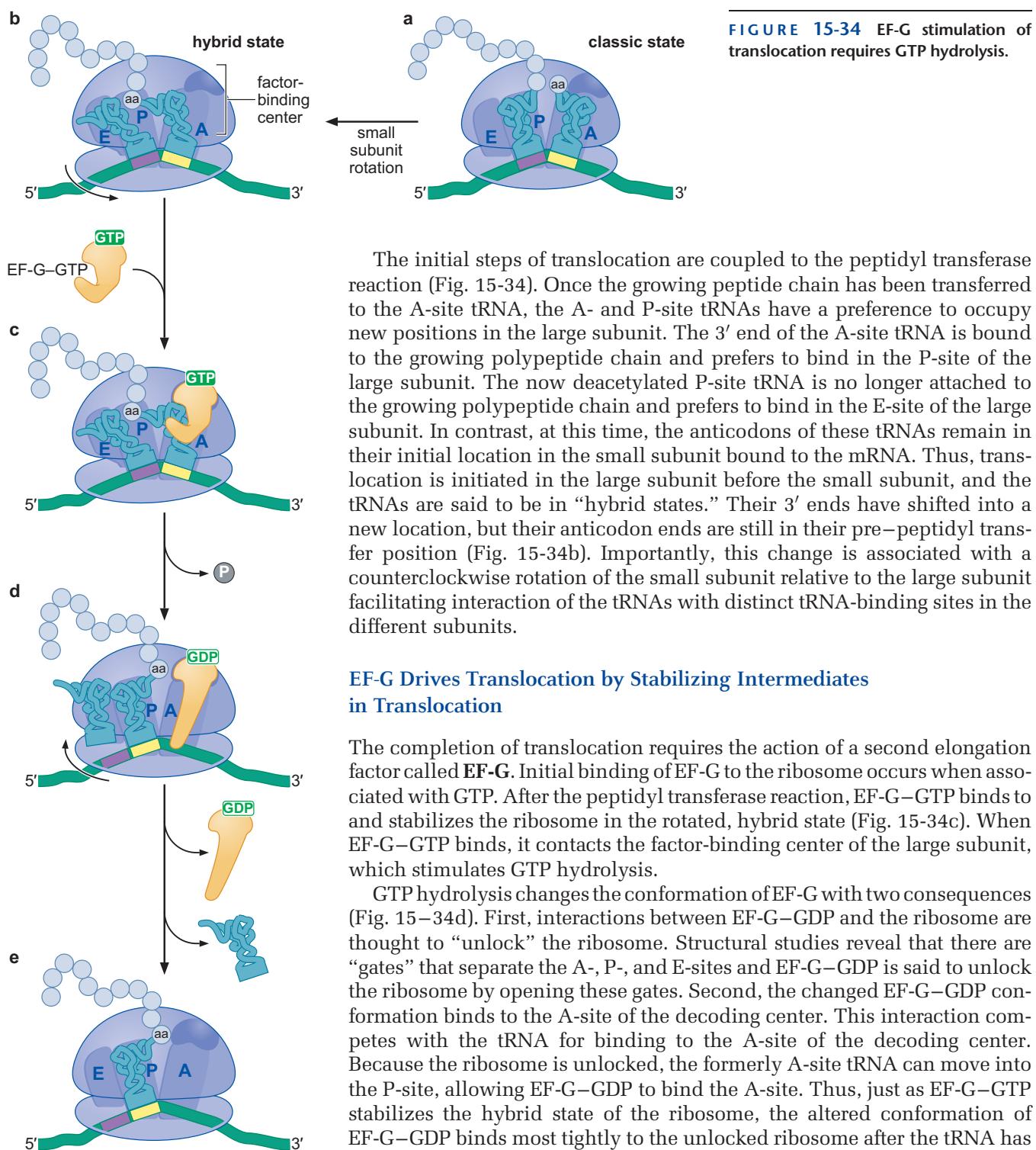
The 2'-OH of the final “A” in the peptidyl-tRNA is critical for peptide-bond formation. Based on this finding, it has been proposed that the hydrogen component of the 2'-OH participates in a “proton shuttle” illustrated here. In this model, as the bond between the peptidyl-tRNA and the polypeptide chain is broken, the 3' oxygen extracts a hydrogen (yellow highlight) from the 2'-hydroxyl, and the 2' oxygen, in turn, extracts a hydrogen (green highlight) from the amino group attacking the carbonyl. The red arrows show the proposed direction of electron movement during peptide-bond formation.

unexpected RNA as being critical for catalysis: the P-site tRNA. Mutations that remove the 2'-OH of the A residue at the 3' end of the P-site tRNA result in a  $10^6$ -fold reduction in catalysis rates. This “substrate-assisted catalysis” is a particularly interesting finding because it indicates that the peptidyl-tRNAs themselves carry critical catalytic elements. This finding suggests that, before the evolution of the ribosome, tRNAs may have provided critical elements to allow them to catalyze protein synthesis on their own.

On the basis of a number of considerations, it has been proposed that the 2'-OH of the P-site tRNA may act as part of a “proton shuttle” (Fig. 15-33). In this model, the 2'-OH donates a hydrogen to the 3'-OH of the peptidyl-tRNA and accepts a proton from the attacking  $\alpha$ -amino group of the amino acid attached to the A-site tRNA. Importantly, both of these findings strongly support the hypothesis that it is RNA and not protein that catalyzes peptide-bond formation. Nevertheless, there is still much to be learned regarding how the ribosome catalyzes peptide-bond formation.

### Peptide-Bond Formation Initiates Translocation in the Large Subunit

Once the peptidyl transferase reaction has occurred, the tRNA in the P-site is deacylated (no longer attached to an amino acid), and the growing polypeptide chain is linked to the tRNA in the A-site. For a new round of peptide chain elongation to occur, the P-site tRNA must move to the E-site and the A-site tRNA must move to the P-site. At the same time, the mRNA must move by three nucleotides to expose the next codon. These movements are coordinated within the ribosome and are collectively referred to as **translocation**.



The initial steps of translocation are coupled to the peptidyl transferase reaction (Fig. 15-34). Once the growing peptide chain has been transferred to the A-site tRNA, the A- and P-site tRNAs have a preference to occupy new positions in the large subunit. The 3' end of the A-site tRNA is bound to the growing polypeptide chain and prefers to bind in the P-site of the large subunit. The now deacetylated P-site tRNA is no longer attached to the growing polypeptide chain and prefers to bind in the E-site of the large subunit. In contrast, at this time, the anticodons of these tRNAs remain in their initial location in the small subunit bound to the mRNA. Thus, translocation is initiated in the large subunit before the small subunit, and the tRNAs are said to be in “hybrid states.” Their 3' ends have shifted into a new location, but their anticodon ends are still in their pre-peptidyl transfer position (Fig. 15-34b). Importantly, this change is associated with a counterclockwise rotation of the small subunit relative to the large subunit facilitating interaction of the tRNAs with distinct tRNA-binding sites in the different subunits.

#### EF-G Drives Translocation by Stabilizing Intermediates in Translocation

The completion of translocation requires the action of a second elongation factor called **EF-G**. Initial binding of EF-G to the ribosome occurs when associated with GTP. After the peptidyl transferase reaction, EF-G–GTP binds to and stabilizes the ribosome in the rotated, hybrid state (Fig. 15-34c). When EF-G–GTP binds, it contacts the factor-binding center of the large subunit, which stimulates GTP hydrolysis.

GTP hydrolysis changes the conformation of EF-G with two consequences (Fig. 15-34d). First, interactions between EF-G–GDP and the ribosome are thought to “unlock” the ribosome. Structural studies reveal that there are “gates” that separate the A-, P-, and E-sites and EF-G–GDP is said to unlock the ribosome by opening these gates. Second, the changed EF-G–GDP conformation binds to the A-site of the decoding center. This interaction competes with the tRNA for binding to the A-site of the decoding center. Because the ribosome is unlocked, the formerly A-site tRNA can move into the P-site, allowing EF-G–GDP to bind the A-site. Thus, just as EF-G–GTP stabilizes the hybrid state of the ribosome, the altered conformation of EF-G–GDP binds most tightly to the unlocked ribosome after the tRNA has left the A-site. Like dominoes, movement of the A-site tRNA into the P-site forces the P-site tRNA into the E-site. Base pairing between the tRNAs and the mRNA causes the mRNA to move by 3 bp. That this distance is dictated by the tRNA is shown by rare “frameshifting” tRNAs that have four-nucleotide-long anticodons (and can therefore compensate for certain frame-shift mutations) and move the mRNA by four nucleotides instead of three.

Completion of translocation is accompanied by a clockwise rotation of the small subunit back to its starting position. The resulting ribosome structure has dramatically reduced affinity for EF-G–GDP. Release of EF-G

results in the return of the ribosome to a “locked” state in which the tRNAs and mRNA are once again tightly associated with the small subunit decoding center and the gates between the A-, P- and E-sites are closed. Together, these events result in the translocation of the A-site tRNA into the P-site, the P-site tRNA into the E-site, and the movement of the mRNA by exactly 3 bp (Fig. 15-34e). The ribosome is now ready for a new cycle of amino acid addition to begin.

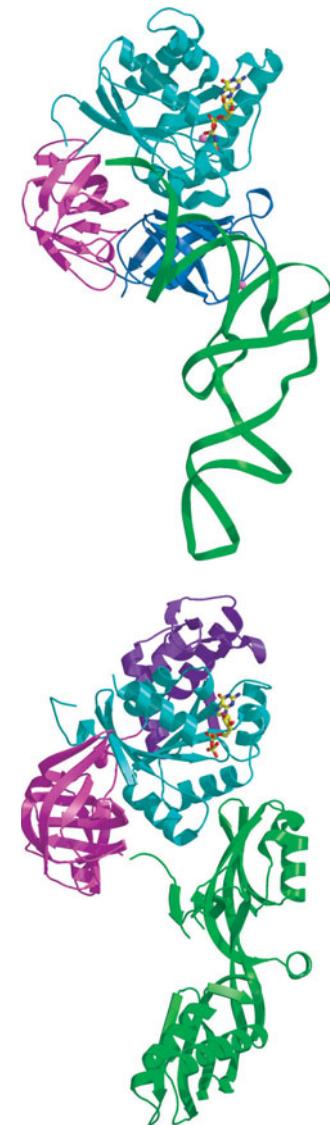
How does EF-G–GDP interact with the A-site of the decoding center so effectively? Crystal structures of EF-G and EF-Tu bound to tRNA reveal a clear answer to this question. EF-G–GDP and EF-Tu–GTP–tRNA have a very similar structure (Fig. 15-35). Recall that EF-Tu–GTP–tRNA also binds to the A-site decoding center. What is most remarkable about this similarity is that even though EF-G is composed of a single polypeptide, its structure mimics that of a tRNA bound to a protein. This is an example of “molecular mimicry” in which a protein takes on the appearance of a tRNA to facilitate association with the same binding site. Intriguingly, structural studies of the eukaryotic analog of EF-G (called eEF-2) have identified two dramatically different conformations of the protein (one of which is bound to the antibiotic sordarin). One conformation is similar to the structure of the EF-G shown in Figure 15-35, whereas the second conformation results from a dramatic movement of the tRNA mimic region relative to the GTP-binding region. The ability to alternate between such distinct conformations is critical for the function of EF-G during translocation.

### EF-Tu–GDP and EF-G–GDP Must Exchange GDP for GTP before Participating in a New Round of Elongation

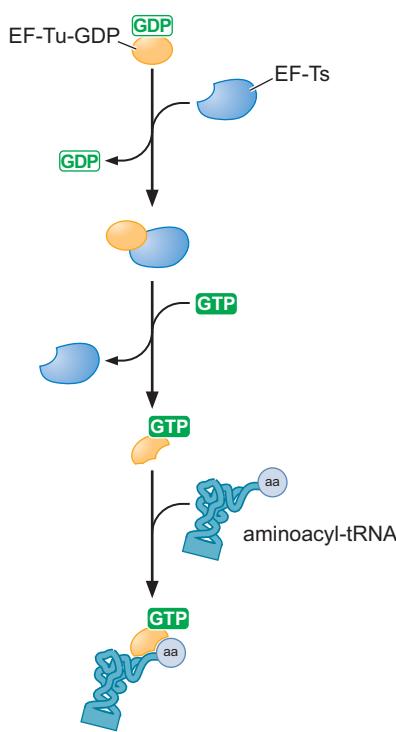
EF-Tu and EF-G are catalytic proteins that are used once for each round of tRNA loading onto the ribosome, peptide-bond formation, and translocation. After GTP hydrolysis, both proteins must release their bound GDP and bind a new molecule of GTP. For EF-G, this is a simple process, because GDP has a lower affinity for EF-G than does GTP. Thus, after GTP hydrolysis, GDP and phosphate are released, and the unbound EF-G rapidly binds a new GTP molecule. In the case of EF-Tu, a second protein is required to exchange GDP for GTP. The elongation factor EF-Ts acts as a **GTP exchange factor** for EF-Tu (Fig. 15-36). After EF-Tu–GDP is released from the ribosome, a molecule of EF-Ts binds to EF-Tu, causing the displacement of GDP. Next, GTP binds to the resulting EF-Tu–EF-Ts complex, causing its dissociation into free EF-Ts and EF-Tu–GTP. Finally, EF-Tu–GTP binds a molecule of charged tRNA, regenerating the EF-Tu–GTP–aminoacyl-tRNA complex, which is once again ready to deliver a charged tRNA to the ribosome.

### A Cycle of Peptide-Bond Formation Consumes Two Molecules of GTP and One Molecule of ATP

We conclude our discussion of elongation by accounting for the energy spent. How many molecules of nucleoside triphosphate does it cost per round of peptide-bond formation (setting aside the energetics of amino acid biosynthesis and the energetics of initiation and termination)? Recall that one molecule of nucleoside triphosphate (ATP) is consumed by the aminoacyl-tRNA synthetase in creating the high-energy acyl bond that links the amino acid to the tRNA. The breakage of this high-energy bond drives the peptidyl transferase reaction that creates the peptide bond. A second molecule of nucleoside triphosphate (GTP) is consumed in the delivery of a charged tRNA to the A-site of the ribosome by EF-Tu and in ensuring that



**FIGURE 15-35** Structural comparison of elongation factors. EF-Tu–GDPNP–Phe–tRNA is shown on the top and EF-G–GDP is shown on the bottom. GDPNP is an analog of GTP that cannot be hydrolyzed that is used to lock the molecule in the GTP-bound conformation during the determination of the 3D structure. Note the similarity between the structure of the green domain in EF-G and the tRNA bound to EF-Tu (also shown in green). (Top structure: Nissen P. et al. 1995. *Science* **270**: 1464–1472. Bottom structure: al-Karadaghi S. et al. 1996. *Structure* **4**: 555–565.) Images prepared with MolScript, BobScript, and Raster3D.



**FIGURE 15-36** EF-Ts stimulates release of GDP from EF-Tu. GDP bound to EF-Tu is released very slowly in isolation. EF-Ts binds EF-Tu–GDP and causes the rapid release of GDP. GTP binding to EF-Tu in the EF-Tu–EF-Ts complex displaces EF-Ts and leaves EF-Tu–GTP, which can then bind a new aminoacyl-tRNA for delivery to the ribosome.

correct codon–anticodon recognition had taken place. Finally, a third nucleoside triphosphate is consumed in the EF-G-mediated process of translocation. Thus, making a peptide bond costs the cell two molecules of GTP and one of ATP, with one nucleoside triphosphate being consumed for each step in the translation elongation process. Interestingly, of the three molecules, only one (ATP) is energetically connected to peptide-bond formation. The energy of the other two molecules (GTP) is spent to ensure the accuracy and order of events during translation (see Box 15-4, GTP-binding Proteins, Conformational Switching, and the Fidelity and Ordering of the Events of Translation).

Throughout the discussion of translation elongation, we have not distinguished between prokaryotes and eukaryotes. Although the eukaryotic factors analogous to EF-Tu (eEF1) and EF-G (eEF2) are named differently, their functions are remarkably similar to their prokaryotic counterparts.

## TERMINATION OF TRANSLATION

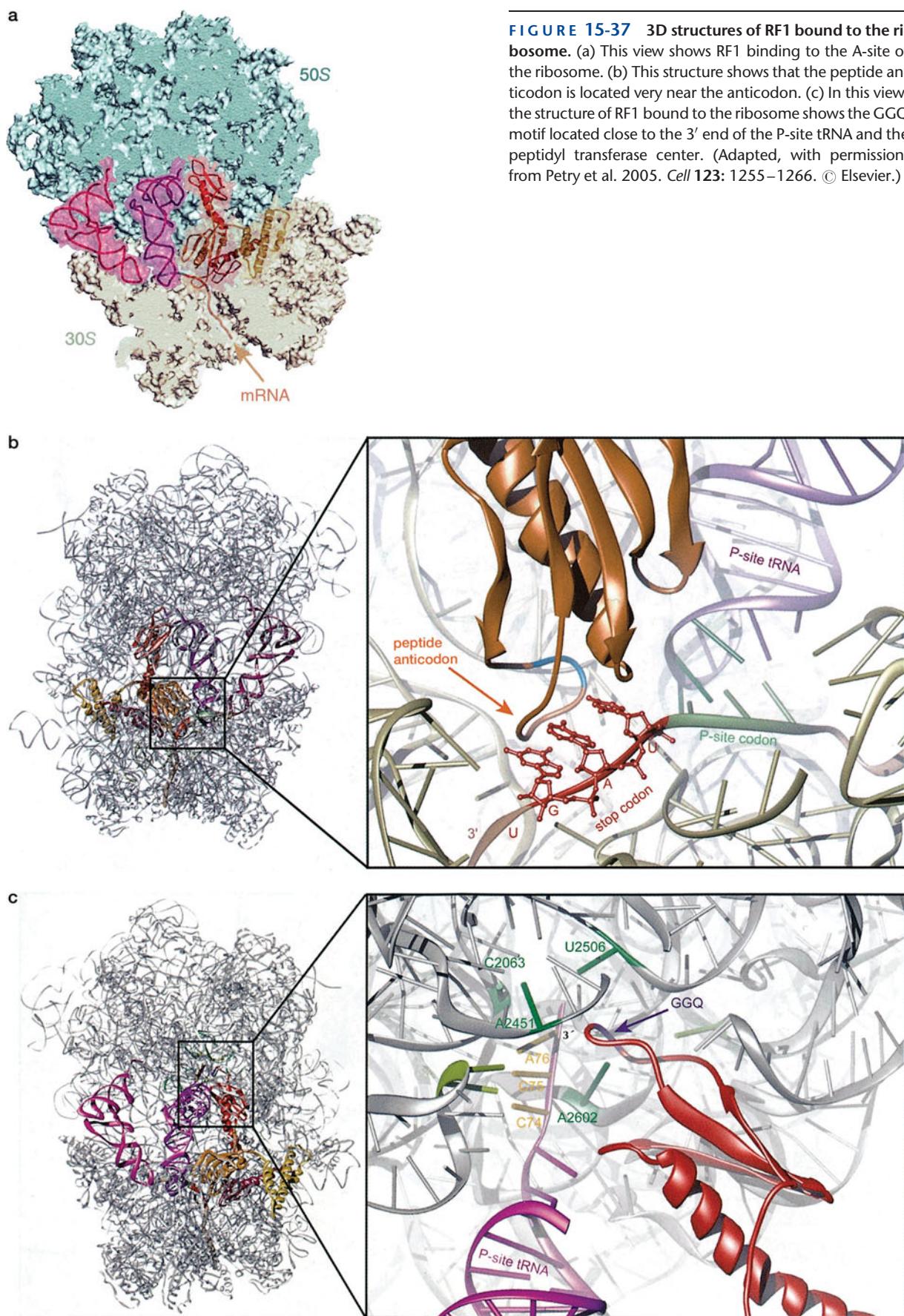
### Release Factors Terminate Translation in Response to Stop Codons

The ribosome's cycle of aminoacyl-tRNA binding, peptide-bond formation, and translocation continues until one of the three stop codons enters the A-site. It was initially postulated that there would be one or more chain-terminating tRNAs that would recognize these codons. However, this is not the case. Instead, stop codons are recognized by proteins called **release factors (RFs)** that activate the hydrolysis of the polypeptide from the peptidyl-tRNA.

There are two classes of release factors. Class I release factors recognize the stop codons and trigger hydrolysis of the peptide chain from the tRNA in the P-site. Prokaryotes have two class I release factors called RF1 and RF2. RF1 recognizes the stop codon UAG, and RF2 recognizes the stop codon UGA. The third stop codon, UAA, is recognized by both RF1 and RF2. In eukaryotic cells, there is a single class I release factor called eRF1 that recognizes all three stop codons. Class II release factors stimulate the dissociation of the class I factors from the ribosome after release of the polypeptide chain. Prokaryotes and eukaryotes have only one class II factor called RF3 and eRF3, respectively. Like EF-G, IF2, and EF-Tu, class II release factors are regulated by GTP binding and hydrolysis.

### Short Regions of Class I Release Factors Recognize Stop Codons and Trigger Release of the Peptidyl Chain

How do release factors recognize stop codons? Because release factors are composed entirely of protein, protein–RNA interaction must mediate stop codon recognition. Experiments in which short coding regions were genetically swapped between RF1 and RF2 (which have different stop-codon specificity) identified a three-amino-acid sequence that is critical for release factor specificity. Exchange of these three amino acids between RF1 and RF2 swaps the stop-codon specificity of the two complexes. For this reason, this three-amino-acid sequence is called a *peptide anticodon* and must interact with and recognize stop codons. A 3D structure of RF1 bound to the ribosome confirms that RF1 binds to the A-site of the ribosome (Fig. 15-37a). In this structure, the peptide anticodon is located very near the anticodon, but it is likely that there are additional protein regions that contribute to codon recognition (Fig. 15-37b).



**FIGURE 15-37** 3D structures of RF1 bound to the ribosome. (a) This view shows RF1 binding to the A-site of the ribosome. (b) This structure shows that the peptide anticodon is located very near the anticodon. (c) In this view, the structure of RF1 bound to the ribosome shows the GGQ motif located close to the 3' end of the P-site tRNA and the peptidyl transferase center. (Adapted, with permission, from Petry et al. 2005. *Cell* 123: 1255–1266. © Elsevier.)

## ► ADVANCED CONCEPTS

**Box 15-4 GTP-Binding Proteins, Conformational Switching, and the Fidelity and Ordering of the Events of Translation**

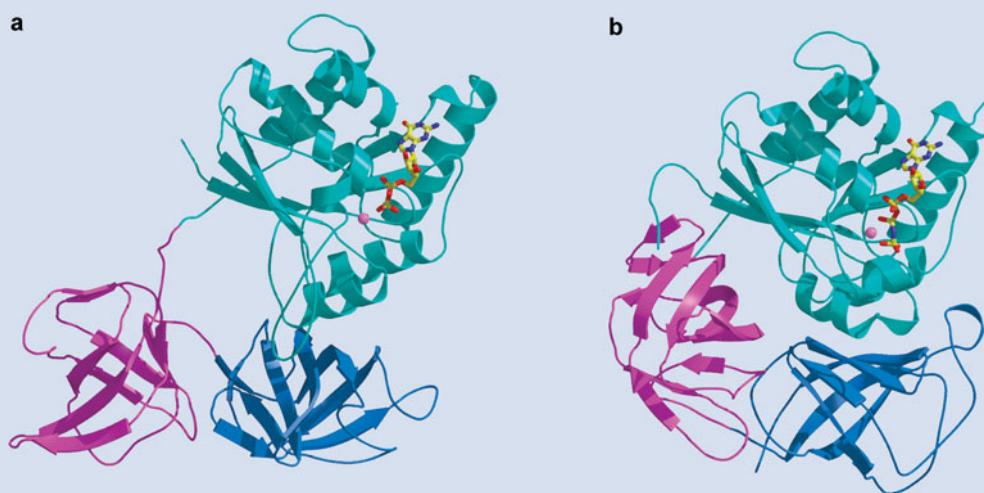
GTP is used throughout translation to control key events. The energy of GTP hydrolysis is not coupled to chemical modification as ATP is in the coupling of amino acids to tRNAs. Instead, the energy of GTP hydrolysis is used to control the order and fidelity of events during translation. How is this accomplished?

A key feature of the GTP-binding proteins involved in translation is that their conformation changes depending on the guanine nucleotide (GDP vs. GTP) to which they are bound. This can be seen for EF-Tu in Box 15-4 Figure 1, which shows the 3D structure of EF-Tu bound to GTP or GDP. EF-Tu undergoes a major conformational change when it binds to GTP that results in the formation of its tRNA-binding site. In particular, one domain of EF-Tu (shown in magenta in Box 15-4 Fig. 1) shifts its location relative to the other domains of the protein depending on the nucleotide that is bound. This change in domain location, as well as changes in the conformation of the other two domains (shown in turquoise and dark blue), results in the formation of a new surface on EF-Tu that binds tightly to charged tRNAs (see EF-Tu bound to a tRNA in Fig. 15-35). Thus, depending on the form of guanine nucleotide bound, these factors can have different functions or bind to different proteins/RNAs. For example, EF-Tu–GTP can bind to an aminoacyl-tRNA, but EF-Tu–GDP cannot.

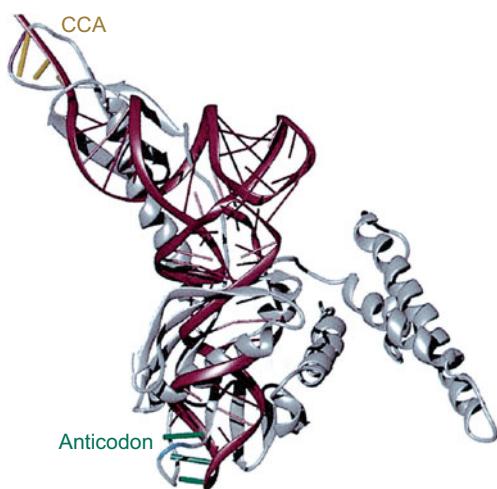
By coupling GTP hydrolysis to the completion of key events in translation, the order of these events can be tightly controlled. For EF-Tu, the GTP-dependent association of EF-Tu with aminoacyl-tRNAs ensures that peptide-bond formation does not occur before correct codon–anticodon pairing. Formation of the correct base pairs triggers GTP hydrolysis. Once bound to GDP, EF-Tu is released from the aminoacyl-tRNA allowing peptide-bond formation to ensue.

The mechanism that activates GTP hydrolysis by each of the GTP-regulated auxiliary proteins is the same. In each case, GTPase activity is stimulated through an interaction with a specific region of the large subunit called the factor-binding center. This interaction is not of sufficient affinity to occur in isolation. Instead, each GTP-controlled translation factor must make several other critical interactions with the ribosome to stabilize the precise association with the factor-binding center that leads to GTPase activation. Indeed, as we have seen for EF-Tu, this interaction is highly sensitive to the exact nature of the interactions between EF-Tu, the aminoacyl-tRNA, the mRNA, and the ribosome. Thus, the interaction with the factor-binding center monitors all of the other interactions of these proteins and RNAs with the ribosome. Only when correct codon–anticodon pairing is achieved does the GTP-binding site of EF-Tu interact productively with the factor-binding center, leading to GTP hydrolysis and the associated changes in protein conformation.

The use of GTP during translation is analogous to the use of ATP by the sliding clamp loaders (see Chapter 9, Box 9-3). Recall that in that case, ATP binding was required to assemble an initial complex with the sliding clamp, but ATP hydrolysis and release of the sliding clamp could only occur when the clamp loader bound the primer:template junction. In translation, GTP is required for the initial association of the GTP-regulated factors with the ribosome (and, in some instances, other RNAs and proteins), and GTP hydrolysis occurs only when the factor has correctly interacted with the ribosome. As in the case of the sliding clamp, GTP hydrolysis generally results in the release of the translation factor from the ribosome.



**BOX 15-4 FIGURE 1** Comparison of EF-Tu bound to GDP and GTP. (a) EF-Tu bound to GDP. (b) EF-Tu bound to GTP. The GTP-binding domain is turquoise. The rotation of the magenta domain and the changes in the structure of the turquoise and blue domains lead to the formation of a strong tRNA-binding site when GTP is bound (see Fig. 15-35). GTP is depicted in stick representation. (a, Polekhina G. et al. 1996. *Structure* 4: 1141–1151; b, Kjeldgaard M. et al. 1993. *Structure* 1: 35–50.) Images prepared with MolScript, BobScript, and Raster3D.



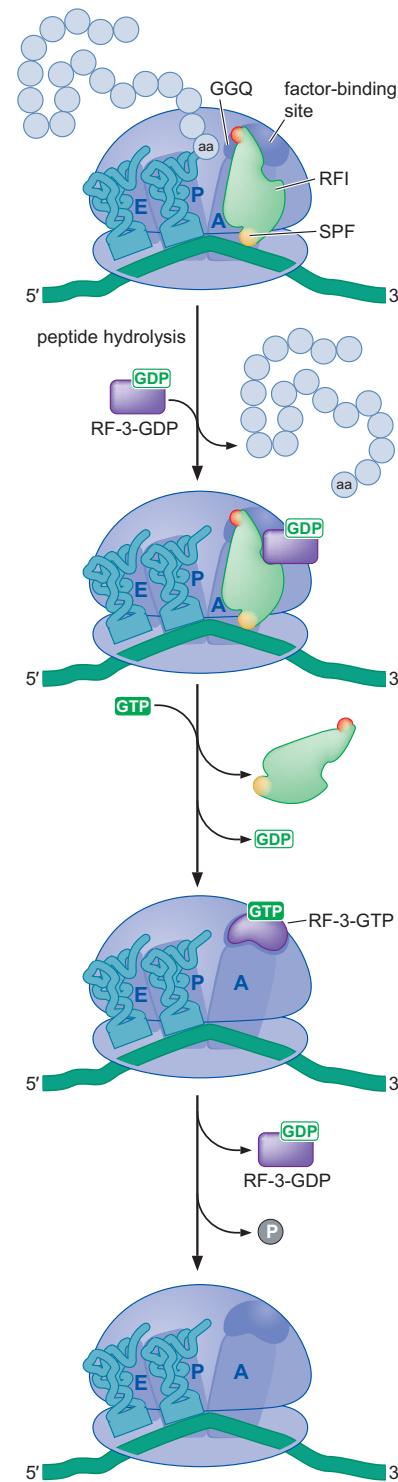
**FIGURE 15-38** Comparison of the structures of RF1 to a tRNA. The tRNA (dark red) and RF1 (gray) are shown occupying the same space. (Redrawn, with permission, from Petry et al. 2005. *Cell* 123: 1255–1266, Fig. 3E. © Elsevier.)

A region of class I release factors that stimulates polypeptide release has also been identified. All class I factors share a conserved three-amino-acid sequence (glycine glycine glutamine, GGQ) that is essential for polypeptide release. Moreover, the structure of RF1 bound to the ribosome confirms that the GGQ motif is located in close proximity to the peptidyl transferase center (Fig. 15-38c). It remains unclear whether the GGQ motif is directly involved in the hydrolysis of the polypeptide from the peptidyl-tRNA or if it induces a change in the peptidyl transferase center that allows the center itself to catalyze hydrolysis. Studies of the conserved bases found adjacent to the CCA ends in the peptidyl transferase center (e.g., A2541 or A2602) indicate that several of these residues are required for peptide hydrolysis. Indeed, these bases appear to play a more important role in peptide release than they do in peptide-bond formation. A likely explanation for this difference is that only proximal RNA residues can position a small water molecule for hydrolysis, but residues at many sites in the ribosomes can help position the larger tRNAs for catalysis.

Together, these studies have led to the hypothesis that class I release factors functionally mimic a tRNA, having a peptide anticodon that interacts with the stop codon and a GGQ motif that reaches into the peptidyl transferase center. Comparison of the structure of RF1 to a tRNA reveals how the protein functionally mimics a tRNA (Fig. 15-38). Just as the CCA 3' terminus and the anticodon loop occupy extreme ends of each tRNA, the GGQ and the peptide anticodon loop occupy extreme ends of RF1.

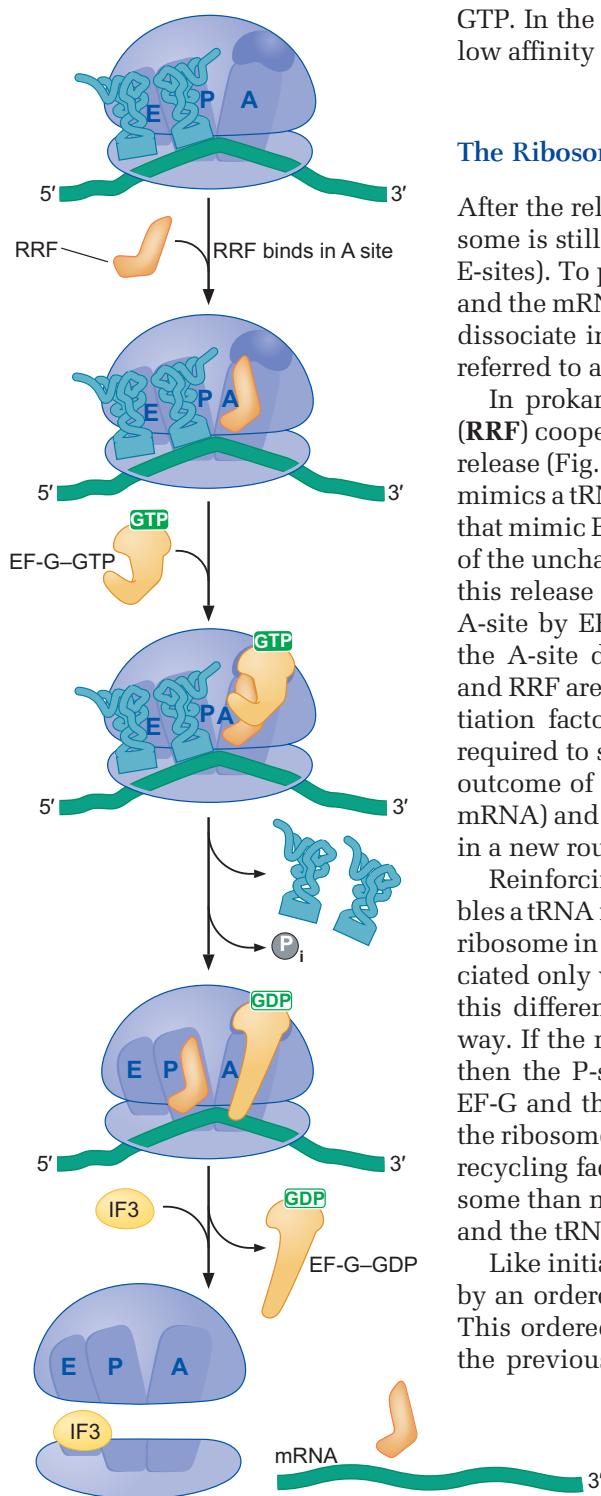
### GDP/GTP Exchange and GTP Hydrolysis Control the Function of the Class II Release Factor

Once the class I release factor has triggered the hydrolysis of the peptidyl-tRNA linkage, it must be removed from the ribosome (Fig. 15-39). This step is stimulated by the class II release factor, RF3. RF3 is a GTP-binding protein but, unlike the other GTP-binding proteins involved in translation, this factor has a higher affinity for GDP than GTP. Thus, free RF3 is predominantly in the GDP-bound form. RF3-GDP binds to the ribosome in a manner that depends on the presence of a class I release factor. After the class I release factor stimulates polypeptide release, a change in the conformation



**FIGURE 15-39** Polypeptide release is catalyzed by two release factors. The class I release factor (shown here as RF1) recognizes the stop codon and stimulates polypeptide release through a GGQ motif that is localized to the peptidyl transferase center. The class II release factor (RF3) binds only after polypeptide release and drives the dissociation of the class I release factor.

of the ribosome and the class I release factor stimulates RF3 to exchange its bound GDP for a GTP. That is, these factors act as a GTP exchange factor for RF3 in much the same way that EF-Ts does for EF-Tu. The binding of GTP to RF3 leads to the formation of a high-affinity interaction with the ribosome that favors the rotated hybrid state we discussed as a translocation intermediate above. This change in conformation displaces the class I factor from the ribosome. These changes also allow RF3 to associate with the factor-binding center of the large subunit. As with other GTP-binding proteins involved in translation, this interaction stimulates the hydrolysis of GTP. In the absence of a bound class I factor, the resulting RF3·GDP has a low affinity for the ribosome and is released.



### The Ribosome Recycling Factor Mimics a tRNA

After the release of the polypeptide chain and the release factors, the ribosome is still bound to the mRNA and two deacylated tRNAs (in the P- and E-sites). To participate in a new round of polypeptide synthesis, the tRNAs and the mRNA must be removed from the ribosome, and the ribosome must dissociate into its large and small subunits. Collectively, these events are referred to as **ribosome recycling**.

In prokaryotic cells, a factor known as the **ribosome recycling factor (RRF)** cooperates with EF-G and IF3 to recycle ribosomes after polypeptide release (Fig. 15-40). RRF binds to the empty A-site of the ribosome, where it mimics a tRNA. RRF also recruits EF-G–GTP to the ribosome, and, in events that mimic EF-G function during elongation, the EF-G stimulates the release of the uncharged tRNAs bound in the P- and E-sites. Although exactly how this release occurs is unclear, it is thought that RRF is displaced from the A-site by EF-G in a manner similar to the displacement of a tRNA from the A-site during elongation. Once the tRNAs are removed, EF-G–GDP and RRF are released from the ribosome along with the mRNA. IF3 (the initiation factor) may also participate in the release of the mRNA and is required to separate the two ribosomal subunits from each other. The final outcome of these events is a small subunit bound to IF3 (but not tRNA or mRNA) and a free large subunit. The released ribosome can now participate in a new round of translation.

Reinforcing the view that the RRF is a mimic of tRNA, RRF, in fact, resembles a tRNA in its 3D structure. Despite this similarity, RRF interacts with the ribosome in a manner very different from that of a tRNA. RRF is closely associated only with the large subunit portion of the A-site. We can rationalize this difference between the recycling factor and tRNAs in the following way. If the ribosome recycling factor precisely mimicked an A-site tRNA, then the P-site tRNA would be moved into the E-site by EF-G. Instead, EF-G and the recycling factor lead to the release of the P-site tRNA from the ribosome directly from the P-site. It is likely that EF-G and the ribosome recycling factor cause a more dramatic change in the structure of the ribosome than normally occurs during translocation, allowing both the mRNA and the tRNAs to be released.

Like initiation and elongation, the termination of translation is mediated by an ordered series of interdependent factor binding and release events. This ordered nature of translation ensures that no one step occurs before the previous step is complete. For example, EF-Tu cannot escort a new

**FIGURE 15-40** RRF and EF-G combine to stimulate the release of tRNA and mRNA from a terminated ribosome.

tRNA into the A-site until EF-G completes translocation. Similarly, RF3 cannot bind to the ribosome unless a class I release factor has already recognized a stop codon. There is a weakness to this orderly approach to translation: if any step cannot be completed, then the entire process stops. It is just this Achilles' heel that antibiotics exploit when they target the translation process (see Box 15-5, Antibiotics Arrest Cell Division by Blocking Specific Steps in Translation).

Although there are class I and II release factors in eukaryotic cells, their structure and amino acid sequence are unrelated and only the class I factor functions analogously. Like RF1 and RF2, eRF1 recognizes all three stop codons and brings a GGQ motif into the peptidyl transferase center leading to polypeptide release. Unlike the prokaryotic RF3 that catalyzes RF1/RF2 release, eRF3 delivers eRF1 to the ribosome. eRF3·GTP binds to eRF1 away from the ribosome and, like EF-Tu and a charged tRNA, escorts eRF1 to the ribosome (Fig. 15-41). Also like EF-Tu, if eRF1 recognizes a stop codon, eRF3 · GTP binds the factor-binding center, stimulating GTP hydrolysis. eRF3 · GDP is rapidly released from the ribosome, and eRF1 moves into the peptidyl-transferase center in a manner thought to be analogous to tRNA accommodation. Interestingly, there is no evidence for a ribosome-recycling factor in eukaryotic cells, nor does eEF2 (the eukaryotic EF-G) participate in ribosome recycling. Instead, current models suggest that after stimulating peptide hydrolysis from the P-site tRNA, eRF1 (in conjunction with an ATPase called Rli1) also participates in ribosome recycling based on the similarity of eRF1 and eRF3 to two proteins shown to stimulate ribosome disassembly at stalled ribosomes called Dom34 and Hbs1 (see later discussion).

## REGULATION OF TRANSLATION

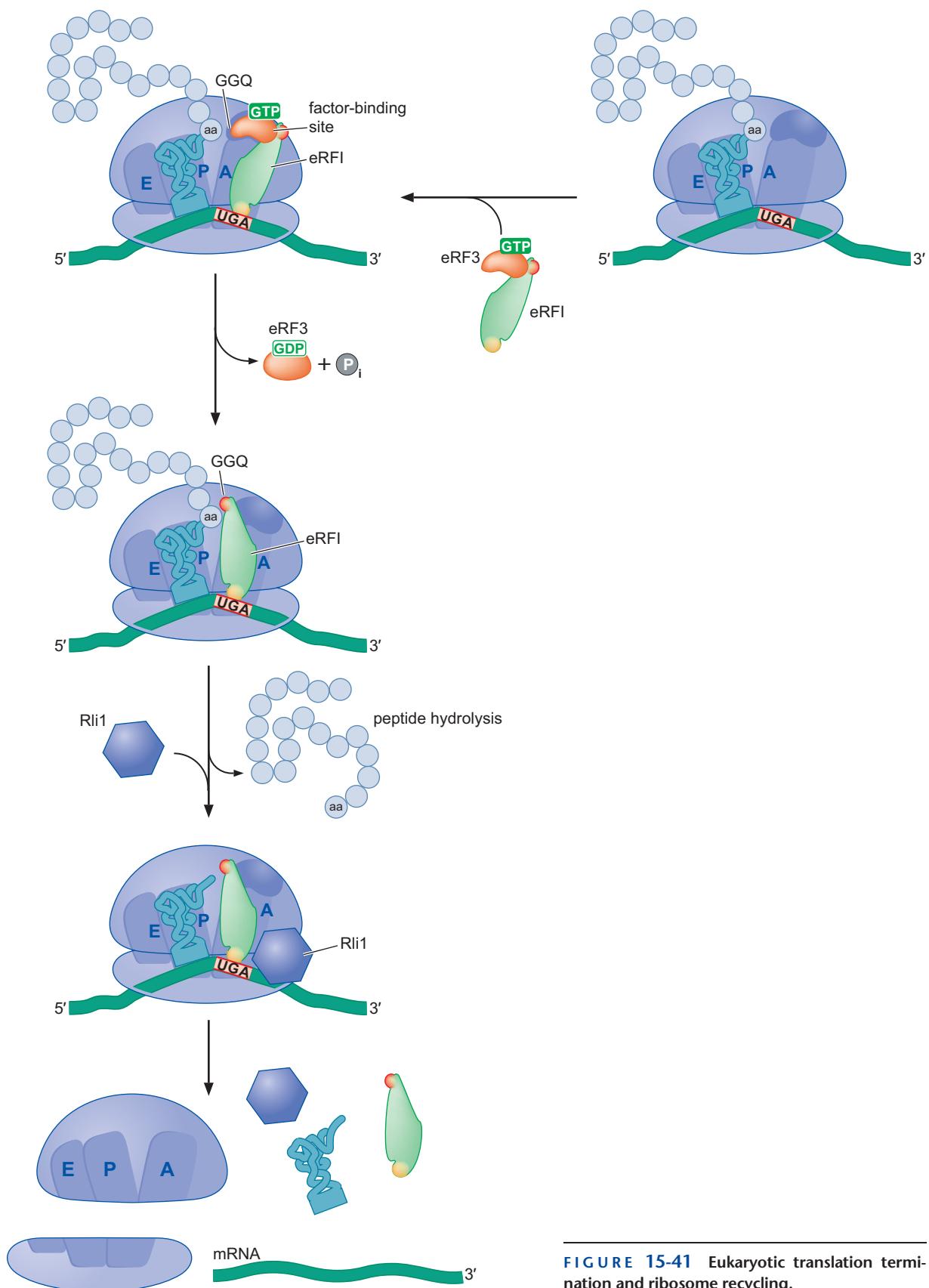
---

Although the expression of many genes is regulated at the level of mRNA transcription, it is becoming increasingly clear that many genes are also regulated at the level of protein synthesis. One advantage of control of translation over transcription is the ability to respond very rapidly to external stimuli. Regulation at the level of protein synthesis eliminates the time required to alter the levels of mRNA transcription (and in eukaryotes also mRNA processing and transport to the cytoplasm), thereby allowing a more rapid change in protein levels. As with other types of regulation, translational control typically functions at the level of initiation. It is generally more efficient to regulate a pathway at an earlier step rather than starting a process and then stopping it. In the case of translation, regulation at the level of initiation also eliminates the production of incomplete proteins that might have altered function.

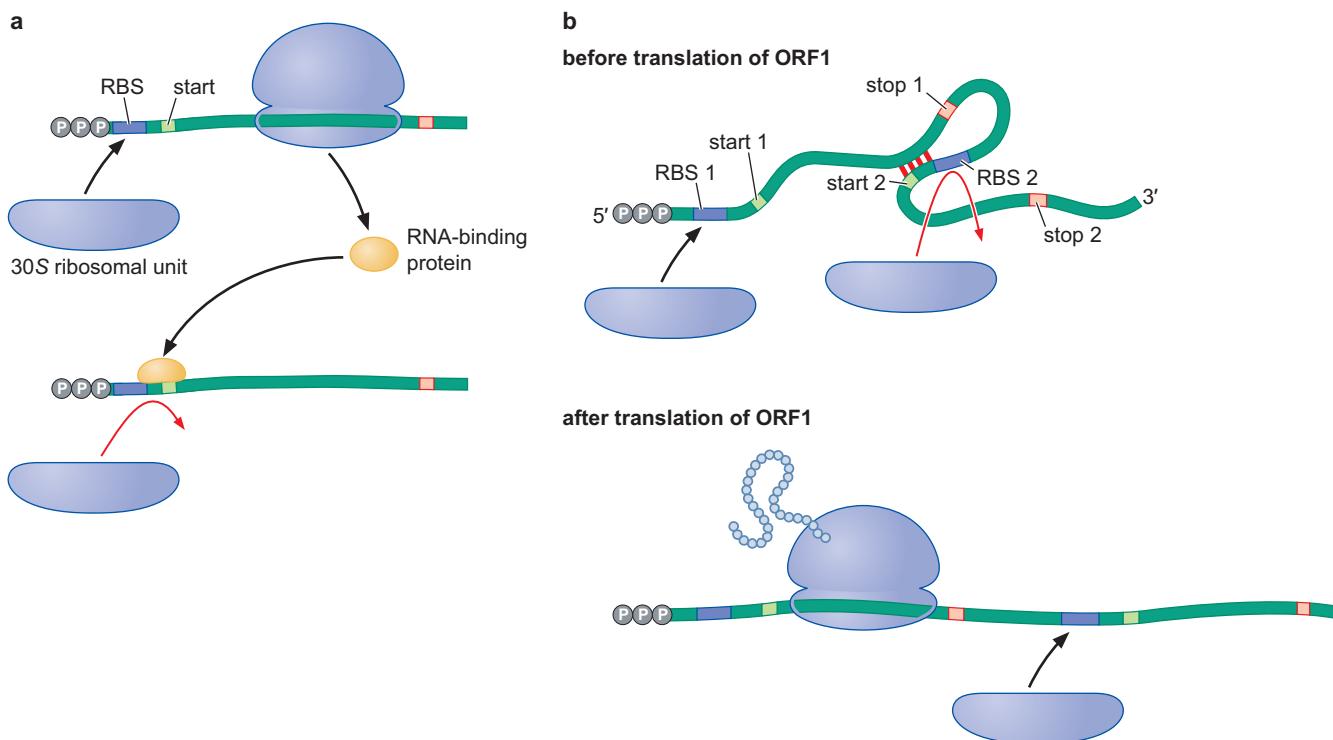
In this section, we first describe general mechanisms used by bacteria and eukaryotic cells to regulate translation. We then describe specific examples in which this type of regulation is used.

### Protein or RNA Binding near the Ribosome-Binding Site Negatively Regulates Bacterial Translation Initiation

The primary target of regulators of bacterial initiation is to interfere with the recognition of the RBS by the 30S subunit. In general, the mechanism of these inhibitors is to associate with sequences near the RBS and physically inhibit base pairing between the RBS and the 16S rRNA (Fig. 15-42a). These repressors are often RNA-binding proteins that recognize RNA structures



**FIGURE 15-41** Eukaryotic translation termination and ribosome recycling.



**FIGURE 15-42** Regulation of bacterial translation initiation by inhibiting 30S subunit binding. (a) Protein binding to sites near the RBS prevents access of the 16S rRNA to the RBS. In this case, the protein encoded by the mRNA binds to its own RBS. (b) Intramolecular base pairing of the mRNA can interfere with base pairing by the 16S rRNA. In many cases, this inhibition is modulated by the translation of other genes in the same operon. If the region of the mRNA that is base pairing to the RBS proximal region is within an ORF, when that ORF is translated, the interfering base pairing is disrupted, allowing a second ribosome to recognize the previously blocked RBS.

that form adjacent to the RBS. Although they do not bind directly to the RBS, the bound proteins are large enough to prevent the 30S subunit from gaining access to the RBS. Indeed, it is important that these repressors do not bind the RBS directly because such a protein would run the risk of inhibiting the translation of a large proportion of proteins in the cell.

RNA molecules can also act as inhibitors of translation using similar mechanisms. This regulation occurs most often when an mRNA base-pairs with itself to mask one or more RBSs (Fig. 15-42b). This masking can prevent translation of the associated ORF until the interaction is disrupted. In many instances, disruption occurs as a consequence of translating another gene in the operon. In this case, the region of the mRNA that is interacting with the RBS proximal region is within another ORF, and the passage of the ribosome disrupts the base pairing, thereby allowing another ribosome to recognize the unmasked RBS.

### Regulation of Prokaryotic Translation: Ribosomal Proteins Are Translational Repressors of Their Own Synthesis

We now present an example of regulation of translation in bacteria that illustrates how the cell uses these mechanisms to control correct expression of ribosomal protein genes. Coordinating the expression of ribosomal proteins

► MEDICAL CONNECTIONS

**Box 15-5 Antibiotics Arrest Cell Division by Blocking Specific Steps in Translation**

Antibiotics represent a powerful tool to fight disease. The most widely used antibiotics in medicine kill bacteria but have little or no effect on eukaryotic cells and hence are not toxic to the patient. Since their discovery in the first half of the last century, antibiotics have helped make previously untreatable infections such as tuberculosis, bacterial pneumonia, syphilis, and gonorrhea largely curable (although the emergence of antibiotic-resistant bacteria is becoming an increasing obstacle to effective treatment). Antibiotics have many different kinds of targets in the bacterial cell, but ~40% of the known antibiotics are inhibitors of the translation machinery (Box 15-5 Table 1). In general, these antibiotics bind a component of the translation apparatus and inhibit its function. Because different antibiotics arrest translation at different steps and do so in a precise manner (e.g., just before EF-Tu release), these agents have become useful tools in studies of the mechanism of protein synthesis. Thus, in addition to their obvious medical benefits, antibiotics have come to play an important role in helping us understand the workings of the translation machinery.

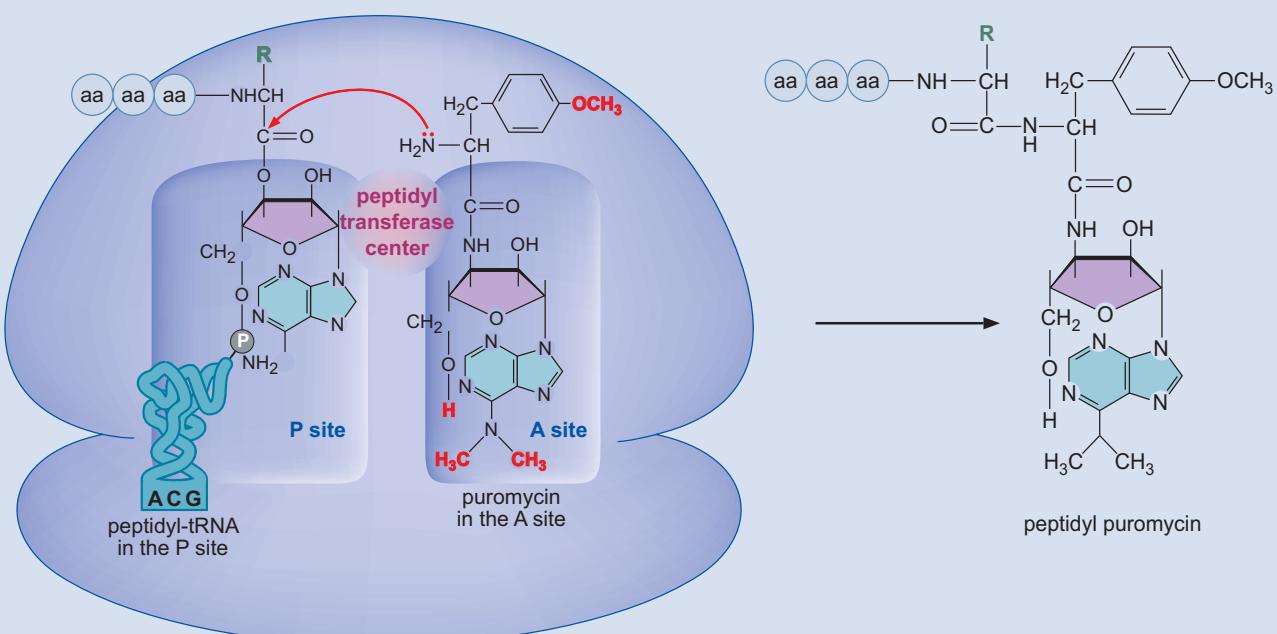
Puromycin is one antibiotic commonly used in studies of translation. It binds to the large subunit region of the A-site. Once bound, puromycin can substitute for an aminoacyl-tRNA

in the peptidyl transferase reaction (Box 15-5 Fig. 1). Because puromycin is very small compared with a tRNA, its binding to the A-site is not sufficient to retain the polypeptide chain on the ribosome. Thus, peptidyl chains that are transferred to puromycin dissociate from the ribosome as an incomplete, puromycin-bound polypeptide. In other words, puromycin causes polypeptide synthesis to terminate prematurely. Other antibiotics target other features of the ribosome, such as the peptide exit tunnel, the peptidyl transferase center, the factor-binding center, the decoding center, and regions critical for translocation (Box 15-5 Table 1).

Yet other antibiotics are inhibitors of translation factors. For example, kirromycin and fusidic acid are inhibitors of the elongation factors EF-Tu and EF-G, respectively. In both cases, the antibiotic interacts with the GTP-bound form of the translation factor and prevents changes in conformation that would normally occur after GTP hydrolysis. Thus, kirromycin arrests ribosomes with bound EF-Tu · GDP aminoacyl-tRNA. Similarly, fusidic acid arrests ribosomes with bound EF-G · GDP. In both cases, the next step in translation is prevented by the failure to release the elongation factor.

**BOX 15-5 TABLE 1** Antibiotics: Targets and Consequences

Antibiotic/Toxin	Target Cells	Molecular Target	Consequence
Tetracycline	Prokaryotic cells	A-site of 30S subunit	Inhibits aminoacyl-tRNA binding to A-site
Hygromycin B	Prokaryotic and eukaryotic cells	Near A-site of 30S subunit	Prevents translocation of A-site tRNA to P-site
Paromycin	Prokaryotic cells	Adjacent to A-site codon–anticodon interaction site in 30S subunit	Increases error rate during translation by decreasing selectivity of codon–anticodon pairing
Chloramphenicol	Prokaryotic cells	Peptidyl transferase center of 50S subunit	Blocks correct positioning of A-site aminoacyl-tRNA for peptidyl transfer reaction
Puromycin	Prokaryotic and eukaryotic cells	Peptidyl transferase center of large ribosomal subunit	Chain terminator; mimics 3' end of aminoacyl-tRNA in A-site and acts as acceptor for nascent polypeptide chain
Erythromycin	Prokaryotic cells	Peptide exit tunnel of 50S subunit	Blocks exit of growing polypeptide chain from the ribosome; arrests translation
Fusidic acid	Prokaryotic cells	EF-G	Prevents release of EF-G–GDP from the ribosome
Thiostrepton	Prokaryotic cells	Factor-binding center of 50S subunit	Interferes with the association of IF2 and EF-G with factor-binding center
Kirromycin		EF-Tu	Prevents conformational changes associated with GTP hydrolysis and therefore EF-Tu release
Ricin and $\alpha$ -sarcin (protein toxins)	Prokaryotic and eukaryotic cells	Chemically modifies RNA in factor-binding center of large ribosomal subunit	Prevents activation of translation factor GTPases
Diphtheria toxin	Eukaryotic cells	Chemically modifies EF-Tu	Inhibits EF-Tu function
Cycloheximide	Eukaryotic cells	Peptidyl transferase center of 60S subunit	Inhibits peptidyl transferase activity

**Box 15-5** (Continued)

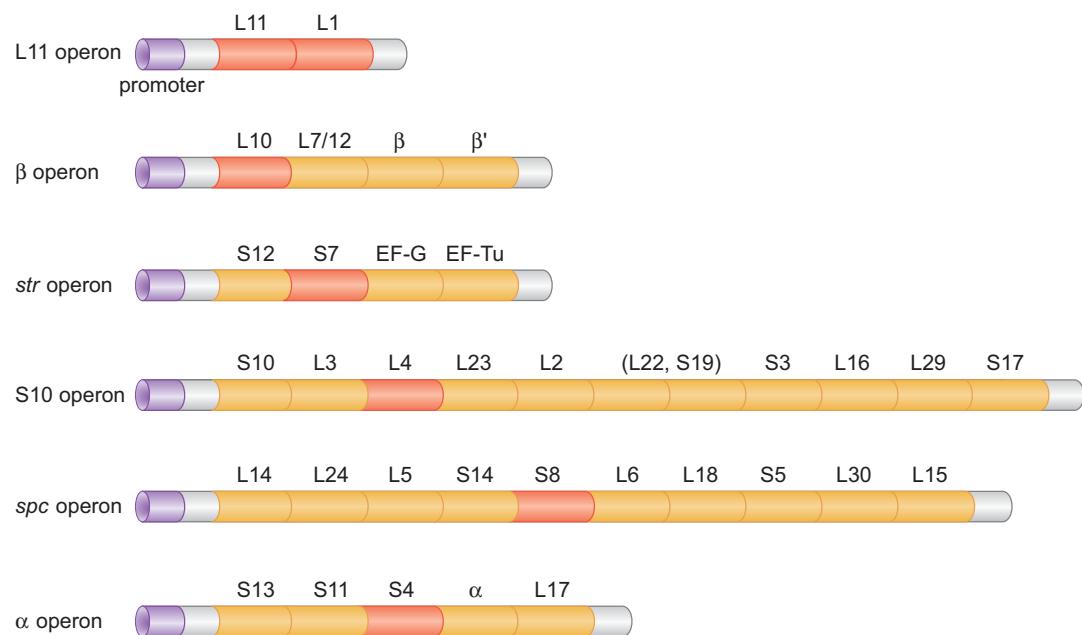
**BOX 15-5 FIGURE 1** Puromycin terminates translation by mimicking a tRNA in the A-site. Puromycin binds in the A-site and participates in peptide-bond formation. Once completed, puromycin and any associated polypeptide diffuse out of the ribosome.

with rRNA expression poses an interesting regulatory problem for the cell. As we discussed above, each ribosome contains more than 50 distinct proteins that should be produced at the same rate as the rRNAs to which they bind. Furthermore, the rate at which a cell makes protein and the number of ribosomes it needs are closely tied to the cell's growth rate. Changes in growth conditions quickly lead to an appropriate increase or decrease in the rate of synthesis of all ribosomal components. How is this coordinated regulation accomplished?

Coordinate regulation of ribosomal protein genes is simplified by their organization into several operons, each containing genes for up to 11 ribosomal proteins (Fig. 15-43). As with other operons, these gene clusters are regulated at the level of RNA synthesis (as we discuss in Chapter 18); however, the most important control of ribosomal protein synthesis is at the level of *translation* of the mRNA. This can be illustrated by a simple experiment. When extra copies of a ribosomal protein operon are introduced into the cell, the amount of mRNA increases correspondingly, but synthesis of ribosomal proteins stays nearly the same. Thus, the cell compensates for extra mRNA by reducing its use as a template for protein synthesis.

The tight control of the translation of ribosomal protein mRNAs is the result of autorepression. For each ribosomal protein operon, one (or a complex of two) of the encoded ribosomal proteins binds that operon's mRNA near the translation initiation sequence of one of the most 5'-proximal genes. Binding of the ribosomal protein sterically inhibits association of the ribosomal small subunit with the nearby RBS, thereby inhibiting translation initiation.

It is easy to see how ribosomal protein binding prevents translation of the initial gene in the operon. But how does this affect the downstream genes that, in some cases, have their own RBSs? Such "polar" effects can occur

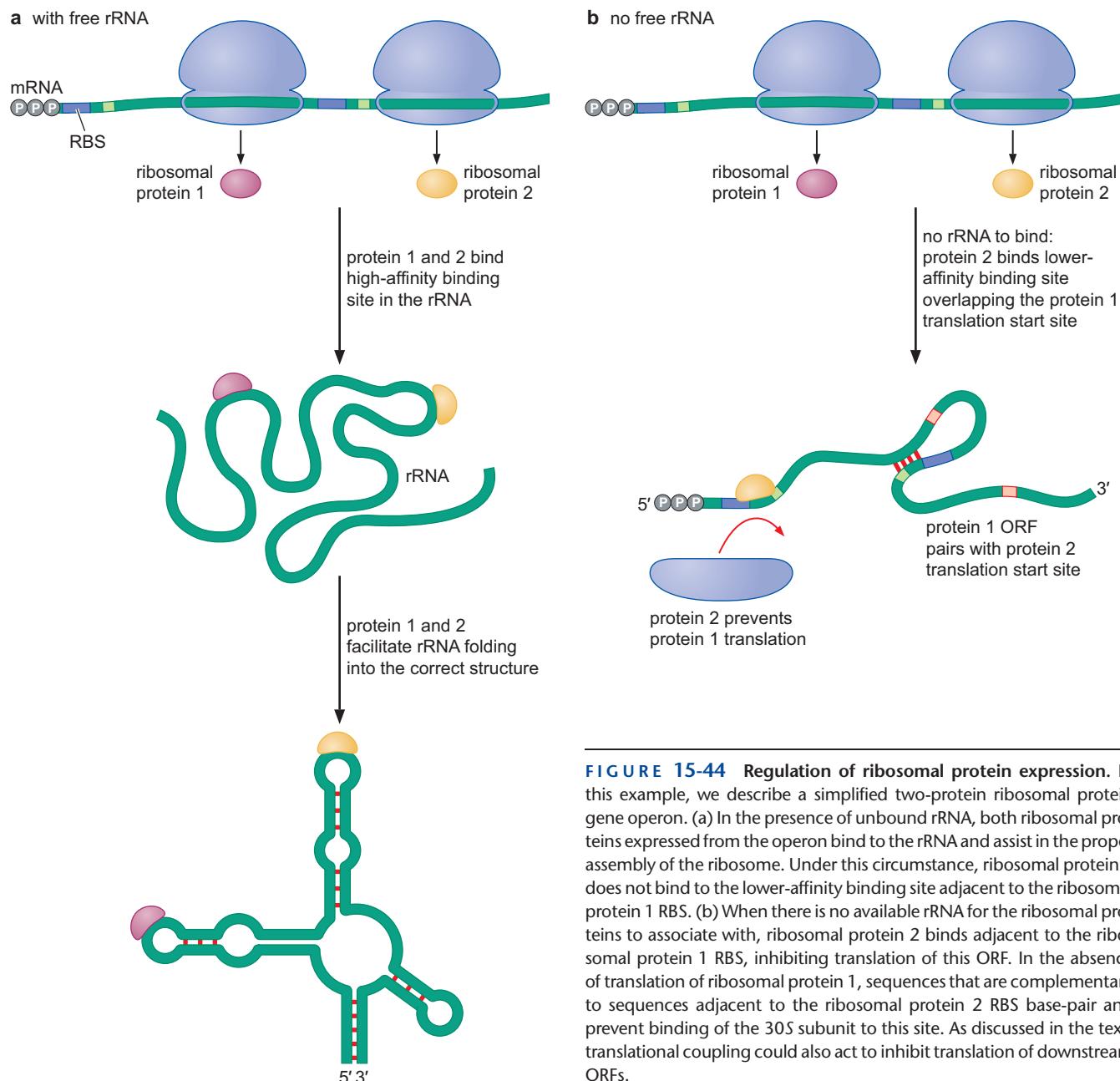


**FIGURE 15-43** *E. coli* ribosomal protein operons. The protein that acts as a translational repressor of the other proteins is shaded red. The promoter is shown in purple, and each ORF is labeled according to the ribosomal protein encoded (e.g., L14 is large ribosomal protein 14). (Adapted, with permission, from Nomura M. et al. 1984. *Annu. Rev. Biochem.* **53**: 75–117. © Annual Reviews.)

through multiple mechanisms. As we discussed above in the chapter, translational coupling may occur when the stop codon of an upstream gene is located very close to the start codon of a downstream gene. This proximity can create a situation in which translation of the upstream gene is required for translation of the downstream gene. A second mechanism exploits the folding of mRNAs into particular structures. The ribosomal protein operon mRNAs frequently are folded into structures that only allow recognition of internal RBSs if earlier genes in the mRNA are being translated. For example, suppose a region in the coding region of the first gene in the mRNA were to base-pair with a site near the RBS of the second gene. Under these circumstances, the 16S rRNA could only recognize this RBS after the inhibitory base pairing is disrupted by a ribosome translating through the first coding region (Fig. 15-42b).

How is expression of the ribosomal proteins coupled to the amount of rRNA in the cell? In each case, the regulatory ribosomal protein that binds the mRNA also recognizes a very strong binding site on the appropriate rRNA (Fig. 15-44). If this binding site is unoccupied, then the ribosomal protein will preferentially bind there. On the other hand, if all of these rRNA-binding sites are occupied, then the regulatory protein will bind to the second, lower-affinity binding site on its own mRNA. Thus, only when the ribosomal protein is present in excess to its target rRNA will it bind its own mRNA. This simple competitive binding event ensures that ribosomal protein synthesis is inhibited only when the regulatory ribosomal protein is in excess.

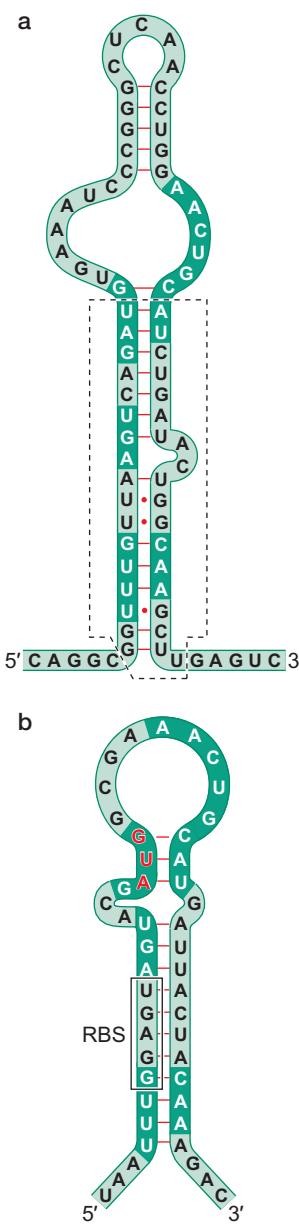
Not surprisingly, in several instances, the two binding sites for the regulatory ribosomal protein are related to each another. In the case of the S8 ribosomal protein, the two binding sites share substantial similarities (Fig. 15-45). The sequence of the binding site in the mRNA reveals a clear mechanism by which S8 inhibits translation. The binding site in



**FIGURE 15-44** Regulation of ribosomal protein expression. In this example, we describe a simplified two-protein ribosomal protein gene operon. (a) In the presence of unbound rRNA, both ribosomal proteins expressed from the operon bind to the rRNA and assist in the proper assembly of the ribosome. Under this circumstance, ribosomal protein 2 does not bind to the lower-affinity binding site adjacent to the ribosomal protein 1 RBS. (b) When there is no available rRNA for the ribosomal proteins to associate with, ribosomal protein 2 binds adjacent to the ribosomal protein 1 RBS, inhibiting translation of this ORF. In the absence of translation of ribosomal protein 1, sequences that are complementary to sequences adjacent to the ribosomal protein 2 RBS base-pair and prevent binding of the 30S subunit to this site. As discussed in the text, translational coupling could also act to inhibit translation of downstream ORFs.

the messenger includes the initiating AUG. Thus, mRNA bound by excess protein S8 (in this example) cannot attach to ribosomes to initiate translation. The differences in the two binding sites explain how binding to the rRNA can be stronger than to mRNA; thus translation is repressed only when the need for the S8 protein in ribosome assembly is satisfied.

This strategy for translational inhibition is not restricted to ribosomal proteins. Other RNA-binding proteins regulate their expression by binding to their own mRNAs, including some aminoacyl-tRNA synthetases. In addition, there are instances in which mRNAs fold into different structures that favor or inhibit translation depending on the cellular conditions (e.g., temperature or metabolite levels) (see Chapter 20).



**FIGURE 15-45** Ribosomal protein S8 binds 16S rRNA and its own mRNA. The comparison shows the regions of the two RNAs bound by the ribosomal protein S8 (encoded by the *spc* operon) (Fig. 15-43). (a) The region of the 16S rRNA bound by the S8 protein. (b) The translation initiation site of ribosomal protein S8 that is bound when there is no available 16S rRNA to bind. Shared sequences are shaded in dark green. The dashed lines box off the region of the 16S rRNA protected by the S8 protein. The AUG (in red) and the RBS binding site (boxed) for the S8 protein mRNA are indicated (b). (Adapted, with permission, from Cerretti D.P. et al. 1988. *J. Mol. Biol.* 204: 309–329. © Elsevier.)

### Global Regulators of Eukaryotic Translation Target Key Factors Required for mRNA Recognition and Initiator tRNA Ribosome Binding

Under conditions of reduced nutrients or other cellular stresses, it is often useful for eukaryotic cells to reduce translation globally. In these instances, two early steps in eukaryotic translation initiation are targeted for inhibition: recognition of the mRNA or initiator tRNA binding to the 40S subunit. Recall from our earlier discussion of initiation of eukaryotic translation that these events occur independently of one another, but inhibition of either eliminates new protein synthesis. In each case, the mechanism of inhibition is controlled by phosphorylation.

One common mechanism of inhibition is mediated by phosphorylation of eIF2. Recall that eIF2 bound to GTP is required to deliver the initiator tRNA to the P-site of the 40S subunit of the eukaryotic ribosome. Several protein kinases have been identified that phosphorylate the  $\alpha$  subunit of eIF2. Phosphorylation of this subunit inhibits the action of a GTP-exchange factor for eIF2, called eIF2B, leading to reduced levels of eIF2–GTP. Similar to the action of EF-Ts on EF-Tu–GDP, eIF2B stimulates eIF2–GDP to release its bound GDP and bind GTP. Because eIF2 bound to GTP is required to escort the initiator tRNA to the 40S subunit, reduced levels of eIF2–GTP limit initiation of translation. The known eIF2 $\alpha$  kinases are activated by several different cellular conditions including amino acid starvation (see later discussion), viral infection, and elevated temperature.

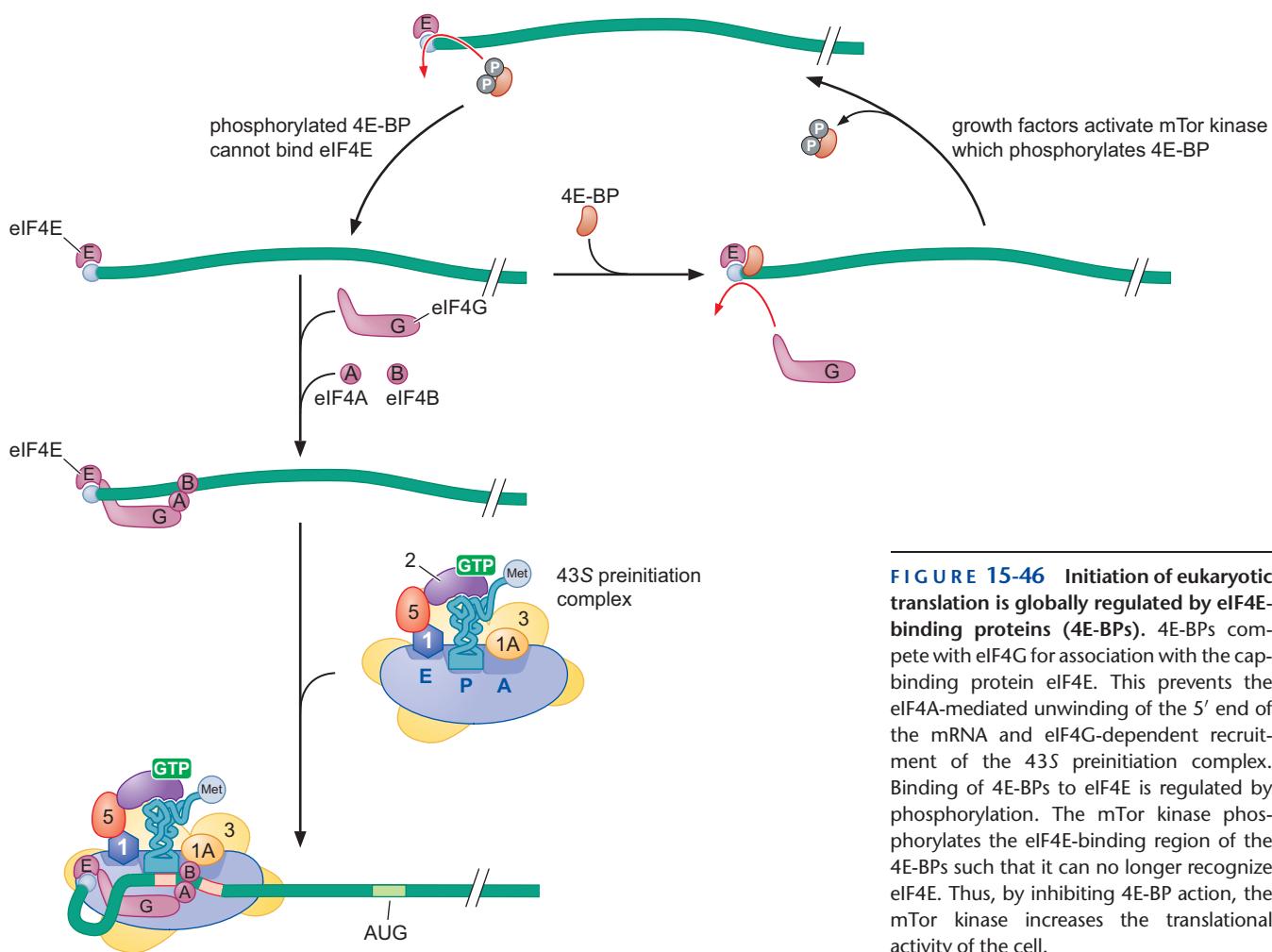
A second mechanism to globally inhibit translation initiation targets the 5'-cap-binding protein: eIF4E. Recall that after binding to the 5' cap, eIF4E binds to eIF4G. The short domain of eIF4G that is recognized by eIF4E is also found in a small family of proteins called eIF4E-binding proteins (4E-BPs). These proteins compete with eIF4G for binding to eIF4E and therefore act as general inhibitors of translation initiation (Fig. 15-46). Like eIF2, the 4E-BPs are also regulated by phosphorylation. In their unphosphorylated state, 4E-BPs bind to eIF4E tightly and inhibit translation. In contrast, phosphorylation of 4E-BPs inhibits their binding to eIF4E (Fig. 15-46).

Phosphorylation of 4E-BPs is mediated by a key cellular protein kinase called mTor. Growth factors, hormones, and other factors that stimulate cell division activate this kinase and therefore increase the overall translational capacity of the cell. These observations have led to the hypothesis that the control of translation capacity is carefully coordinated with cell proliferation. Indeed, overexpression of eIF4E can result in cancerous transformation of cells, and inhibitors of mTor (e.g., rapamycin) are effective chemotherapy agents. Although we have discussed these regulatory mechanisms in the context of the global control of translation, both are also used to regulate the translation of specific mRNAs in the cell as we shall see later.

### Spatial Control of Translation by mRNA-Specific 4E-BPs

In addition to globally regulating translation, binding to eIF4E is also used to regulate the translation of specific mRNAs. For example, the correct establishment of the anterior–posterior axis of the *Drosophila melanogaster* oocyte (egg) and developing embryo requires the correct localization of many proteins within a large shared cytoplasm (see Chapter 21 for a complete description of these events). In several instances, spatially restricted translation of these critical regulatory proteins plays a key role in controlling their localization.

The Oskar protein is carefully localized to the posterior regions of the oocyte before fertilization. Despite this, *Oskar* mRNA is synthesized by attached nurse cells of the ovary of the mother fly and deposited into the *anterior* of the oocyte before fertilization. *Oskar* mRNA is then transported



**FIGURE 15-46** Initiation of eukaryotic translation is globally regulated by eIF4E-binding proteins (4E-BPs). 4E-BPs compete with eIF4G for association with the cap-binding protein eIF4E. This prevents the eIF4A-mediated unwinding of the 5' end of the mRNA and eIF4G-dependent recruitment of the 43S preinitiation complex. Binding of 4E-BPs to eIF4E is regulated by phosphorylation. The mTor kinase phosphorylates the eIF4E-binding region of the 4E-BPs such that it can no longer recognize eIF4E. Thus, by inhibiting 4E-BP action, the mTor kinase increases the translational activity of the cell.

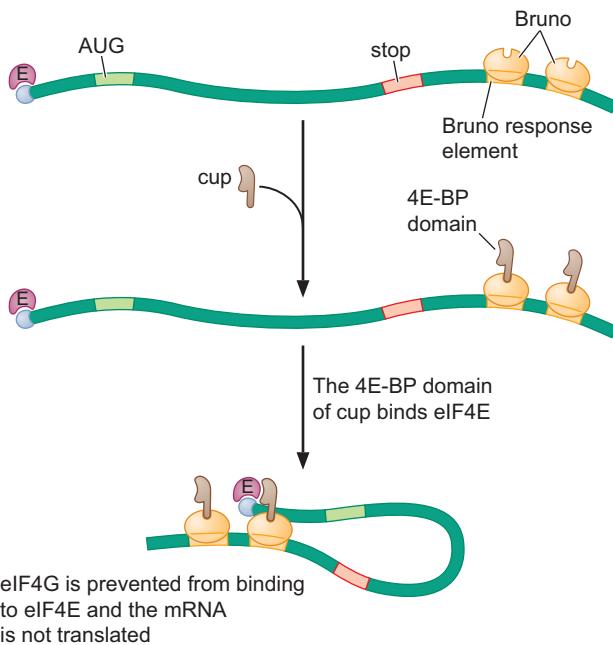
to the posterior region of the oocyte. For the cell to restrict Oskar expression to the posterior region, it is critical that *Oskar* mRNA not be translated as it moves from the anterior to the posterior region of the oocyte.

The action of a 4E-BP called Cup is critical to specifically repressing translation of *Oskar* mRNA (Fig. 15-47). The *Oskar* mRNA contains several sequences in the 3'-UTR that specifically bind to a protein called Bruno. Bruno, in turn, binds to Cup, recruiting this 4E-BP to *Oskar* mRNA. When localized to *Oskar* mRNA, Cup outcompetes eIF4G for binding to eIF4E, thus inhibiting translation of the mRNA. Cup is not abundant enough to act generally on all translation as do the global 4E-BPs described above. Nevertheless, when localized to a particular mRNA, Cup becomes a very effective inhibitor of translation. This mechanism is not exclusive to *Oskar*. The Nanos protein in *Drosophila* is also regulated by recruitment of Cup to its mRNA. Similarly, an mRNA-binding protein called CPEB recruits a 4E-BP called Maskin to a number of mRNAs whose translation is inhibited during vertebrate oocyte development.

### An Iron-Regulated, RNA-Binding Protein Controls Translation of Ferritin

Regulating iron levels in the human body is critical. Many proteins use iron as a cofactor, including the oxygen transport proteins hemoglobin and myoglobin, as well as many of the proteins involved in oxidative

**FIGURE 15-47** The eIF4E-binding protein Cup acts to specifically inhibit Oskar mRNA translation. As Oskar mRNA is transported from the anterior to the posterior of the *Drosophila* oocyte, it is critical that it is not translated. The inhibition of Oskar translation is mediated by two proteins. The RNA-binding protein Bruno binds to multiple sequences in the 3'-UTR of Oskar mRNA called Bruno response elements (BREs). Bruno then recruits the 4E-BP Cup to the mRNA. When localized to the mRNA, Cup outcompetes eIF4G for binding to eIF4E, inhibiting translation of this mRNA.



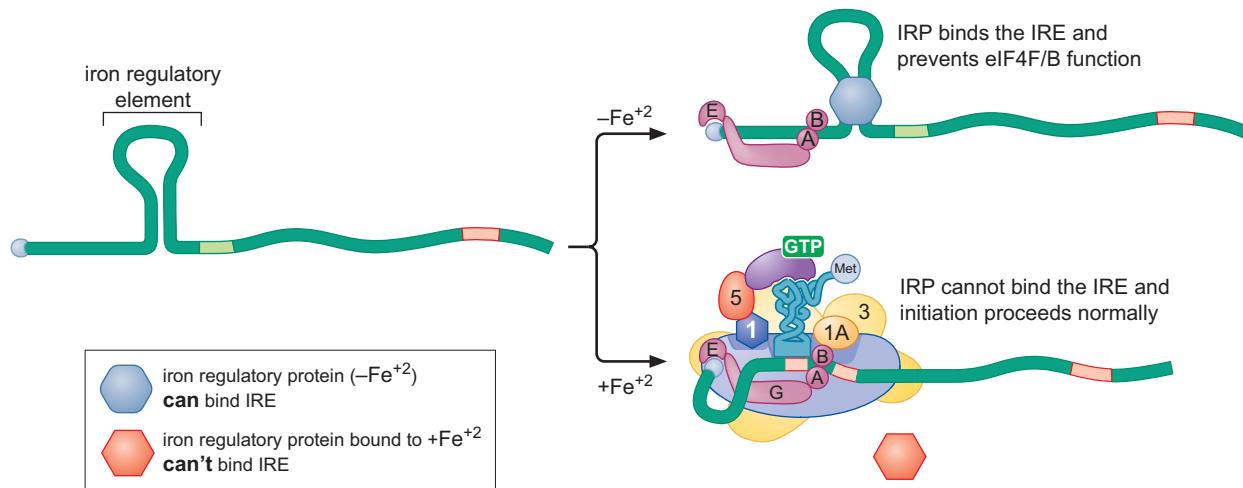
phosphorylation. Consistent with the important role of iron in oxygen transport and energy production, a shortage of iron in the human body (called anemia) results in an overall feeling of weakness. On the other hand, excess iron is toxic to cells and can contribute to liver damage, heart failure, and diabetes.

The iron-binding protein Ferritin is the major regulator of iron levels in the human body. Ferritin stores and releases iron in a controlled manner, thereby maintaining proper iron homeostasis. Thus, the levels of Ferritin must respond rapidly to the levels of free iron in the body. The need to respond rapidly to changes in the levels of free iron has resulted in the regulation of Ferritin expression at the level of protein synthesis.

Ferritin translation is regulated by iron-binding proteins called iron regulatory proteins (IRPs). These proteins are also RNA-binding proteins that recognize a specific hairpin structure formed at the 5' end of the *ferritin* mRNA called the iron regulatory element (IRE) (Fig. 15-48). Importantly, the ability of these proteins to recognize the IRE is controlled by the levels of iron in the cell. In iron-deficient cells, the concentration of iron is too low to bind the IRPs. In the absence of bound iron, these proteins bind tightly to the IRE and inhibit the ability of eIF4A/B to unwind the IRE hairpin structure. The continued presence of the hairpin acts as a steric block to 43S complex mRNA binding. In contrast, when the concentration of free iron in the cell is elevated, the IRPs bind iron. When bound to iron, the IRPs lose their ability to bind to the IRE and, therefore, are not able to inhibit translation.

### Translation of the Yeast Transcriptional Activator Gcn4 Is Controlled by Short Upstream ORFs and Ternary Complex Abundance

Gcn4 is a yeast transcriptional activator that regulates the expression of genes encoding enzymes that direct amino acid biosynthesis. Although it is a transcriptional activator, Gcn4 is itself regulated at the level of translation. In the presence of low levels of amino acids, *Gcn4* mRNA is translated (and thus the biosynthetic enzymes are expressed). But in the presence of



**FIGURE 15-48** Regulation of Ferritin translation by iron. The 5'-UTR of the *ferritin* genes includes a stem–loop structure called the iron regulatory element (IRE). The iron regulatory protein (IRP) binds tightly to this site when it is not bound to  $\text{Fe}^{2+}$ . By stabilizing the stem–loop structure of the IRE, IRP prevents eIF4A from removing this structure from the end of the *ferritin* mRNA. Under these conditions, association of the 43S preinitiation complex with the mRNA cannot occur and the *ferritin* genes are not translated. When iron levels are elevated and Ferritin protein is needed, IRP binds to  $\text{Fe}^{2+}$ , which inhibits its ability to bind to the IRE and, therefore, allows translation of the Ferritin protein.

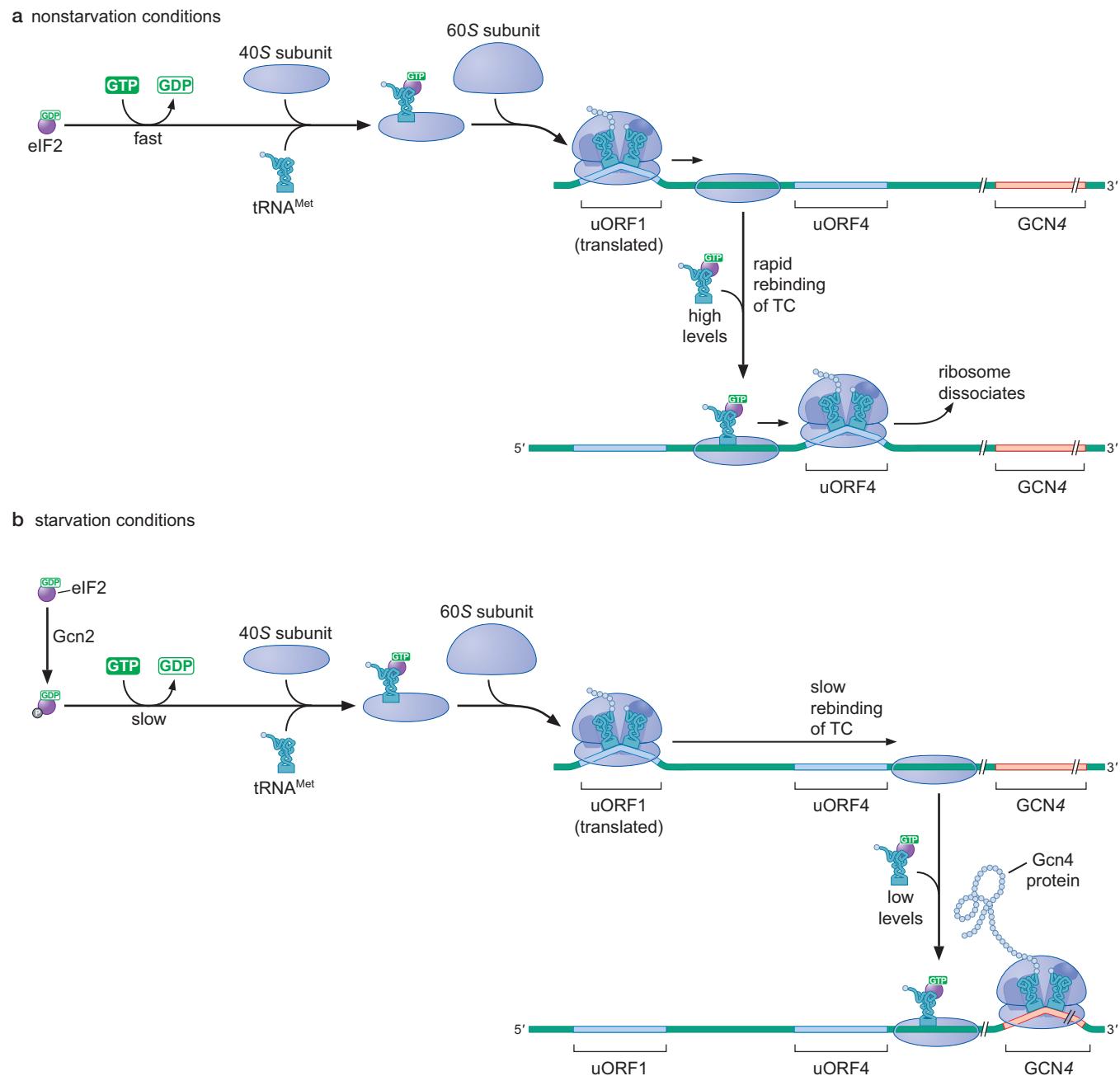
high levels of amino acids, *Gcn4* mRNA is not translated. How is this regulation achieved?

Unlike the structure of the typical eukaryotic message, the mRNA encoding the *Gcn4* protein contains four small open reading frames (called uORFs) upstream of the coding sequence for *Gcn4*. The most upstream of these short ORFs (uORF1) is efficiently recognized by ribosomes that scan along the message from the 5' end. Once they have translated uORF1, a unique property of this ORF allows 50% of the small subunits of the ribosome to remain bound to the RNA and resume scanning for downstream initiation (AUG) codons (Fig. 15-49) (see Box 15-3).

Which downstream AUG is recognized by the scanning small subunit is controlled by when the small subunit binds to eIF2 complexed with an initiator tRNA. Recall that in the absence of an initiator tRNA in the P-site, the 40S subunit cannot recognize an AUG sequence in the mRNA. Thus, before initiating translation at any downstream ORF, the scanning 40S ribosome subunit must bind eIF2·Met-tRNA<sub>i</sub><sup>Met</sup> (hereafter referred to as TC, for ternary complex).

When amino acids are not limiting, TC rebinds the scanning ribosomes soon after they complete translation of uORF1 (Fig. 15-49a). Once rebound by TC, the small subunits can recognize an AUG and reinitiate translation at one of the other uORFs (2, 3, or 4). Unlike uORF1, after translating these uORFs, the ribosome fully dissociates from the mRNA and fails to reach the *Gcn4* ORF. Thus, no *Gcn4* protein is made.

Under conditions of amino acid starvation, a combination of events reduces the rate at which the TC binds to the 40S subunit. Limited amino acids lead to an abundance of uncharged tRNAs, which, in turn, activates an eIF2 $\alpha$  kinase called Gcn2. As described for the global control of translation by eIF2 $\alpha$  kinases, when Gcn2 phosphorylates eIF2, the population of eIF2-GTP is reduced because the interaction of eIF2 with the GTP-exchange factor eIF2B is inhibited. Because eIF2 can only bind Met-



**FIGURE 15-49** Translational control of Gcn4 in response to amino acid starvation. As described in detail in the text, the ORF encoding the yeast activator Gcn4 is preceded by four short ORFs called uORFs (here, only uORF1 and uORF4 are shown). The first of these upstream ORFs is translated initially and, because of special properties of this ORF, approximately half of the 40S subunits are retained after translation termination to continue scanning the Gcn4 mRNA. (a) When amino acids are abundant, eIF2B stimulates eIF2 to exchange GDP for GTP rapidly. This allows for rapid binding of eIF2–GTP–Met-tRNA<sup>Met</sup> to the 40S subunit and the ability to recognize one of the three other short ORFs. Translation of any one of these uORFs results in full termination of translation. (b) Under starvation conditions, phosphorylation of eIF2 by the eIF2 $\alpha$  kinase Gcn2 reduces the ability of eIF2B to stimulate GTP binding to eIF2. Reduced levels of eIF2-GTP result in slower binding of eIF2–GTP–Met-tRNA<sup>Met</sup> to the 40S subunit. This reduced rate of initiator tRNA binding increases the chance that the scanning ribosome will pass uORF4 before being able to recognize an AUG and therefore favors the translation of Gcn4. (Modified, with permission, from Hinnebusch A.G. 1997. *J. Biol. Cell* **272**: 21661–21664, Fig. 1. © American Society for Biochemistry & Molecular Biology.)

tRNA<sub>i</sub><sup>Met</sup> in the presence of GTP, these conditions lead to less TC and reduce the rate of TC binding to the 40S subunit (Fig. 15-49b). The reduced rate of binding means that 40S subunit scanning continues farther along the mRNA without the ability to detect an AUG. If the ribosome scans through the AUGs for uORF2–4 before rebinding the TC, then these ORFs will not be translated. The start codons for uORF2–4 are relatively close to uORF1, whereas the AUG of the Gcn4 ORF is much further downstream. This additional distance provides a larger window of time for the TC to bind to the small subunit before the Gcn4 ORF is encountered, thereby increasing the odds that the ribosome translates it. Indeed, removing the spacer RNA between the uORFs and the Gcn4 start codon results in progressively less Gcn4 protein expression. Thus, in the presence of limiting TC, Gcn4 is produced and can switch on the genes needed to synthesize additional amino acids in the cell. The use of next-generation sequencing technologies to analyze global translation patterns—that is, measuring all mRNAs that are being translated—is described in Box 15-6, Ribosome and Polysome Profiling.

## ► TECHNIQUES

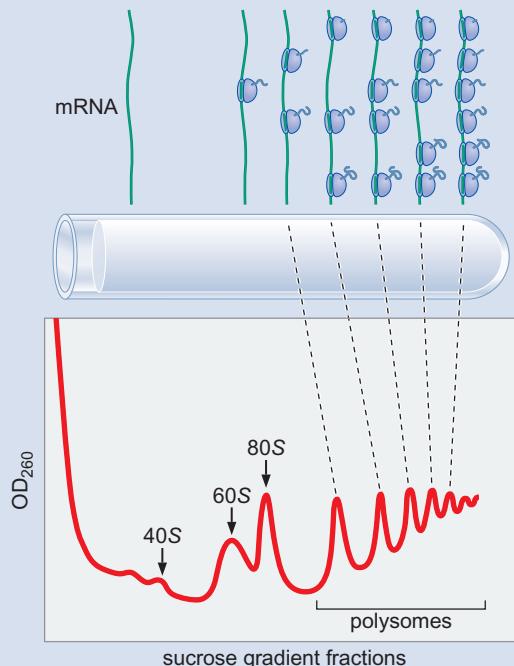
### Box 15-6 Ribosome and Polysome Profiling

There are many ways to measure the level of gene expression in a cell. Many assays focus on determining the amount of the primary product of gene expression, the RNA. On the other hand, the existence of an mRNA in a cell does not mean that the protein encoded by the mRNA is being expressed. To determine if a protein is actually being synthesized in a cell, one must measure the translation of that mRNA.

Early studies to determine the extent of translation of an mRNA focused on the association of the mRNA with polysomes. In these studies, researchers added the drug cycloheximide to cells to arrest translation at the translocation stage. The treated cells are then broken open, and the resulting cell extract is separated by centrifugation on a sucrose gradient (Box 15-6 Fig. 1). Because of the very large size of the ribosome, the ribosome-associated mRNAs migrate much faster during centrifugation relative to free mRNA. Moreover, mRNAs that are part of a polysome migrate faster still. Measurement of the total protein across the gradient reveals several peaks corresponding to mRNA associated with increasing numbers of ribosomes (i.e., as polysomes). Measuring the presence of mRNA across these fractions (e.g., by Northern blot or RT-PCR) reveals whether that mRNA is being translated.

More recently, this assay has been adapted to take advantage of new sequencing technologies and gain a global measurement of mRNAs that are being translated in a cell population. Called ribosome profiling, as with the polysome profile assay, this assay starts with treating the cell population with cycloheximide and making a cell extract. The next step of the assay takes advantage of the finding that an elongating ribosome protects the ~28-base stretch of the mRNA that is in the decoding center from RNase I digestion (Box 15-6 Fig. 2a). Thus, before sucrose gradient centrifugation, the cell extract is

treated with RNase I. After centrifugation, the fractions containing the 80S particles (the ribosome plus its protected mRNA) are isolated and denatured, and the 28-base RNAs associated are



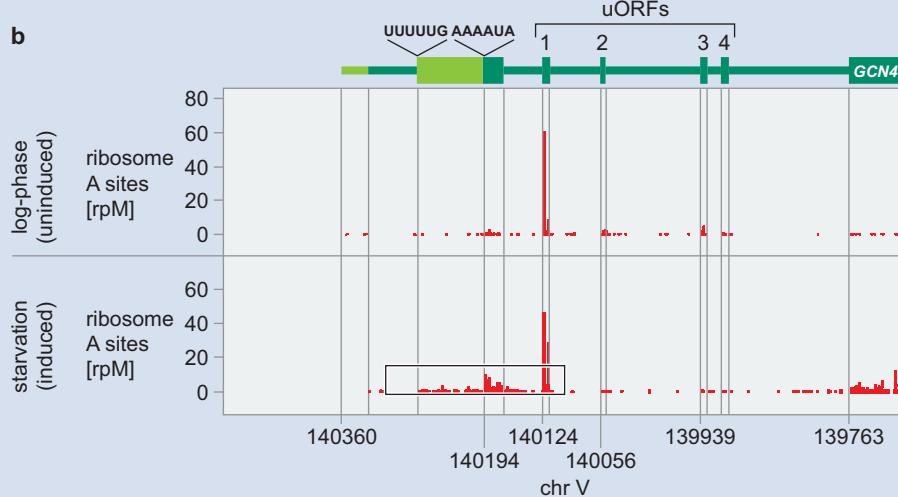
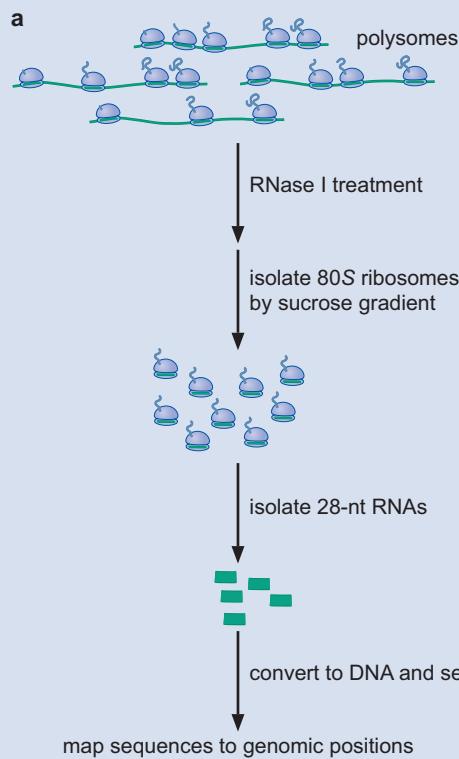
**BOX 15-6 FIGURE 1** Polysome profiling. mRNAs with different numbers of associated ribosomes (polysomes) can be separated from free mRNA and single ribosomal subunits by sucrose gradient centrifugation. The presence of an mRNA in the polysome fraction is indicative of that mRNA being actively translated.

**Box 15-6 (Continued)**

isolated by gel purification. After conversion to double-stranded DNA, the RNAs are subjected to deep sequencing. By analyzing the resulting DNA sequences, one can readily determine not only the mRNAs that were being actively translated at the time of cycloheximide treatment, but also exactly where each ribosome was located. Indeed, the method is precise enough to determine which codons were in the A- and P-sites!

This approach has rapidly become an important method to analyze gene expression. By comparing the numbers of

sequence reads within each gene between two cell populations, one can identify the changes in the proteins being produced in two cell populations (e.g., growing in high or low amino acid levels) (Box 15-6 Fig. 2b). In addition, owing to its precision, this method can also reveal attributes of translation. For example, this technique has been used to determine that many of the sites of translational pausing in *E. coli* cells occur because of the presence of an RBS-like sequence within ORFs.



**BOX 15-6 FIGURE 2** Ribosome profiling. (a) Polysomes treated with RNase I are separated on a sucrose gradient to isolate monoribosomes (80S particles). These ribosomes have only 28 nucleotides of mRNA that was protected from degradation by the ribosome. The resulting mRNA fragments are gel-purified and converted to DNA (using reverse transcriptase) and subjected to deep DNA sequencing. The resulting sequences are mapped onto the genomic sequence of the organism to reveal the sites of ribosome engagement, representing the sites of translation. The more sequences isolated from a particular region, the more actively it is being translated. (b) Representative data from ribosome profiling of the *S. cerevisiae* *GCN4* gene under rich media and amino acid starvation growth conditions. The plot is a histogram of a normalized number of times a particular sequence was present under each condition (reads per million bases or rpM). Note the different extents of translation of the uORFs and the *Gcn4* ORF in the two conditions.

## TRANSLATION-DEPENDENT REGULATION OF mRNA AND PROTEIN STABILITY

At some frequency, mRNAs will be made that are mutant or damaged. Such defective mRNAs can arise from mistakes in transcription or from damage that occurs after they are synthesized. For example, because they are single-stranded, mRNAs are more susceptible to breakage. Such damaged mRNAs have the possibility of making incomplete or incorrect proteins that could have negative effects on the cell. In some cases, such as point mutations that change only a single amino acid, there is little that can be done to eliminate the mutant mRNA or its protein product. However, in other cases described later, the process of translation is used to detect defective mRNAs and eliminate them and their protein products.

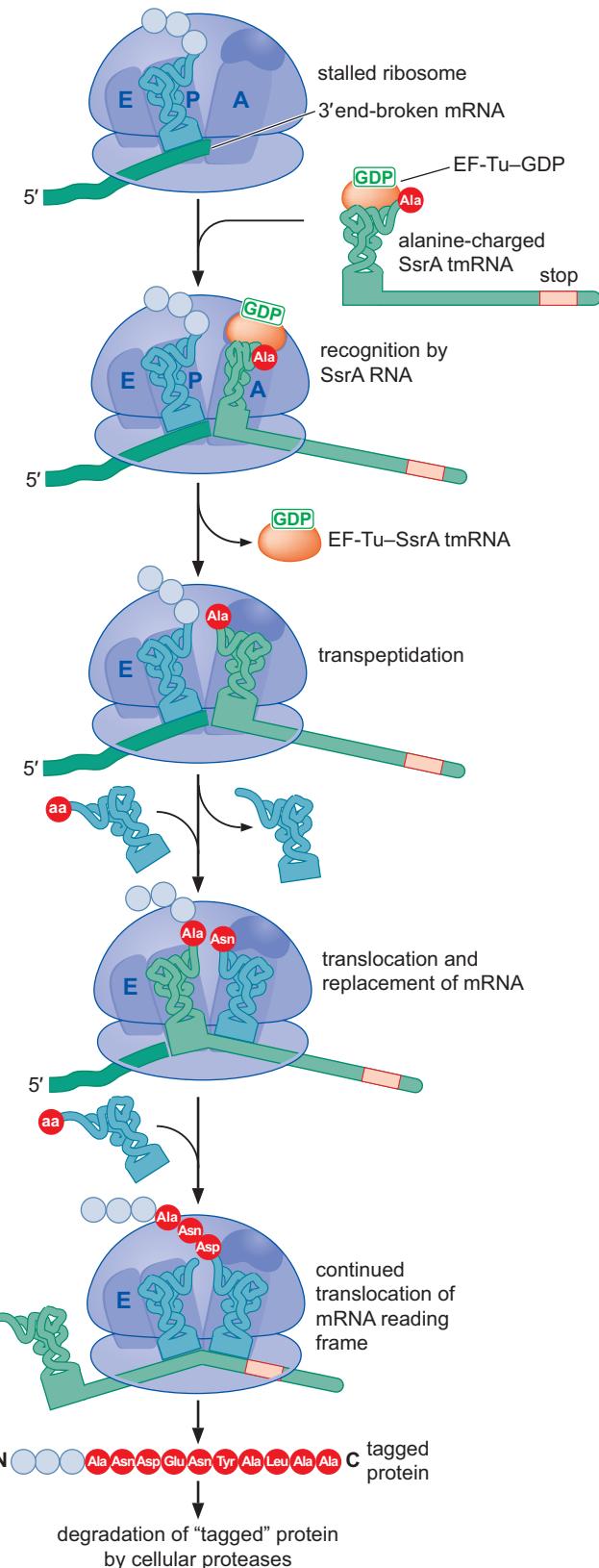
### The SsrA RNA Rescues Ribosomes That Translate Broken mRNAs

Normally, a stop codon is required to release the ribosome from an mRNA. But what happens to a ribosome that initiates translation of an mRNA fragment that lacks a termination codon in the appropriate reading frame? Such an mRNA can be generated by incomplete transcription or nuclease action. Translation of this type of mRNA can initiate normally and continue until the 3' end of the mRNA is reached. At this point, the ribosome cannot proceed. There is no codon to bind either an aminoacyl-tRNA or a release factor. Without some mechanism to release them from these defective mRNAs, many ribosomes would be permanently trapped, removing them from polypeptide synthesis. In prokaryotic cells, such stalled ribosomes are rescued by the action of a chimeric RNA molecule that is part tRNA and part mRNA, appropriately called a **tmRNA**.

SsrA is a 457-nucleotide tmRNA that includes a region at its 3' end that strongly resembles tRNA<sup>Ala</sup> (Fig. 15-50). This similarity allows the SsrA RNA to be charged with alanine and to bind EF-Tu–GTP. When a ribosome is stalled at the 3' end of an mRNA, the SsrA<sup>Ala</sup>–EF-Tu–GTP complex binds to the A-site of the ribosome and participates in the peptidyl transferase reaction, as would any other tRNA. Translocation of the peptidyl-SsrA RNA results in the release of the broken mRNA. Remarkably, translocation of the SsrA RNA also results in a portion of this RNA entering the mRNA-binding channel of the ribosome. This portion of the SsrA RNA acts as an mRNA and encodes 10 codons followed by a stop codon.

The net result of SsrA binding is that when the defective mRNA is released from the ribosome, the protein encoded by the incomplete mRNA is fused to a 10-amino-acid “peptide tag” at its carboxyl terminus, and the ribosome is recycled. Interestingly, the 10-amino-acid tag is recognized by cellular proteases that rapidly degrade the tag and the truncated polypeptide to which it is attached. Thus, translation products arising from broken mRNAs are rapidly cleared to prevent these defective proteins from harming the cell.

How does the SsrA RNA bind only to stalled ribosomes? Because of the large size of SsrA (it is more than four times longer than a standard tRNA), it cannot bind to the A-site during normal elongation. In contrast, when the 3' end of the mRNA is missing, additional room is created in the A-site to accommodate the larger RNA (Fig. 15-50). Thus, only ribosomes stalled at or very near the 3' end of an mRNA are binding sites for the SsrA RNA. SsrA has recently been revealed to be the target of one of the drugs used in combination for the treatment of tuberculosis (see Box 15-7, A Frontline Drug in Tuberculosis Therapy Targets SsrA Tagging).



**FIGURE 15-50** The tmRNA SsrA rescues ribosomes stalled on prematurely terminated mRNAs. The SsrA RNA mimics a tRNA but can only bind a ribosome that is stalled at the 3' end of an mRNA. Once bound, the SsrA RNA substitutes part of its sequence to act as a new "mRNA."

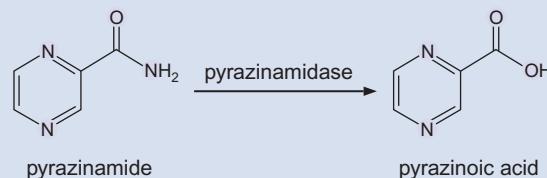
## MEDICAL CONNECTIONS

### Box 15-7 A Frontline Drug in Tuberculosis Therapy Targets SsrA Tagging

One of the great scourges of humankind, tuberculosis, dates back to antiquity, with evidence of the disease seen in Egyptian mummies. Its excruciating symptoms include chronic cough, blood in the sputum, and weight loss (hence the old name “consumption”). The pathogen that causes tuberculosis is the bacterium *Mycobacterium tuberculosis*. One-third of the world’s population is thought to be infected with *M. tuberculosis*, but in a majority of these individuals, the bacterium remains in a latent state, and the individuals are disease-free. About 10% of infected individuals develop active disease over the course of their lifetime, resulting in about 8 million people with tuberculosis at any time and roughly 1.5 million deaths yearly. Individuals with HIV/AIDS and other immune compromised individuals are at particularly high risk for tuberculosis.

Tuberculosis is treated with a combination of four frontline drugs: rifampicin, isoniazid, ethambutol, and pyrazinamide. The mode of action of three of these drugs is well understood. Rifampicin is an inhibitor of bacterial RNA polymerases (Chapter 18), and isoniazid and ethambutol each inhibit the synthesis of different components of the *M. tuberculosis* cell envelope. The mechanism of action of pyrazinamide, however, has remained a mystery since its therapeutic effects were discovered more than half a century ago. Pyrazinamide is a pro-drug that is converted to the active agent pyrazinoic acid upon entry into the bacterial cell by a deaminase (pyrazinamidase) (Box 15-7 Fig. 1). Recently, a team of scientists from the United States, China, and South Korea has identified SsrA (see the text and Fig. 15-50) as a target of pyrazinoic acid.

Shi et al. (2001) tackled the problem of how pyrazinoic acid works by immobilizing it on a column and performing affinity chromatography to identify proteins from *M. tuberculosis* that bind to the drug. One of the proteins identified was the largest protein in the small subunit of the ribosome, ribosomal protein S1 or RpsA. RpsA plays an essential role in translation, binding mRNA upstream of the Shine–Dalgarno sequence and helping anchor mRNA to the small subunit during



**BOX 15-7 FIGURE 1** Conversion of pyrazinamide to pyrazinoic acid.

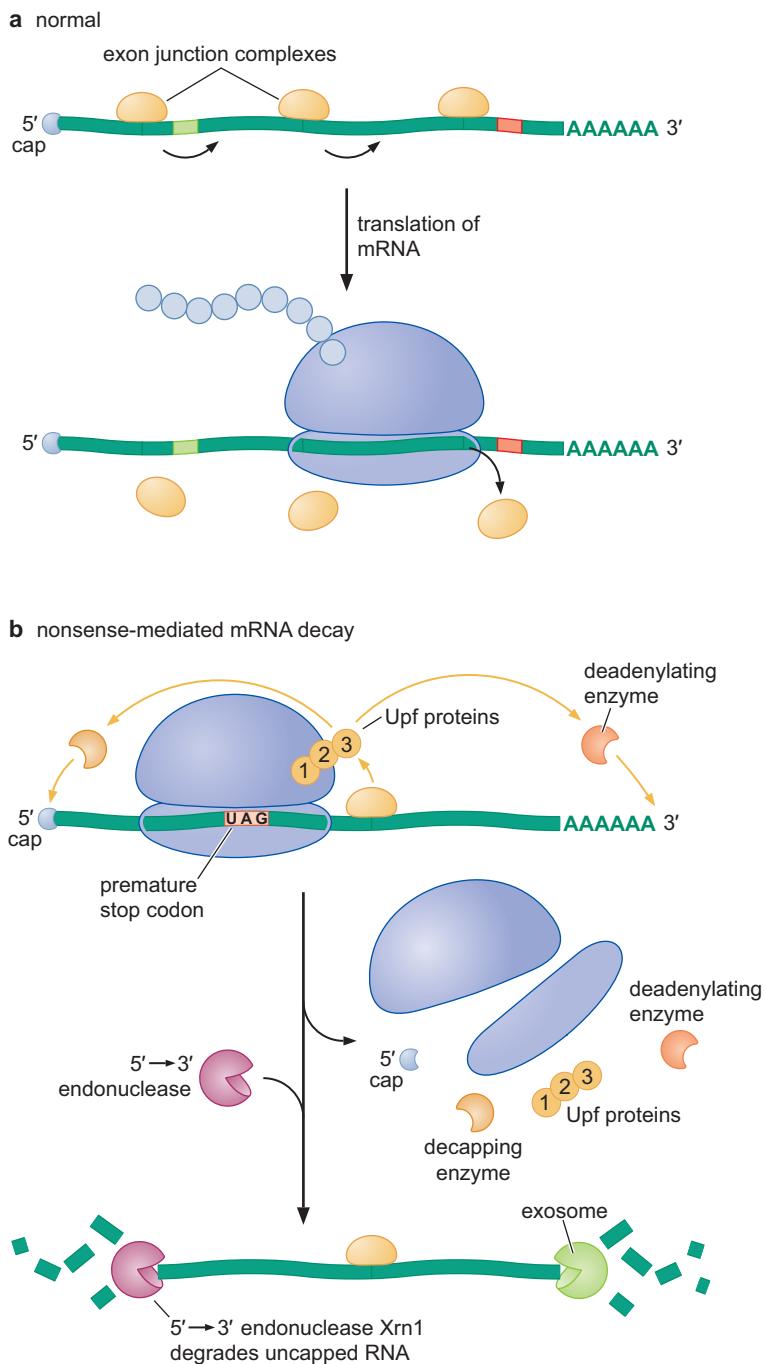
translation initiation. But RpsA is also implicated in SsrA tagging. SsrA RNA is delivered to the stalled ribosome in a complex with a dedicated SsrA-binding protein (SmpB) and, intriguingly, RpsA (as well as EF-Tu-GTP). Shi et al. discovered that mutants that show resistance to the drug are altered in RpsA. They also found that pyrazinoic acid blocks the binding of SsrA RNA to wild-type RpsA but not to the pyrazinoic acid–resistant mutant RpsA. Finally, and most importantly, they found that the drug inhibits SsrA tagging in an *in vitro* protein synthesis assay with a modified mRNA that requires SsrA function. As a control, pyrazinoic acid had little effect on protein synthesis with a normal mRNA template, showing that it was not inhibiting the normal function of RpsA.

The pyrazinamide story underscores the growing connections between chemical biology and molecular biology in modern medicine. Important new medical drugs are sometimes discovered by screening libraries of small molecules for activities against proteins that mediate cellular processes known to be important in disease. Conversely, drugs uncovered simply on the basis of their therapeutic effects can lead to previously unappreciated targets governing the working of the cell. Finally, knowing now that pyrazinamide exerts its therapeutic benefit by binding to RpsA opens the door to the discovery of new classes of anti-*M. tuberculosis* compounds that target SsrA tagging.

### Eukaryotic Cells Degrade mRNAs That Are Incomplete or Have Premature Stop Codons

Translation is tightly linked to the process of mRNA decay in eukaryotic cells (Fig. 15-51). This linkage is illustrated by two mechanisms that monitor the integrity of mRNAs that are being translated. For example, when an mRNA contains a premature stop codon (known as a nonsense codon) (see Chapter 16), the mRNA is rapidly degraded by a process called **nonsense-mediated mRNA decay** (Fig. 15-51b). In mammals, recognition of mRNAs with premature stop codons relies on the assembly of protein complexes within the ORF of the mRNA. These exon–junction complexes are assembled on the mRNA as a consequence of splicing and are located just upstream of each exon–exon boundary (see Chapter 14). Ordinarily, when the first ribosome translates an mRNA, these complexes are dis-

**FIGURE 15-51** Eukaryotic mRNAs with premature stop codons are targeted for degradation. (a) Translation of a normal mRNA displaces all of the exon–junction complexes. (b) Nonsense-mediated decay. Translation of an mRNA with a premature stop codon does not displace one or more of the exon–junction complexes. This results in the recruitment of the Upf1, Upf2, and Upf3 proteins to the ribosome. Once bound to the ribosome, these proteins activate a decapping enzyme that removes the 5' cap and a deadenylating enzyme that removes the poly-A tail of the mRNA. The uncapped and deadenylated mRNA is then rapidly degraded by 5'-to-3' (Xrn1) and 3'-to-5' (exosome) exonucleases that are normally unable to degrade the mRNA because of the presence of the 5' cap and poly-A tail.



placed as the mRNA enters the decoding center of the ribosome. However, if a premature stop codon is present in the mRNA (because of mutation of the gene or mistakes in transcription or splicing), then the ribosome is released before the displacement of all of the exon–junction complexes. Under these conditions, the exon–junction complexes and the eRF3 that is bound to the prematurely terminating ribosome recruit a set of proteins to the prematurely terminating ribosome. These proteins recruit and/or activate multiple enzymes that cleave the mRNA, remove the cap at the 5' end of the mRNA, or remove the poly-A tail at the 3' end. Because the mRNA is ordinarily protected from degradation by the 5' cap and the poly-A tail, any of these events result in the exposure of

unprotected 5' or 3' ends, leading to rapid degradation of the mRNA by 5'→3' and 3'→5' exonucleases.

A different process called **nonstop-mediated decay** rescues ribosomes that translate mRNAs that lack a stop codon (Fig. 15-52a). Unlike their prokaryotic counterparts, eukaryotic mRNAs terminate with a poly-A tail. When an mRNA lacking a stop codon is translated, the ribosome translates through the poly-A tail (because there is no stop codon to cause it to terminate before reaching the tail). This results in the addition of multiple lysines to the end of the protein (AAA is the codon for lysine) and stalling of the ribosome at the end of the mRNA. The stalled ribosome is bound by two proteins related to eRF1 and eRF3 (Dom34 and Hbs1) that stimulate ribosome dissociation and release the peptidyl-tRNA and mRNA. A second eRF3-related factor, Ski7, recruits a 3'→5' exonuclease that degrades the “non-stop” mRNA. In addition to these events, as in nonsense-mediated decay, the non-stop mRNA is also cut by an endonuclease. Importantly, proteins that contain polylysine at their carboxyl termini are unstable, leading to the rapid degradation of proteins derived from non-stop mRNAs. Thus, like the situation in prokaryotes, proteins synthesized from mRNAs lacking stop codons are rapidly removed from the cell.

A third mRNA surveillance mechanism related to non-stop-mediated decay is no-go decay (Fig. 15-52b). This mechanism recognizes ribosomes that are stalled on an mRNA. This can occur as a result of a stable mRNA secondary structure in the coding region of the mRNA or because of a stretch of codons for which there are few corresponding tRNAs in the cell. Although the reason for the mRNA stalling are distinct from non-stop decay, the consequences are very similar. Dom34 and Hbs1 bind to the stalled ribosome and stimulate its dissociation into the large and small subunits. In addition, endonucleolytic cleavage of the mRNA is also observed. As for nonsense and non-stop decay, the identity of the endonuclease is unknown.

A fascinating feature of nonsense-, non-stop- and no-go-mediated mRNA decay is that each of these processes requires translation of the damaged mRNA to detect the defect and degrade the mRNA. In the absence of translation, the damaged mRNAs are not rapidly degraded and have normal stability. Thus, although indirect, eukaryotic cells rely on translation as a mechanism to proofread their mRNAs.

## SUMMARY

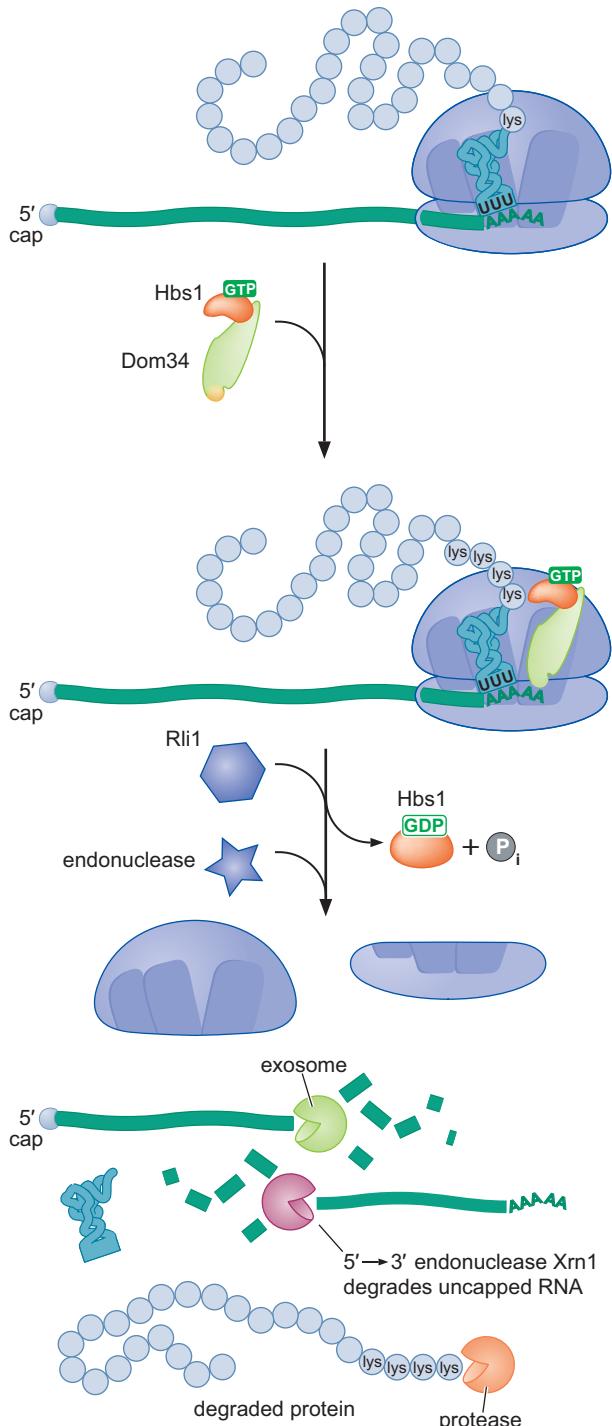
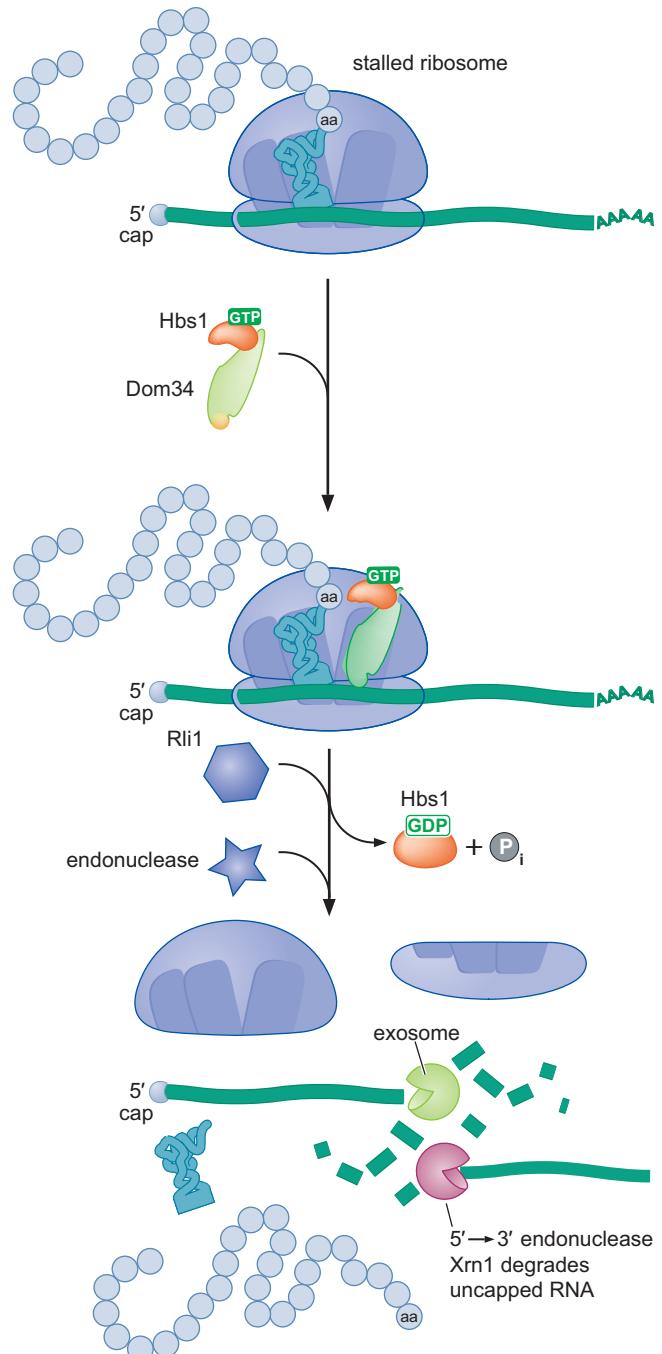
---

Proteins are synthesized on RNA templates known as messenger RNAs (mRNAs) in a process known as translation. Translation involves the decoding of nucleotide sequence information into the linear sequence of amino acids of the polypeptide chain. The machinery for protein synthesis consists of four principal components: the mRNA; adaptor RNAs known as transfer RNAs (tRNAs); aminoacyl-tRNA synthetases that attach amino acids to the tRNAs; and the ribosome, which is a multi-subunit complex of protein and RNA that catalyzes peptide-bond formation.

The mRNA contains the coding sequence for protein and recognition elements for the initiation and termination of translation. The coding sequence is known as an open reading frame (ORF), and consists of a series of three-nucleotide-long units known as codons that are in register with each other. An ORF specifies a single polypeptide chain. Each ORF begins with a start codon and ends with a stop codon.

The start codon is usually AUG or GUG in prokaryotes and always AUG in eukaryotes. In prokaryotes, the start codon is preceded by a region of sequence complementarity to the 16S rRNA component of the ribosome, which is responsible for aligning the ribosome over the start codon. In eukaryotes, the mRNA contains a special structure at its 5' terminus known as the 5' cap, which is responsible for recruiting the ribosome. Eukaryotic mRNAs terminate in a string of A residues known as the poly-A tail, which enhances the efficiency of translation. Prokaryotic mRNAs often contain two or more ORFs; these mRNAs are referred to as being polycistronic. Eukaryotic mRNAs usually contain only a single ORF and are called monocistronic.

tRNAs are the physical interface between codons in the mRNA and the amino acids that are added to the growing polypeptide chain. tRNAs are L-shaped molecules with a loop at one end that displays the anticodon and a 3'-protruding

**a** nonstop-mediated decay**b** no-go-mediated mRNA decay

**FIGURE 15-52** Eukaryotic mRNAs with premature stop codons are targeted for degradation. (a) Nonstop-mediated decay. In the absence of a stop codon, the poly-A tail of the mRNA is translated, leading to the addition of polylysine (AAA encodes Lys) to the end of the protein. Upon reaching the 3' end of the template, the stalled ribosome is recognized by a complex of Dom34 and Hbs1. After delivering Dom34 to the ribosome, Hbs1 hydrolyzes GTP and is released. In combination with the Rli1 ATPase, Dom34 acts to disassemble the ribosome into its two subunits and recruit an endonuclease that cuts the mRNA upstream of the ribosome. The resulting mRNA fragments are degraded by 5'→3' and 3'→5' exonucleases. The protein with the polylysine at the end is also subject to proteolysis. (b) No-go-mediated decay. As with nonstop-mediated decay, no-go-mediated decay is initiated when the ribosome stalls. In this case, the stall is induced by an RNA secondary structure or a stretch of codons demanding charged tRNAs that are present in low abundance (often referred to as rare codons). The stalled ribosome is recognized by Dom34 and Hbs1, and the ribosome is released and the mRNA degraded in a similar manner to non-stop-mediated decay.

5'-CCA-3' sequence at the other end. The anticodon is complementary to the codon, which it recognizes by base pairing. Amino acids are attached to the terminal residue of the 5'-CCA-3' via an acyl linkage between the carbonyl group of the amino acid and the 2'- or 3'-hydroxyl of the terminal ribose.

Aminoacyl-tRNA synthetases attach amino acids to tRNAs in a two-step process known as charging. A single aminoacyl tRNA synthetase is responsible for charging all tRNAs for a specific amino acid. Synthetases recognize the correct tRNAs by interactions with both ends of these L-shaped molecules. Synthetases are responsible for charging their cognate tRNAs with the correct amino acid and do so with high fidelity. Some aminoacyl-tRNA synthetases achieve increased accuracy by means of a proofreading mechanism.

The ribosome consists of a large subunit, which contains the site of peptide-bond formation (the peptidyl transferase center), and a small subunit, which contains the site of mRNA decoding (decoding center). Each subunit is composed of one or more RNAs and multiple proteins. The RNAs not only are a principal structural feature of the subunits, but also are responsible for the principal functions of the ribosome. The intact ribosome contains three tRNA-binding sites that reach between the two subunits: the A-site, where the charged tRNA enters the ribosome; the P-site that contains the peptidyl-tRNA; and the E-site, where deacylated tRNAs exit the ribosome.

Translation of one protein involves a cycle of association and dissociation of the small and large subunits. In this ribosome cycle, the small and then the large subunit assembles at the beginning of an ORF and then dissociates into free subunits when translation of the ORF is complete. The mRNA is translated starting at the 5' end of the ORF, and the polypeptide chain is synthesized in an amino-terminal to carboxy-terminal direction.

Translation takes place in three principal steps: initiation, elongation, and termination. Initiation in prokaryotes involves the recruitment of the small ribosomal subunit to the mRNA through the interaction of the ribosome-binding site (RBS) with the 16S rRNA. This interaction is facilitated by three auxiliary proteins (called initiation factors IF1, IF2, and IF3) that help to keep the two ribosomal subunits apart and recruit a special initiator tRNA to the start codon. Pairing between the anticodon of the charged initiator tRNA and the start codon triggers the recruitment of the large subunit, the release of the initiation factors, and the placement of the charged initiator tRNA in the P-site. This is the prokaryotic initiation complex, and it is poised to accept a charged tRNA into the A-site and perform the formation of the first peptide bond.

Eukaryotic mRNAs recruit the small subunit through recognition of the 5' cap and the action of numerous auxiliary initiation factors. One set of factors functions like the prokaryotic initiation factors to recruit the initiator tRNA to the small subunit. A distinct set of factors unique to eukaryotic cells recognizes the 5' cap and prepares the mRNA to bind to the initiation-factor-bound small subunit (the 43S preinitiation complex). After binding to the mRNA, the small subunit scans downstream until it encounters an AUG, which it recognizes as the start codon. As in prokaryotes, only when the starting AUG is recognized does the large ribosomal subunit associate with the mRNA.

The first step of the elongation phase of translation is the introduction of a charged tRNA into the A-site. This is catalyzed by the GTP-binding protein EF-Tu in prokaryotes and its equivalent in eukaryotes. Multiple mechanisms ensure that proper base pairing has taken place between the codon and the anticodon before the aminoacyl group is allowed to enter the peptidyl transferase center. Next, peptide-bond formation takes place through the transfer of the peptidyl chain from the tRNA in the P-site to the aminoacyl-tRNA in the A-site. Peptide-bond formation is catalyzed by RNA in the peptidyl transferase center of the large subunit, as well as the 2'-OH of the P-site tRNA. This ribozyme stimulates the nucleophilic attack of the amino group of the aminoacyl-tRNA in the A-site on the carbonyl group that attaches the growing polypeptide chain to the tRNA in the P-site. Finally, the ribosome translocates to the next vacant codon in a process that is driven by both the peptidyl transferase reaction and the action of the elongation factor EF-G (or its eukaryotic equivalent). As a result of translocation, the deacylated tRNA in the P-site is shifted into the E-site, where it exits the ribosome, and the peptidyl-tRNA in the A-site is shifted into the now vacant P-site. The adjacent codon in the mRNA is shifted into the now vacant A-site, which is poised to accept the delivery of a charged tRNA by EF-Tu.

Translation terminates when the ribosome encounters a stop codon, which is recognized by one of two class I release factors in prokaryotes and a single class I release factor in eukaryotes. The release factor triggers the hydrolysis of the polypeptide from the peptidyl-tRNA and hence the release of the completed polypeptide. In prokaryotes, the class II release factor, a ribosome recycling factor, and an initiation factor (IF3 in prokaryotes) complete termination by causing the release of the mRNA and the deacylated tRNAs and the dissociation of the ribosome into its large and small subunits. Translational termination in eukaryotes requires eRF3 to deliver eRF1 to the ribosome. The mechanism of ribosome recycling is unclear but is likely to involve eRF1 playing the role that RRF plays in prokaryotic cells. The ribosome cycle is now complete, and the small subunit is ready to commence a new cycle of polypeptide synthesis.

The expression of many genes is regulated at the level of translation initiation. In bacterial cells, this regulation generally occurs by inhibiting binding of the small subunit to the RBS. This inhibition can be mediated by either protein or RNA binding to mRNA sequences near the RBS. Global levels of eukaryotic translation are regulated by 4E-BP proteins, which bind to eIF4E and compete for its ability to bind eIF4G, and eIF2 $\alpha$  kinases, which inhibit the ability of eIF2 to exchange GDP for GTP. Regulation of the translation of specific eukaryotic mRNAs is sometimes mediated by small uORFs that limit access of the small subunit to a downstream ORF. Modified versions of both of these mechanisms are adapted to regulate specific genes.

Translation is also used by both bacteria and eukaryotic cells to monitor the integrity of mRNAs and eliminate mutant mRNAs and their protein products. mRNAs lacking stop codons result in the synthesis of proteins that are recognized by cellular proteases and degraded. Eukaryotic mRNAs with a premature stop codon, a stalled ribosome, or lacking a stop codon are detected, and the associated mRNA is degraded.

## BIBLIOGRAPHY

---

### Books

Mathews M.B., Sonenberg N., and Hershey J.W.B. 2007. *Translational control in biology and medicine*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

### tRNA and Aminoacyl-tRNA Synthetases

Arnez J.G. and Moras D. 1997. Structural and functional considerations of the aminoacylation reaction. *Trends Biochem. Sci.* **2**: 189–232.  
Ling J., Reynolds N., and Ibbá I. 2009. Aminoacyl-tRNA synthesis and translational quality control. *Annu. Rev. Microbiol.* **63**: 61–78.

### The Ribosome

Dunkle J.A., Wang L., Feldman M.B., Pulk A., Chen V.B., Kapral G.J., Noeske J., Richardson J.S., Blanchard S.C., and Cate J.H.D. 2011. Structures of the bacterial ribosome in classical and hybrid states of tRNA binding. *Science* **332**: 981–984.  
Frank J. and Gonzalez R.L. 2010. Structure and dynamics of a processive Brownian motor: The translating ribosome. *Annu. Rev. Biochem.* **79**: 381–412.  
Korostelev A., Trakhanov S., Laurberg M., and Noller H.F. 2006. Crystal structure of a 70S ribosome–tRNA complex reveals functional interactions and rearrangements. *Cell* **126**: 1065–1077.  
Moore P.B. and Steitz T.A. 2005. The ribosome revealed. *Trends Biochem. Sci.* **30**: 281–283.  
Poehlsgaard J. and Douthwaite S. 2005. The bacterial ribosome as a target for antibiotics. *Nat. Rev. Microbiol.* **3**: 870–881.  
Ramakrishnan V. 2002. Ribosome structure and the mechanism of translation. *Cell* **108**: 557–572.  
Rodnina M.V., Beringer M., and Wintermeyer W. 2006. How ribosomes make peptide bonds. *Trends Biochem. Sci.* **32**: 20–26.  
Selmer M., Dunham C.M., Murphy F.V., Weixlbaumer A., Petry S., Kelley A.C., Weir J.R., and Ramakrishnan V. 2006. Structure of the 70S ribosome complexed with mRNA and tRNA. *Science* **313**: 1935–1942.

### Translation

Balvay L., Soto Rifo R., Ricci E.P., Decimo D., and Ohlmann T. 2009. Structural and functional diversity of viral IRESes. *Biochim. Biophys. Acta* **1789**: 542–557.

## QUESTIONS

### MasteringBiology®

*For instructor-assigned tutorials and problems, go to MasteringBiology.*

For answers to even-numbered questions, see Appendix 2: Answers.

**Question 1.** Compare and contrast the features of a prokaryotic mRNA to a eukaryotic mRNA.

**Question 2.** Describe the significance of the sequence 5'-CCA-3' at the 3' terminus of every tRNA.

**Question 3.** Write the equations for the first and second steps of tRNA charging using the amino acid threonine and the threonyl-tRNA synthetase. Also write the overall equation.

**Question 4.** Write the equation for editing Ser-tRNA<sup>Thr</sup> by the threonyl-tRNA synthetase.

Broderson D.E. and Ramakrishnan V. 2003. Shapes can be seductive. *Nat. Struct. Biol.* **10**: 78–80.

Dever T.E. and Green R. 2012. The elongation, termination, and recycling phases of translation in eukaryotes. *Cold Spring Harb. Perspect. Biol.* **4**: a013706.

Hernández G. 2008. Was the initiation of translation in early eukaryotes IRES-driven? *Trends Biochem. Sci.* **33**: 58–64.

Laursen B.S., Sorenson H.P., Mortenson K.K., and Sperling-Peterson H.U. 2005. Initiation of protein synthesis in bacteria. *Microbiol. Mol. Biol. Rev.* **60**: 101–123.

Nilsson J. and Nissen P. 2005. Elongation factors on the ribosome. *Curr. Opin. Struct. Biol.* **15**: 349–354.

Nissen P., Kjeldgaard M., and Nyborg J. 2000. Macromolecular mimicry. *EMBO J.* **19**: 489–495.

Sonenberg N. and Hinnebusch A.G. 2009. Regulation of translation initiation in eukaryotes: Mechanisms and biological targets. *Cell* **136**: 731–745.

Weinger J.S., Parnell K.M., Forner S., Green R., and Strobel S.A. 2004. Substrate-assisted catalysis of peptide bond formation by the ribosome. *Nat. Struct. Mol. Biol.* **11**: 1101–1106.

### Regulation of Translation

Gebauer F. and Hentze M.W. 2004. Molecular mechanisms of translational control. *Nat. Rev. Mol. Cell Biol.* **5**: 827–835.

Ingolia N.T., Ghaemmaghami S., Newman J.R.S., and Weissman J.S. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**: 218–223.

Janssen B.D. and Hayes C.S. 2012. The tmRNA ribosome-rescue system. *Adv. Protein Chem. Struct. Biol.* **86**: 151–191.

Richter J.D. and Sonenberg N. 2006. Regulation of cap-dependent translation by eIF4E inhibitory proteins. *Nature* **433**: 477–480.

Shi W., Zhang X., Jiang X., Yuan H., Lee J.S., Barry C.E. 3rd, Wang H., Zhang W., and Zhang Y. 2011. Pyrazinamide inhibits trans-translation in *Mycobacterium tuberculosis*. *Science* **333**: 1630–1632.

van Hoof A. and Wagner E.J. 2011. A brief survey of mRNA surveillance. *Trends Biochem. Sci.* **36**: 585–592.

**Question 5.** Explain how the tyrosyl-tRNA synthetase distinguishes tyrosine from phenylalanine to avoid mischarging.

**Question 6.** Which tRNA synthetase would you expect to have an editing pocket to hydrolyze an aminoacyl-tRNA mischarged with glycine? Explain your choice.

**Question 7.** Multiple ribosomes can translate the same mRNA at the same time. Describe how this is advantageous for the cell.

**Question 8.** Describe one experiment that supports the statement that an rRNA and not a protein component of the ribosome catalyzes the peptidyl transferase reaction.

**Question 9.** Explain where the energy comes from for peptide bond formation.

**Question 10.** Explain how structural studies revealed how two different complexes, prokaryotic EF-G-GDP and EF-Tu-GTP-tRNA, interact with the A site of the decoding center at different points during elongation.

**Question 11.** Calculate the energetic cost of nucleoside triphosphates consumed during one round of elongation after initiation is completed. Describe how the nucleoside triphosphate is used.

**Question 12.** Explain a general mechanism for how antibiotics inhibit translation and how they specifically target bacterial cells.

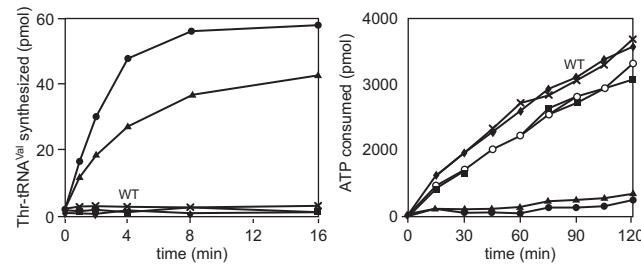
**Question 13.** Describe two mechanisms for prokaryotic cells to inhibit initiation of translation as a means to regulate translation.

**Question 14.** The valyl-tRNA synthetase (ValRS) normally adenylylates and transfers valine to the tRNA<sup>Val</sup>. At some frequency, the ValRS mischarges tRNA<sup>Val</sup> with threonine. Researchers wanted to determine what amino acids in the editing pocket of the ValRS are important for editing mischarged Thr-tRNA<sup>Val</sup>.

A. Why does the ValRS mischarge the tRNA<sup>Val</sup> with threonine rather than another amino acid?

To study the critical residues in the editing pocket, the researchers made amino acid substitutions at different posi-

tions within the editing pocket. They measured the amount of Thr-tRNA<sup>Val</sup> synthesized or ATP consumed in an in vitro charging reaction that included a mutant or wild-type ValRS. The mutant labeled F264A means that the ValRS includes an alanine at the 264th position rather than phenylalanine. The data are shown below.

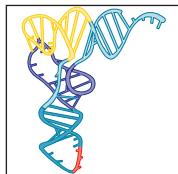


Enzyme activities of ValRS mutants. Wild type, ♦; D279A, ●; K270A, ▲; other mutants shown as ■, ○, × and \*. (Adapted, with permission, from Fukunaga R. and Yokoyama S. 2005. *J. Biol. Chem.* **280**: 29937–29945, Fig. 5 B,C, p. 29943. © The American Society for Biochemistry and Molecular Biology.)

- B. Given the data on the left, which mutant(s) likely have the most significant loss of editing function relative to the wild-type ValRS? Why?
- C. Explain why ATP consumption is higher for wild-type ValRS compared to the K270A ValRS.

*This page intentionally left blank*

CHAPTER 16



# The Genetic Code

AT THE VERY HEART OF THE CENTRAL DOGMA is the concept of information transfer from the linear sequence of the four-letter alphabet of the polynucleotide chain into the 20-amino-acid language of the polypeptide chain. As we have seen, the translation of genetic information into amino acid sequences takes place on ribosomes and is mediated by special adaptor molecules known as transfer RNAs (tRNAs). These tRNAs recognize groups of three consecutive nucleotides known as codons. With four possible nucleotides at each position, the total number of permutations of these triplets is 64 ( $4 \times 4 \times 4$ ), a value well in excess of the number of amino acids. Which of these triplet codons are responsible for specifying which amino acids, and what are the rules that govern their use? In this chapter, we discuss the nature and underlying logic of the genetic code, how the code was “cracked,” and the effect of mutations on the coding capacity of messenger RNA.

## THE CODE IS DEGENERATE

Table 16-1 lists all 64 permutations, with the left-hand column indicating the base at the 5' end of the triplet, the row across the top specifying the middle base, and the right-hand column identifying the base in the 3' position. One of the most striking features of the code is that 61 of the 64 possible triplets specify an amino acid, with the remaining three triplets being chain-terminating signals (see later discussion). This means that many amino acids are specified by more than one codon, a phenomenon called **degeneracy**. Codons specifying the same amino acid are **synonyms**. For example, UUU and UUC are synonyms for phenylalanine, whereas serine is encoded by the synonyms UCU, UCC, UCA, UCG, AGU, and AGC. In fact, when the first two nucleotides are identical, the third nucleotide can be either cytosine or uracil and the codon will still code for the same amino acid. Often, adenine and guanine are similarly interchangeable. However, not all degeneracy is based on equivalence of the first two nucleotides. Leucine, for example, is coded by UUA and UUG, as well as by CUU, CUC, CUA, and CUG (Fig. 16-1). Codon degeneracy, especially the frequent third-place equivalence of cytosine and uracil or guanine and adenine, explains how there can be great variation in the AT/GC ratios in the DNA of various organisms without correspondingly large changes in the relative proportion of amino acids in their proteins. (For example, the genomes of certain bacteria display vastly

## O U T L I N E

The Code Is Degenerate, 573



Three Rules Govern the Genetic Code, 582



Suppressor Mutations Can Reside in the Same or a Different Gene, 584



The Code Is Nearly Universal, 587

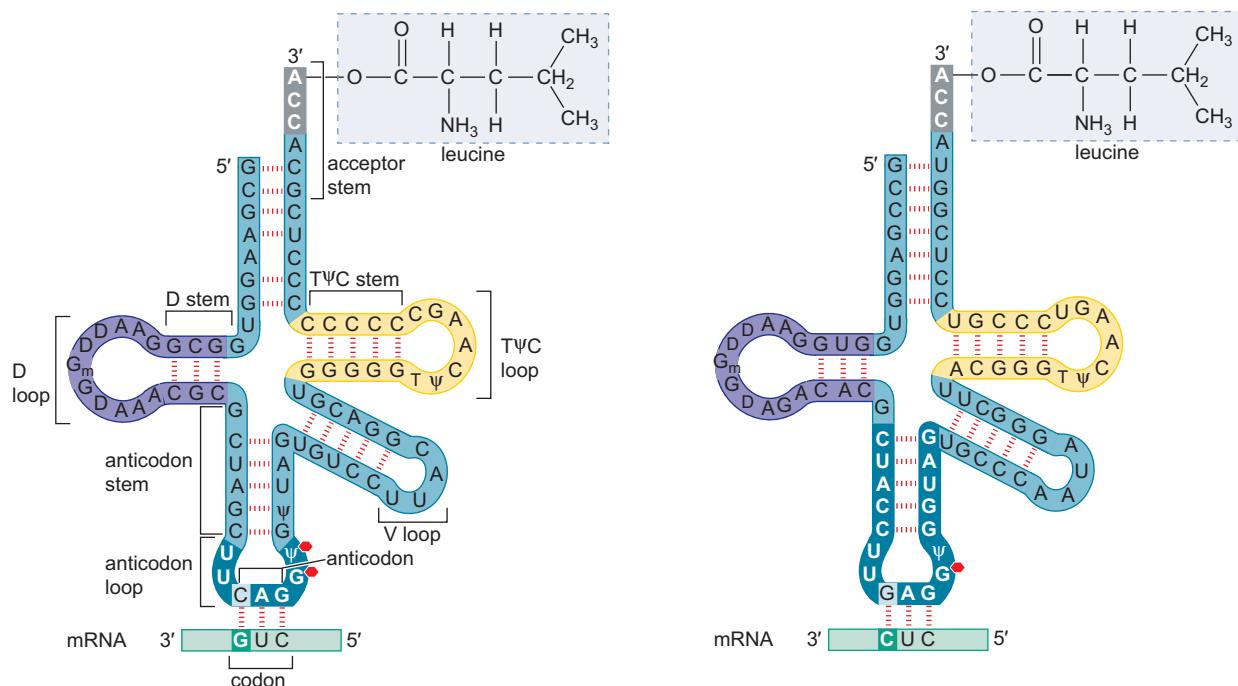


Visit Web Content for Structural Tutorials and Interactive Animations

TABLE 16-1 The Genetic Code

		second position										
		U	C	A	G							
first position (5' end)		U	UUU UUC UUA UUG	Phe Leu	UCU UCC UCA UCG	Ser	UAU UAC UAA* UAG*	Tyr stop stop	UGU UGC UGA* UGG	Cys stop Trp	third position (3' end)	
		C	CUU CUC CUA CUG	Leu	CCU CCC CCA CCG	Pro	CAU CAC CAA CAG	His Gln	CGU CGC CGA CGG	Arg	third position (3' end)	
		A	AUU AUC AUA AUG†	Ile Met	ACU ACC ACA ACG	Thr	AAU AAC AAA AAG	Asn Lys	AGU AGC AGA AGG	Ser Arg	third position (3' end)	
		G	GUU GUC GUA GUG	Val	GCU GCC GCA GCG	Ala	GAU GAC GAA GAG	Asp Glu	GGU GGC GGA GGG	Gly	third position (3' end)	

\* Chain-terminating or “nonsense” codons.

† Also used in bacteria to specify the initiator formyl-Met-tRNA<sup>fMet</sup>.

**FIGURE 16-1** Codon–anticodon pairing of two tRNA<sup>Leu</sup> molecules. Critical stem and loop regions of the tRNA structure are labeled (see Chapter 14). The red hexagons linked to the G (3' to the anticodon) denote methylation at the N1 positions of the base. Note that the codon is shown in a 3' to 5' orientation.

different AT/GC ratios and yet are closely related enough to encode proteins of highly similar amino acid sequences.)

### Perceiving Order in the Makeup of the Code

Inspection of the distribution of codons in the genetic code suggests that the code evolved in such a way as to minimize the deleterious effects of mutations. For instance, mutations in the first position of a codon will often give a similar (if not the same) amino acid. Furthermore, codons with pyrimidines in the second position specify mostly hydrophobic amino acids, whereas those with purines in the second position correspond mostly to polar amino acids (see Table 16-1 and Chapter 5, Fig. 5-4). Hence, because transitions (A:T to G:C or G:C to A:T substitutions) are the most common type of point mutations, a change in the second position of a codon will usually replace one amino acid with a very similar one. Finally, if a codon suffers a transition mutation in the third position, rarely will a different amino acid be specified. Even a transversion mutation in this position will have no consequence about half the time.

Another consistency noticeable in the code is that whenever the first two positions of a codon are both occupied by G or C, each of the four nucleotides in the third position specifies the same amino acid (such as proline, alanine, arginine, or glycine). On the other hand, whenever the first two positions of the codon are both occupied by A or U, the identity of the third nucleotide does make a difference. Since G:C base pairs are stronger than A:U base pairs, mismatches in pairing the third codon base are often tolerated if the first two positions make strong G:C base pairs. Thus, having all four nucleotides in the third position specify the same amino acid may have evolved as a safety mechanism to minimize errors in the reading of such codons.

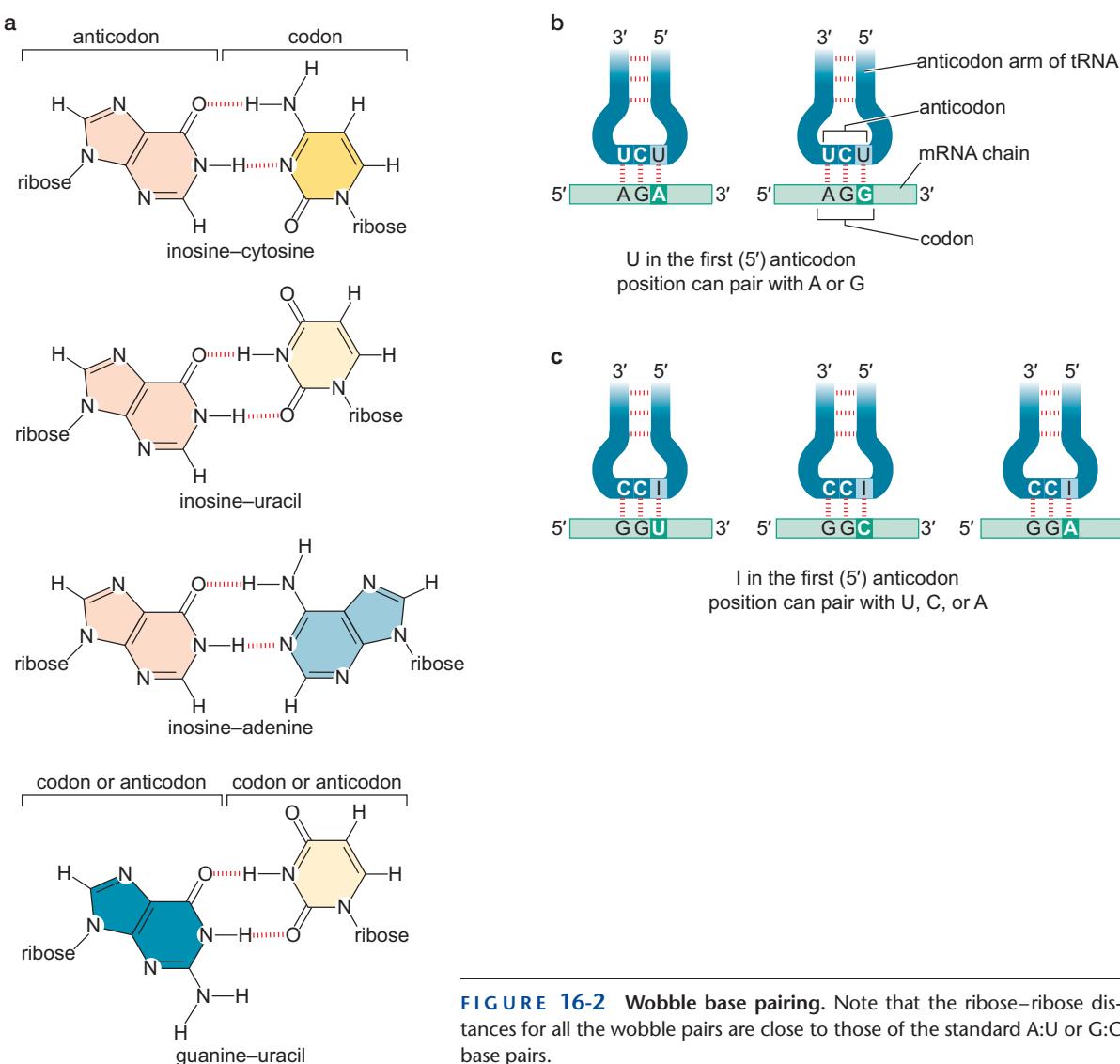
### Wobble in the Anticodon

It was first proposed that a specific tRNA anticodon would exist for every codon. If that were the case, at least 61 different tRNAs, possibly with an additional 3 for the chain-terminating codons, would be present. Evidence began to appear, however, that highly purified tRNA species of known sequence could recognize several different codons. Cases were also discovered in which an anticodon base was not one of the four regular ones, but a fifth base, inosine. Like all the other minor tRNA bases, inosine arises through enzymatic modification of a base present in an otherwise completed tRNA chain. The base from which it is derived is adenine, whose carbon 6 is deaminated to give the 6-keto group of inosine. (Inosine is actually a nucleoside composed of ribose and the base hypoxanthine, but it has come to be referred to as a base in common usage and we do so here.)

In 1966, Francis Crick devised the **wobble concept** to explain these observations. It states that the base at the 5' end of the anticodon is not as spatially confined as the other two, allowing it to form hydrogen bonds with any of several bases located at the 3' end of a codon. Not all combinations are possible, with pairing restricted to those shown in Table 16-2. For example, U at the wobble position can pair with either adenine or guanine, while I can pair with U, C, or A (Fig. 16-2). The pairings permitted by the wobble rules are those that give ribose–ribose distances close to that of the standard A:U or G:C base pairs. Purine–purine (with the exception of I:A pairs) or

**TABLE 16-2** Pairing Combinations with the Wobble Concept

Base in Anticodon	Base in Codon
G	U or C
C	G
A	U
U	A or G
I	A, U, or C



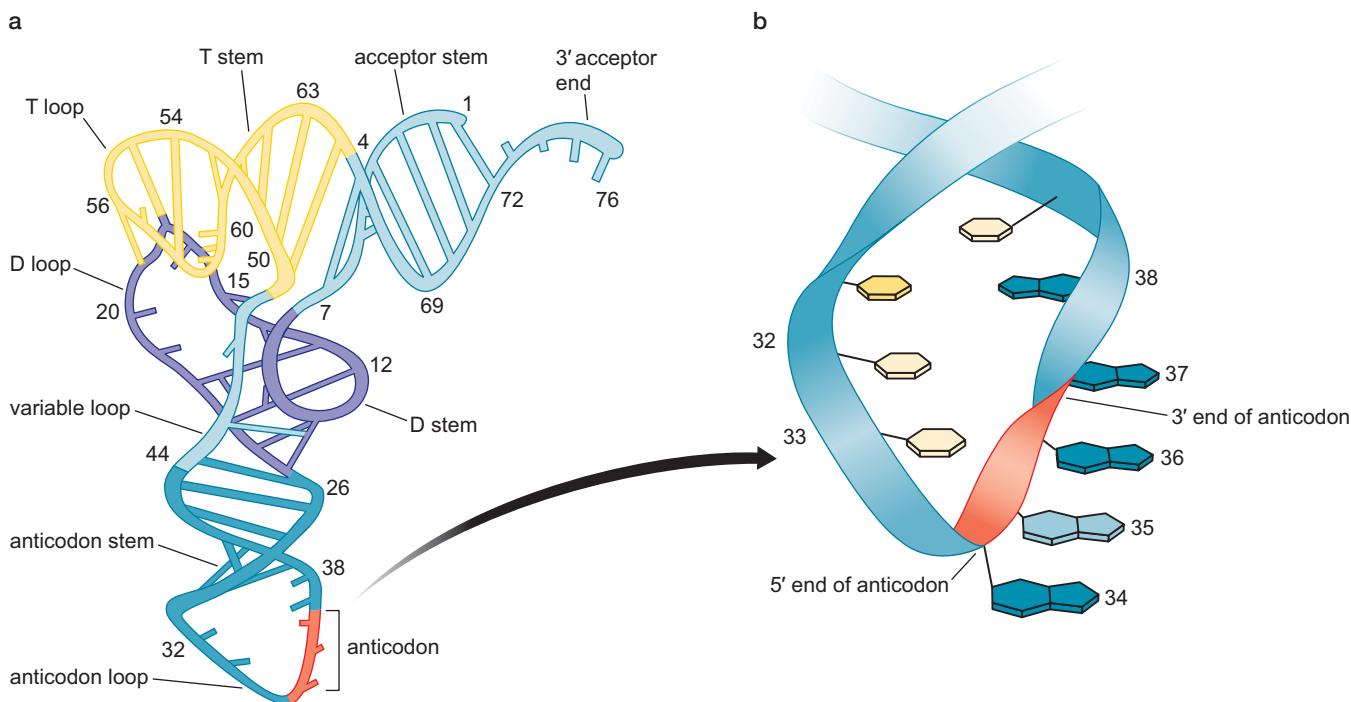
**FIGURE 16-2** Wobble base pairing. Note that the ribose–ribose distances for all the wobble pairs are close to those of the standard A:U or G:C base pairs.

pyrimidine–pyrimidine pairs would give ribose–ribose distances that are too long or too short, respectively.

The wobble rules do not permit any single tRNA molecule to recognize four different codons. Three codons can be recognized only when inosine occupies the first (5') position of the anticodon.

Almost all the evidence gathered since 1966 supports the wobble concept. For example, the concept correctly predicted that at least three tRNAs exist for the six serine codons (UCU, UCC, UCA, UCG, AGU, and AGC). The other two amino acids (leucine and arginine) that are encoded by six codons also have different tRNAs for the sets of codons that differ in the first or second position.

In the three-dimensional structure of tRNA, the three anticodon bases—as well as the two following (3') bases in the anticodon loop—all point in roughly the same direction, with their exact conformations largely determined by stacking interactions between the flat surfaces of the bases (Fig. 16-3). Thus, the first (5') anticodon base is at the end of the stack and is perhaps less restricted in its movements than the other two anticodon bases—hence, wobble in the third (3') position of the codon. By contrast, not only does the third (3') anticodon base appear in the middle of the stack, but



**FIGURE 16-3** Structure of yeast tRNA<sup>Phe</sup>. (a) A view of the L-shaped molecule based on X-ray diffraction data. (b) An enlargement of the anticodon loop. Bases in the anticodon (34–36) are shown in red. The anticodon and the following two bases (37 and 38) on the 3' side are partially stacked. It can be seen that the base at the 5' end of the anticodon is freer to wobble than is the fully stacked base at the 3' end of the anticodon. (Adapted from Kim S.-H. et al. 1974. Proc. Natl. Acad. Sci. 71: 4970.)

the adjacent base is always a bulky modified purine residue. Thus, restriction of its movements may explain why wobble is not seen in the first (5') position of the code.

### Three Codons Direct Chain Termination

As we have seen, three codons do not correspond to any amino acid. Instead, they signify chain termination. As we discussed in Chapter 14, these chain-terminating codons, UAA, UAG, and UGA, are read not by special tRNAs but by specific proteins known as release factors (RF1 and RF2 in bacteria and eRF1 in eukaryotes). Release factors enter the A site of the ribosome and trigger hydrolysis of the peptidyl-tRNA occupying the P site, resulting in the release of the newly synthesized protein.

### How the Code Was Cracked

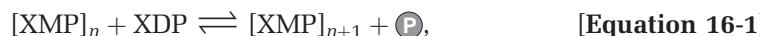
The assignment of amino acids to specific codons is one of the great achievements in the history of molecular biology (see Chapter 2 for an historic account). How were these assignments made? By 1960, the general outline of how messenger RNA (mRNA) participates in protein synthesis had been established. Nevertheless, there was little optimism that we would soon have a detailed understanding of the genetic code itself. It was believed that identification of the codons for a given amino acid would require exact knowledge of both the nucleotide sequences of a gene and the corresponding amino acid order in its protein product. At that time, the elucidation of

the amino acid sequence of a protein, although a laborious process, was already a very practical one. On the other hand, the then-current methods for determining DNA sequences were very primitive. Fortunately, this apparent roadblock did not hold up progress. In 1961, just one year after the discovery of mRNA, the use of artificial messenger RNAs and the availability of cell-free systems for carrying out protein synthesis began to make it possible to crack the code (see Chapter 2).

### Stimulation of Amino Acid Incorporation by Synthetic mRNAs

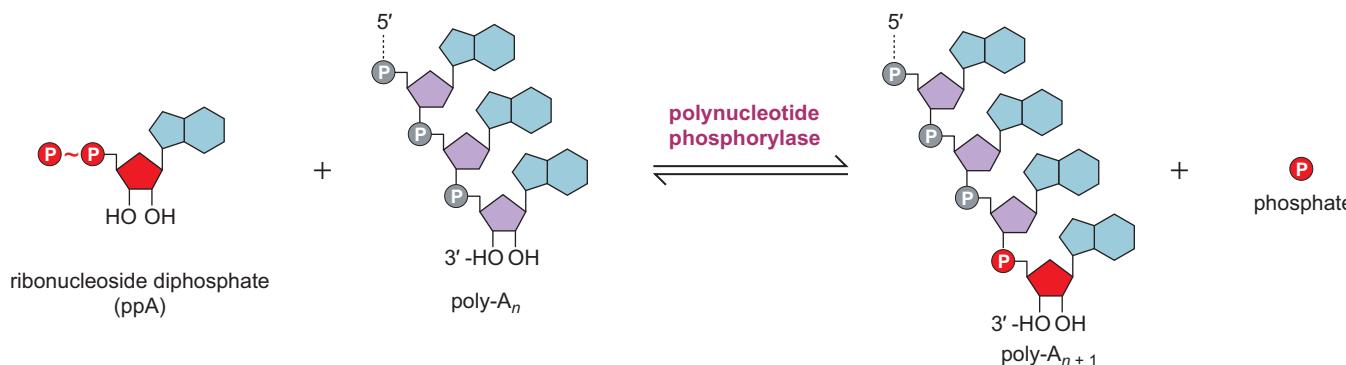
Biochemists found that extracts prepared from cells of *Escherichia coli* that were actively engaged in protein synthesis were capable of incorporating radioactively labeled amino acids into proteins. Protein synthesis in these extracts proceeded rapidly for several minutes and then gradually came to a stop. During this interval, there was a corresponding loss of mRNA owing to the action of degradative enzymes present in the extract. However, the addition of fresh mRNA to extracts that had stopped making protein caused an immediate resumption of synthesis.

The dependence of cell extracts on externally added mRNA provided an opportunity to elucidate the nature of the code using synthetic polyribonucleotides. These synthetic templates were created using the enzyme polynucleotide phosphorylase, which catalyzes the reaction



where X represents the base and  $[\text{XMP}]_n$  represents RNA of length  $n$  nucleotides.

Polynucleotide phosphorylase is normally responsible for breaking down RNA and under physiological conditions favors the degradation of RNA into nucleoside diphosphates. By use of high nucleoside diphosphate concentrations, however, this enzyme can be made to catalyze the formation of internucleotide  $3' \rightarrow 5'$  phosphodiester bonds and thus make RNA molecules (Fig. 16-4). No template DNA or RNA is required for RNA synthesis with this enzyme; the base composition of the synthetic product depends entirely on the ratio of the various ribonucleoside diphosphates added to the reaction mixture. For example, when only adenosine diphosphate is used, the resulting RNA contains only adenylyc acid and is thus called **polyadenylic acid** or **poly-A**. It is likewise possible to make poly-U, poly-C, and poly-G. Addition of two or more different diphosphates produces



**FIGURE 16-4** Polynucleotide phosphorylase reaction. The figure shows the reversible reactions of synthesis or degradation of polyadenylic acid catalyzed by the enzyme polynucleotide phosphorylase.

mixed copolymers such as poly-AU, poly-AC, poly-CU, and poly-AGCU. In all these mixed polymers, the base sequences are approximately random, with the nearest-neighbor frequencies determined solely by the relative concentrations of the reactants. For example, poly-AU molecules with two times as much A as U have sequences like UAAUAUAA-AUAAUAAAAAUUU....

### Poly-U Codes for Polyphenylalanine

Under the right conditions *in vitro*, almost all synthetic polymers will attach to ribosomes and function as templates. Luckily, high concentrations of magnesium were used in the early experiments. A high magnesium concentration circumvents the need for initiation factors and the special initiator fMet-tRNA, allowing chain initiation to take place without the proper signals in the mRNA. Poly-U was the first synthetic polyribonucleotide discovered to have mRNA activity. It selects phenylalanyl tRNA molecules exclusively, thereby forming a polypeptide chain containing only phenylalanine (polyphenylalanine). Thus, we know that a codon for phenylalanine is composed of a group of three uridylic acid residues, UUU. (That a codon has three nucleotides was known from genetic experiments, as indicated in Chapters 2 and 7, and later.) On the basis of analogous experiments with poly-C and poly-A, CCC was assigned as a proline codon and AAA as a lysine codon. Unfortunately, this type of experiment did not tell us what amino acid GGG specifies. The guanine residues in poly-G firmly hydrogen-bond to each other and form multistranded triple helices that do not bind to ribosomes.

### Mixed Copolymers Allowed Additional Codon Assignments

Poly-AC molecules can contain eight different codons, CCC, CCA, CAC, ACC, CAA, ACA, AAC, and AAA, whose proportions vary with the copolymer A/C ratio. When AC copolymers attach to ribosomes, they cause the incorporation of asparagine, glutamine, histidine, and threonine—in addition to the proline previously assigned to CCC codons and the lysine previously assigned to AAA codons. The proportions of these amino acids incorporated into polypeptide products depend on the A/C ratio. Thus, since an AC copolymer containing much more A than C promotes the incorporation of many more asparagine than histidine residues, we conclude that asparagine is coded by two As and one C and that histidine is coded by two Cs and one A (Table 16-3). Similar experiments with other copolymers allowed several additional assignments. Such experiments, however, did not reveal the order of the different nucleotides within a codon. There is no way of knowing from random copolymers whether the histidine codon containing two Cs and one A is ordered CCA, CAC, or ACC.

### Transfer RNA Binding to Defined Trinucleotide Codons

A direct way of ordering the nucleotides within some of the codons was developed in 1964. This method utilized the fact that even in the absence of all the factors required for protein synthesis, specific aminoacyl-tRNA molecules can bind to ribosome–mRNA complexes. For example, when poly-U is mixed with ribosomes, only phenylalanyl tRNA will attach. Correspondingly, poly-C promotes the binding of prolyl-tRNA. Most impor-

**TABLE 16-3** Amino Acid Incorporation into Proteins

Amino Acid	Observed Amino Acid Incorporation	Tentative Codon Assignments	Calculated Triplet Frequency				Sum of Calculated Triplet Frequencies
			3A	2A1C	1A1C	3C	
<b>Poly-AC (5:1)</b>							
Asparagine	24	2A1C		20			20
Glutamine	24	2A1C		20			20
Histidine	6	1A2C			4.0		4
Lysine	100	3A	100				100
Proline	7	1A2C, 3C			4.0	0.8	4.8
Threonine	26	2A1C, 1A2C		20	4.0		24
<b>Poly-AC (1:5)</b>							
Asparagine	5	2A1C		3.3			3.3
Glutamine	5	2A1C		3.3			3.3
Histidine	23	1A2C			16.7		16.7
Lysine	1	3A	0.7				0.7
Proline	100	1A2C, 3C			16.7	83.3	100
Threonine	21	2A1C, 1A2C		3.3	16.7		20

The amino acid incorporation into proteins was observed after adding random copolymers of A and C to a cell-free extract. The incorporation is given as a percentage of the maximal incorporation of a single amino acid. The copolymer ratio was then used to calculate the frequency with which a given codon would appear in the poly-nucleotide product. The relative frequencies of the codons are a function of the probability that a particular nucleotide will occur in a given position of a codon. For example, when the A/C ratio is 5:1, the ratio of AAA/AAC =  $5 \times 5 : 5 \times 1 = 125:25$ . If we thus assign to the 3A codon a frequency of 100, then the 2A and 1C codon is assigned a frequency of 20. By correlating the relative frequencies of amino acid incorporation with the calculated frequencies with which given codons appear, tentative codon assignments can be made.

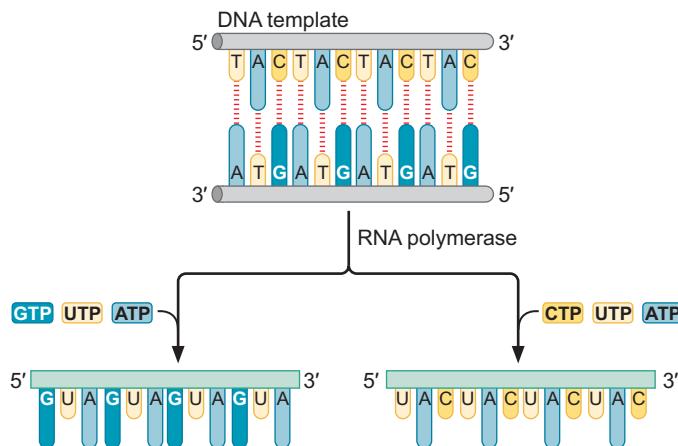
tantly, this specific binding does not demand the presence of long mRNA molecules. In fact, the binding of a trinucleotide to a ribosome is sufficient. The addition of the trinucleotide UUU results in phenylalanyl-tRNA attachment, whereas if AAA is added, lysyl-tRNA specifically binds to ribosomes. The discovery of this trinucleotide effect provided a relatively easy way of determining the order of nucleotides within many codons. For example, the trinucleotide 5'-GUU-3' promotes valyl-tRNA binding, 5'-UGU-3' stimulates cysteinyl-tRNA binding, and 5'-UUG-3' causes leucyl-tRNA binding (Table 16-4). Although all 64 possible trinucleotides were synthesized with the hope of definitely assigning the order of every codon, not all codons were determined in this way. Some trinucleotides bind to ribosomes much less efficiently than UUU or GUU, making it impossible to know whether they code for specific amino acids.

**TABLE 16-4** Binding of Aminoacyl-tRNA Molecules to Trinucleotide-Ribosome Complexes

Trinucleotide	AA-tRNA Bound					
5'-UUU-3'	UUC					Phenylalanine
UUA	UUG	CUU	CUC	CUA	CUG	Leucine
AAU	AUC	AUA				Isoleucine
AUG						Methionine
GUU	GUC	GUU	GUG	UCU <sup>a</sup>		Valine
UCU	UCC	UCA	UCG			Serine
CCU	CCC	CCA	CCG			Proline
AAA	AAG					Lysine
UGU	UGC					Cysteine
GAA	GAG					Glutamic acid

AA, aminoacyl.

<sup>a</sup>Note that this codon was misassigned by this method.



**FIGURE 16-5** Preparing oligoribonucleotides. Using a combination of organic synthesis and copying by DNA polymerase I, double-stranded DNA with simple repeating sequences can be generated. RNA polymerase will then synthesize long polyribonucleotides corresponding to one or the other DNA strand, depending on the choice of ribonucleoside triphosphate added to the reaction mixture.

### Codon Assignments from Repeating Copolymers

At the same time that the trinucleotide binding technique became available, organic chemical and enzymatic techniques were being used to prepare synthetic polyribonucleotides with known repeating sequences (Fig. 16-5). Ribosomes start protein synthesis at random points along these regular copolymers; yet they incorporate specific amino acids into polypeptides. For example, the repeating sequence CUCUCUCU... is the messenger for a regular polypeptide in which leucine and serine alternate. Similarly, UGUGUG... promotes the synthesis of a polypeptide containing two amino acids, cysteine and valine. And ACACAC... directs the synthesis of a polypeptide alternating threonine and histidine. The copolymer built up from repetition of the three-nucleotide sequence AAG (AAGAAGAAG) directs the synthesis of three types of polypeptides: polylysine, polyarginine, and polyglutamic acid. Poly-AUC behaves in the same way, acting as a template for polyisoleucine, polyserine, and polyhistididine (Table 16-5). Further codon assignments were obtained from repeating tetranucleotide sequences.

The sum of all these observations permitted the assignments of specific amino acids to 61 out of the possible 64 codons (see Table 16-1), with the remaining three chain-terminating codons, UAG, UAA, and UGA, not spec-

**TABLE 16-5** Assignment of Codons Using Repeating Copolymers Built from Two or Three Nucleotides

Copolymer	Codons Recognized	Amino Acids Incorporated or Polypeptide Made	Codon Assignment
(CU) <sub>n</sub>	CUC UCU CUC ...	Leucine	5'-CUC-3'
		Serine	UCU
(UG) <sub>n</sub>	UGU GUG UGU ...	Cysteine	UGU
		Valine	GUG
(AC) <sub>n</sub>	ACA CAC ACA ...	Threonine	ACA
		Histidine	CAC
(AG) <sub>n</sub>	AGA GAG AGA ...	Arginine	AGA
		Glutamine	GAG
(AUC) <sub>n</sub>	AUC AUC AUC ... UCA UCA UCA ... CAU CAU CAU ...	Polyisoleucine	AUC
		Polyserine	UCA
		Polyhistididine	CAU

ifying any amino acid. (Note, as discussed in the previous chapter, that in the special context of translation initiation in *E. coli*, AUG is used as a start codon to specify N-formyl methionine rather than its usual codon assignment of methionine.)

### THREE RULES GOVERN THE GENETIC CODE

---

The genetic code is subject to three rules that govern the arrangement and use of codons in messenger RNA. The first rule holds that codons are read in a 5' to 3' direction. Thus, in principle and as an example, the coding sequence for the dipeptide NH<sub>2</sub>-Thr-Arg-COOH could be written as 5'-ACGCGA-3' (where 5'-ACG-3' is a threonine codon and 5'-CGA-3' an arginine codon) or as 3'-GCAAGC-5' wherein the codons are written in the same order as before but oppositely to their original orientations. Because messenger RNA is translated in a 5' to 3' direction, however, only the former is the correct coding sequence; if the latter were translated in a 5' to 3' direction, then the resulting peptide would be NH<sub>2</sub>-Arg-Thr-COOH rather than NH<sub>2</sub>-Thr-Arg-COOH.

The second rule is that codons are nonoverlapping and the message contains no gaps. This means that successive codons are represented by adjacent trinucleotides in register. Thus, the coding sequence for the tripeptide NH<sub>2</sub>-Thr-Arg-Ser-COOH is represented by three contiguous and non-overlapping triplets in the sequence 5'-ACGCGAUCU-3'.

The final rule is that the message is translated in a fixed reading frame, which is set by the initiation codon. As you will recall from Chapter 14, translation starts at an initiation codon, which is located at the 5' end of the protein-coding sequence. Because codons are nonoverlapping and consist of three consecutive nucleotides, a stretch of nucleotides could be translated in principle in any of three reading frames. It is the initiation codon that dictates which of the three possible reading frames is used. Thus, for example, the sequence 5'...ACGACGACGACGACGACG...3' could be translated as a series of threonine codons (5'-ACG-3'), a series of arginine codons (5'-CGA-3'), or a series of asparagine codons (5'-GAC-3') depending on the frame of the upstream start codon.

### Three Kinds of Point Mutations Alter the Genetic Code

Now that we have considered the nature of the genetic code, it is instructive to revisit the issue of how the coding sequence of a gene is altered by point mutations (see Chapter 9). An alteration that changes a codon specific for one amino acid to a codon specific for another amino acid is called a **missense mutation**. As a consequence, a gene bearing a missense mutation produces a protein product in which a single amino acid has been substituted for another, as in the classic example of the human genetic disease sickle-cell anemia, in which glutamate 6 in the β-globin subunit of hemoglobin has been replaced with a valine.

A more drastic effect results from an alteration causing a change to a chain-termination codon, which is known as a **nonsense** or **stop mutation**. When a nonsense mutation arises in the middle of a genetic message, an incomplete polypeptide is released from the ribosome owing to premature chain termination. The size of the incomplete polypeptide chain depends on the location of the nonsense mutation. Mutations occurring near the beginning of a gene result in very short polypeptides, whereas mutations near the end produce polypeptide chains of almost normal length. As we

saw in Chapter 14, mRNAs that contain a premature stop codon are rapidly degraded in eukaryotic cells by a process known as nonsense-mediated mRNA decay.

The third kind of point mutation is a **frameshift mutation**. Frameshift mutations are insertions or deletions of one or a small number of base pairs that alter the reading frame. Consider a tandem repeat of the sequence GCU in a frame that would be read as a series of alanine codons (the codons are artificially set apart from each other by a gap for clarity but are, of course, contiguous in a real messenger RNA):

Ala    Ala    Ala    Ala    Ala    Ala    Ala  
5'-GCU    GCU    GCU    GCU    GCU    GCU    GCU-3'

Now imagine the insertion of an A in the message, thereby generating a serine codon (AGC) at the site of the insertion. The resulting frameshift causes triplets downstream of the insertion to be read as cysteines:

Ala    Ala    Ser    Cys    Cys    Cys    Cys  
5'-GCU    GCU    AGC    UGC    UGC    UGC    UGC-3'

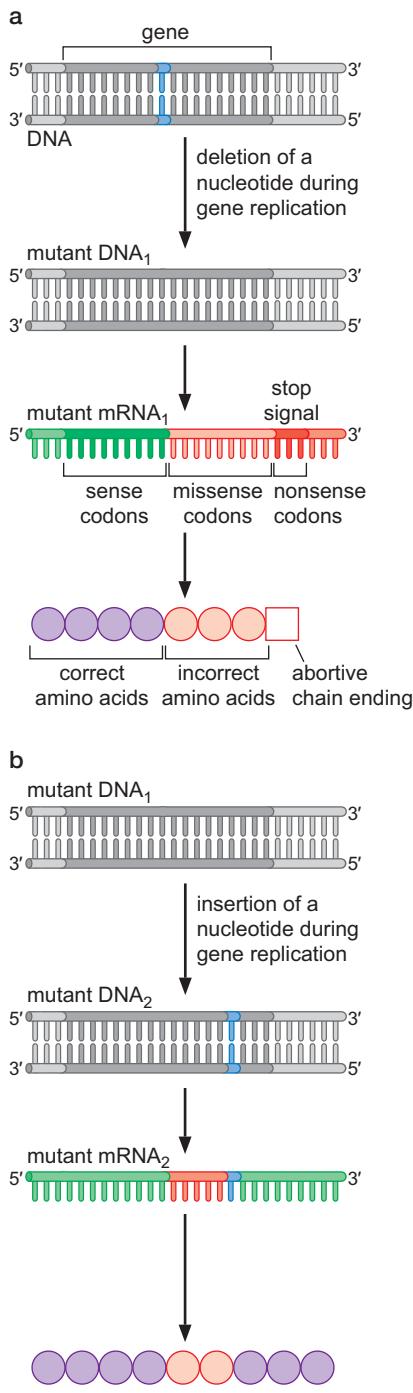
Thus, the insertion (or for that matter the deletion) of a single base drastically alters the coding capacity of the message not only at the site of the insertion but for the remainder of the messenger as well. Likewise, the insertion (or deletion) of two bases would have the effect of throwing the entire coding sequence, at and downstream of the insertions, into a different reading frame.

Finally, consider the instructive case of an insertion of three extra bases at nearby positions in a message. It is obvious that the stretch of message, at and between the three insertions, will be drastically altered. But because the code is read in units of three, mRNA downstream of the three inserted bases will be in its proper reading frame and, hence, completely unaltered:

Ala    Ala    Ser    Cys    Met    Leu    His    Ala    Ala    Ala  
5'-GCU    GCU    AGC    UGC    AUG    CUG    CAU    GCU    GCU    GCU-3'

### Genetic Proof That the Code Is Read in Units of Three

The preceding example is the logic of a classic experiment by Francis Crick, Sydney Brenner, and their coworkers, involving bacteriophage T4 that established that the code is read in units of three and did so purely on the basis of a genetic argument (i.e., without any biochemical or molecular evidence). Genetic crosses were carried out to create a mutant phage harboring three inferred single-base-pair insertion mutations at nearby positions in a single gene. Of course, the three insertions would have scrambled a short stretch of codons but the protein encoded by the gene in question (called *rII*) was able to tolerate the local alteration to its amino acid sequence. This finding indicated that the overall coding capacity of the gene had been chiefly left unaltered despite the presence of three mutations, each of which alone, or any two of which alone, would have drastically altered the reading frame of the gene's message (and rendered its protein product inactive). Because the gene could tolerate three insertions but not one or two (or, for that matter, four), the genetic code must be read in units of three. See Chapters 2 and 22 for a discussion of the historic figures who showed that the code is read in units of three and for a description of the role of bacteriophage T4 as a model system for elucidating the nature of the code.



**FIGURE 16-6** Suppression of frame-shift mutations. (a) A deletion in the nucleotide coding sequence can result in an incomplete, inactive polypeptide chain. (b) The effect of the deletion, shown in panel a, can be overcome by a second mutation, an insertion in the coding sequence. This insertion results in the production of a complete polypeptide chain having two amino acid replacements. Depending on the change in sequence, the protein may have partial or full activity.

## SUPPRESSOR MUTATIONS CAN RESIDE IN THE SAME OR A DIFFERENT GENE

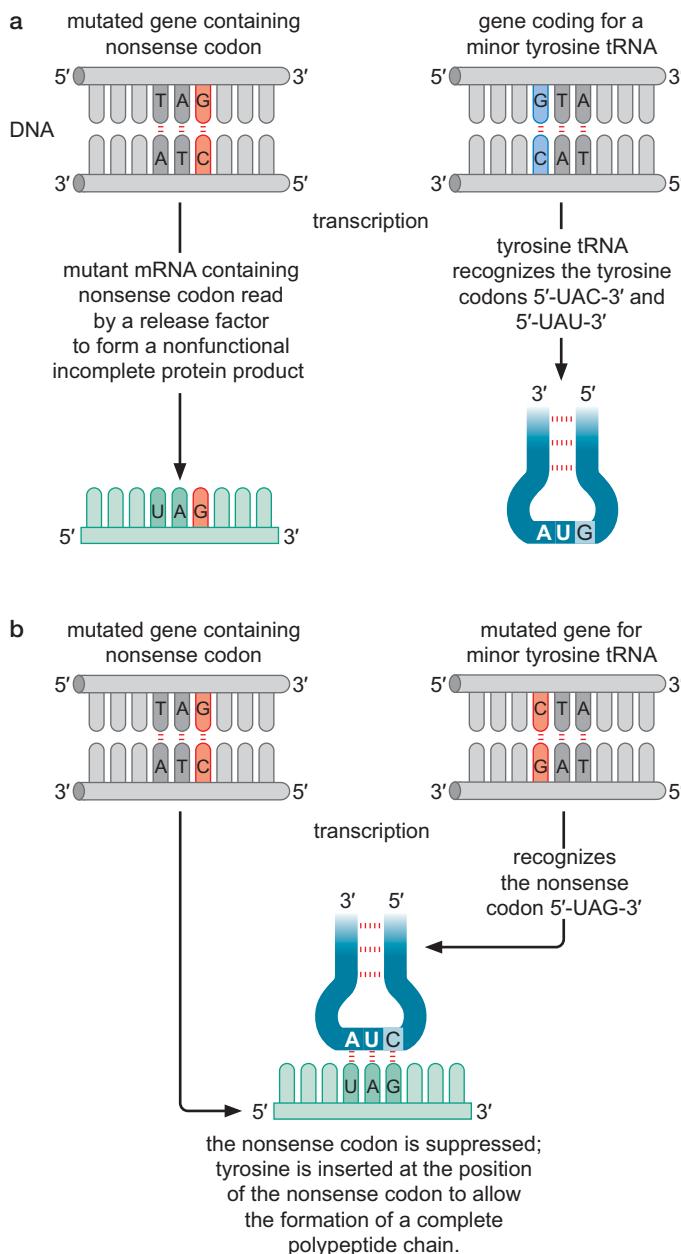
Often, the effects of harmful mutations can be reversed by a second genetic change. Some of these subsequent mutations are easy to understand, being simple **reverse (back) mutations**, which change an altered nucleotide sequence back to its original arrangement. More difficult to understand are the mutations occurring at different locations on the chromosome that suppress the change due to a mutation at site A by producing an additional genetic change at site B. Such **suppressor mutations** fall into two main categories: those occurring within the same gene as the original mutation, but at a different site in this gene (**intragenic suppression**) and those occurring in another gene (**intergenic suppression**). Genes that cause suppression of mutations in other genes are called **suppressor genes**. Both of the types of suppression that we are considering here work by causing the production of good (or partially good) copies of the protein made inactive by the original harmful mutation. For example, if the first mutation caused the production of inactive copies of one of the enzymes involved in making arginine, then the suppressor mutation allows arginine to be made by restoring the synthesis of some good copies of this same enzyme. However, the mechanisms by which intergenic and intragenic suppressor mutations cause the resumption of the synthesis of good proteins are completely different.

As an example of intragenic suppression, consider the case of a missense mutation. Its effect can sometimes be reversed through an additional missense mutation in the same gene. In such cases, the original loss of enzymatic activity is due to an altered three-dimensional configuration resulting from the presence of an incorrect amino acid in the encoded protein sequence. A second missense mutation in the same gene can bring back biological activity if it somehow restores the original configuration around the functional part of the molecule. Figure 16-6 shows another example of intragenic suppression, this time for the case of a frameshift mutation.

### Intergenic Suppression Involves Mutant tRNAs

Suppressor genes do not act by changing the nucleotide sequence of a mutant gene. Instead, they change the way the mRNA template is read. One of the best known examples of suppressor mutations are mutant tRNA genes that suppress the effects of nonsense mutations in protein-coding genes (but mutant tRNAs that suppress missense mutations and even frameshift mutations are also known). In *E. coli*, suppressor genes are known for each of the three stop codons. They act by reading a stop codon as if it were a signal for a specific amino acid. There are, for example, three well-characterized genes that suppress the UAG codon. One suppressor gene inserts serine, another glutamine, and a third tyrosine at the nonsense position. In each of the three UAG suppressor mutants, the anticodon of a tRNA species specific for one of these amino acids has been altered. For example, the tyrosine suppressor arises by a mutation within a tRNA<sup>Tyr</sup> gene that changes the anticodon from GUA (3'-AUG-5') to CUA (3'-AUC-5'), thereby enabling it to recognize UAG codons (Fig. 16-7). The serine and glutamine suppressor tRNAs also arise by single base changes in their anticodons.

The discovery that cells with nonsense suppressors contain mutationally altered tRNAs raised the question of how their codons corresponding to these tRNAs could continue to be read normally. In the case of the tyrosine UAG suppressor, the answer comes from the discovery that three separate genes code for tRNA<sup>Tyr</sup>. One codes for the major tRNA<sup>Tyr</sup> species, whereas



**FIGURE 16-7** Nonsense suppression. The figure shows how a minor tyrosine tRNA species acts to suppress the nonsense codon in mRNA.

the other two are duplicate genes coding for a species present in smaller amounts. One or the other of the two duplicate genes is always the site of the suppressor mutation. No such dilemma exists for UGA suppression, which is mediated by a mutant form of tRNA<sup>Trp</sup>; the suppressing tRNA<sup>Trp</sup> retains its capacity to read UGG (tryptophan) codons while also recognizing UGA stop codons. This is possible because the anticodon was changed from CCA (3'-ACC-5') in the wild type to UCA (3'-ACU-5') in the mutant tRNA<sup>Trp</sup>, and wobble rules, as we have seen, allow recognition of A or G in the 3' position of the codon by U in the 5' position of an anticodon.

### Nonsense Suppressors Also Read Normal Termination Signals

The act of nonsense suppression can be viewed as a competition between the suppressor tRNA and the release factor. When a stop codon comes into the

ribosomal A site, either readthrough or polypeptide chain termination will occur, depending on which arrives first. Suppression of UAG codons is efficient. In the presence of the suppressor tRNA, more than half of the chain-terminating signals are read as specific amino acid codons. *E. coli* can tolerate this misreading of the UAG stop codon because UAG is used infrequently as a chain-terminating codon at the end of open reading frames. In contrast, suppression of the UAA codon usually averages between 1% and 5% and mutant cells producing UAA-suppressing tRNAs grow poorly. This is expected from the fact that UAA is frequently used as a chain-terminating codon and its recognition by a suppressor tRNA would be expected to result in the production of many more aberrantly long polypeptides.

### Proving the Validity of the Genetic Code

The code was cracked, as we have seen, by means of biochemical methods involving the use of cell-free systems for carrying out protein synthesis. But molecular biologists are generally suspicious of a method that relies on *in vitro* analysis alone. So how do we know definitively that the code as depicted in Table 16-1 is true in living cells? Of course, in the modern era of large-scale DNA sequencing, in which the entire nucleotide sequences of the genomes of diverse organisms ranging from microbes to man have been determined, the genetic code has not only been validated but shown to be universal or nearly so (see later discussion). Nonetheless, a classic and instructive experiment in 1966 helped to validate the genetic code well before DNA sequencing was possible. The experiment was based on the construction by genetic recombination of a mutant gene of phage T4 that harbored a mutually suppressing pair of insertion and deletion mutations (similar to the example given in Fig. 16-6). The gene in question encoded a cell-wall-degrading enzyme called lysozyme, chosen because it is small, easy to purify, and its complete amino acid sequence was known. The experimental strategy was to compare the amino acid sequence of the doubly mutant protein with that of wild-type lysozyme.

When the amino acid sequences of the mutant (... NH<sub>2</sub>—Thr Lys **Val His His Leu Met** Ala Ala Lys—COOH ...) and wild type (... NH<sub>2</sub>—Thr Lys **Ser Pro Ser Leu Asn** Ala Ala Lys—COOH ...) were compared, they were found to differ by a stretch of five amino acids (highlighted in bold). This observation suggested that the insertion and deletion mutations had scrambled a short stretch of codons in the message of the mutant. Knowing the consequent effect of the scrambled codons on the amino acid sequence of the protein imposed important constraints on the nature of the genetic code. Specifically, if the genetic code as elucidated in biochemical experiments is valid, then it should be possible to identify a set of codons for the wild-type sequence Ser Pro Ser Leu Asn that, when properly aligned and bracketed with an insertion at one end and a deletion at the other, would specify the mutant amino acid sequence. Indeed, such a solution exists, which requires a deletion of a nucleotide at the 5' end of the coding sequence and the insertion of a nucleotide at the 3' end:

NH <sub>2</sub> —Lys	<b>Ser</b>	<b>Pro</b>	<b>Ser</b>	<b>Leu</b>	<b>Asn</b>	Ala—COOH
5'—AAA	AGU	CCA	UCA	CUU	AAU	GC—3'
5'—AAA	GUC	CAU	CAC	UUA	AUG	GC—3'
NH <sub>2</sub> —Lys	<b>Val</b>	<b>His</b>	<b>His</b>	<b>Leu</b>	<b>Met</b>	Ala—COOH

As you can see, the solution verifies several codon assignments and demonstrates that more than one synonymous codon is used to specify the same

amino acid in vivo (e.g., 5'-CAU-3' and 5'-CAC-3' for histidine). Lastly, and importantly, you should be able to convince yourself from the solution that translation proceeds in a 5' to 3' direction. (Hint: see if you can account for the two amino acid sequences in their proper NH<sub>2</sub> to COOH order when you align each of the codons in your solution in a 3' to 5' orientation.)

## THE CODE IS NEARLY UNIVERSAL

---

The results of large-scale sequencing of genomes have largely confirmed the expected universality of the genetic code. The universality of the code has had a huge impact on our understanding of evolution as it made it possible to directly compare protein-coding sequences among all organisms for which a genome sequence is available. As we shall see in Chapter 21, powerful computer programs are available that can search for and identify similarities among predicted coding sequences from a wide range of organisms. The universality of the code also helped to create the field of genetic engineering by making it possible to express cloned copies of genes encoding useful protein products in surrogate host organisms, such as the production of human insulin in bacteria (see Chapter 21).

To understand the conservative nature of the code, consider what might happen if a mutation changed the genetic code. Such a mutation might, for example, alter the sequence of the serine tRNA molecule of the class that corresponds to UCU, causing them to recognize UUU sequences instead. This would be a lethal mutation in haploid cells containing only one gene directing the production of tRNA<sup>Ser</sup>, for serine would not be inserted into many of its normal positions in proteins. Even if there were more than one gene for tRNA<sup>Ser</sup> (as in a diploid cell), this type of mutation would still be lethal since it would cause the simultaneous replacement of many phenylalanine residues by serine in cell proteins.

In view of what we have just said, it was completely unexpected to find that in certain subcellular organelles, the genetic code is in fact slightly different from the standard code. This realization came during the elucidation of the entire DNA sequence of the 16,569-bp human mitochondrial genome but is observed for mitochondria in yeast, the fruit fly, and higher plants. Sequences of the regions known to specify proteins have revealed the following differences between the standard and mitochondrial genetic codes (Table 16-6).

- UGA is not a stop signal but codes for tryptophan. Hence, the anticodon of mitochondrial tRNA<sup>Trp</sup> recognizes both UGG and UGA, as if obeying the traditional wobble rules.
- Internal methionine is encoded by both AUG and AUA.
- In mammalian mitochondria, AGA and AGG are not arginine codons (of which there are six in the “universal” code) but specify chain termination. Thus, there are four stop codons (UAA, UAG, AGA, and AGG) in the mammalian mitochondrial code.
- In fruit fly mitochondria, AGA and AGG are also not arginine codons but specify serine.

Perhaps not surprisingly, mitochondrial tRNAs are likewise unusual with respect to the rules by which they decode mitochondrial messages. Only 22 tRNAs are present in mammalian mitochondria, whereas a minimum of 32 tRNA molecules are required to decode the “universal” code according to the wobble rules. Consequently, when an amino acid is

**TABLE 16-6** Genetic Code of Mammalian Mitochondria

		second position					
		U	C	A	G		
first position (5' end)	U	UUU Phe (GAA) <sup>†</sup> UUC UUA Leu (UAA) UUG	UCU Ser (UGA) UCC UCA UCG	UAU Tyr (GUA) UAC UAA stop UAG stop	UGU Cys (GCA) UGC UGA Trp (UCA) UGG	U	C A G
	C	CUU Leu (UAG) CUC CUA CUG	CCU Pro (UGG) CCC CCA CCG	CAU His (GUG) CAC CAA Gln (UUG) CAG	CGU Arg (UCG) CGC CGA CGG	U	C A G
	A	AUU Ile (GAU) AUC AUA Met (CAU) <sup>#</sup> AUG	ACU Thr (UGU) ACC ACA ACG	AAU Asn (GUU) AAC AAA Lys (UUU) AAG	AGU Ser (GCU) AGC AGA stop AGG stop	U	C A G
	G	GUU Val (UAC) GUC GUA GUG	GCU Ala (UGC) GCC GCA GCG	GAU Asp (GUC) GAC GAA Glu (UUC) GAG	GGU Gly (UCC) GGC GGA GGG	U	C A G
							third position (3' end)

\* Differences between the mitochondrial and “universal” genetic code (Table 16-1) are shown by green shading.

† Each group of codons is shaded in gray and is read by a single tRNA whose anticodon, written 5' → 3', is in parentheses. Each four-codon group is read by a tRNA having a U in the first (5') position of the anticodon. Two-codon groups with codons ending in either U/C or A/G are read with GU wobble by tRNAs, with G or U, respectively, in the first position of the anticodon. The anticodons often contain modified bases.

# Note that the C in the first anticodon position engages in unusual pairing.

specified by four codons (with the same first and second positions), only a single mitochondrial tRNA is involved. (Recall that a minimum of two tRNAs would be required by nonmitochondrial systems.) Such mitochondrial tRNAs all have in the 5' (wobble) position of their anticodons a U residue, which is able to engage in pairing with any of the four nucleotides in the third codon position. In cases where purines in the third position of the codon correspond to different amino acids from pyrimidines in that position, a modified U in the first position of the anticodon of the mitochondrial tRNA restricts wobble to pairing with the two purines only.

Exceptions to the “universal” code are not limited to mitochondria but are also found in several prokaryotic genomes and in the nuclear genomes of certain eukaryotes. The bacterium *Mycoplasma capricolum* uses UGA as a tryptophan codon rather than a chain-termination codon. Likewise, some unicellular protozoa use UAA and UAG, which are stop codons in the “universal” code, as glutamine codons. Finally, a codon (CUG) for one amino acid (leucine) in the “universal” code has become a codon for another amino acid (serine) in the yeast *Candida*.

We have just seen that variations in the genetic code may occur that are peculiar to certain organelles and organisms. But differences in the code also may come about with the introduction of novel amino acids into protein sequences that serve, in fact, to expand the code. In Box 16-1 we consider

## ► ADVANCED CONCEPTS

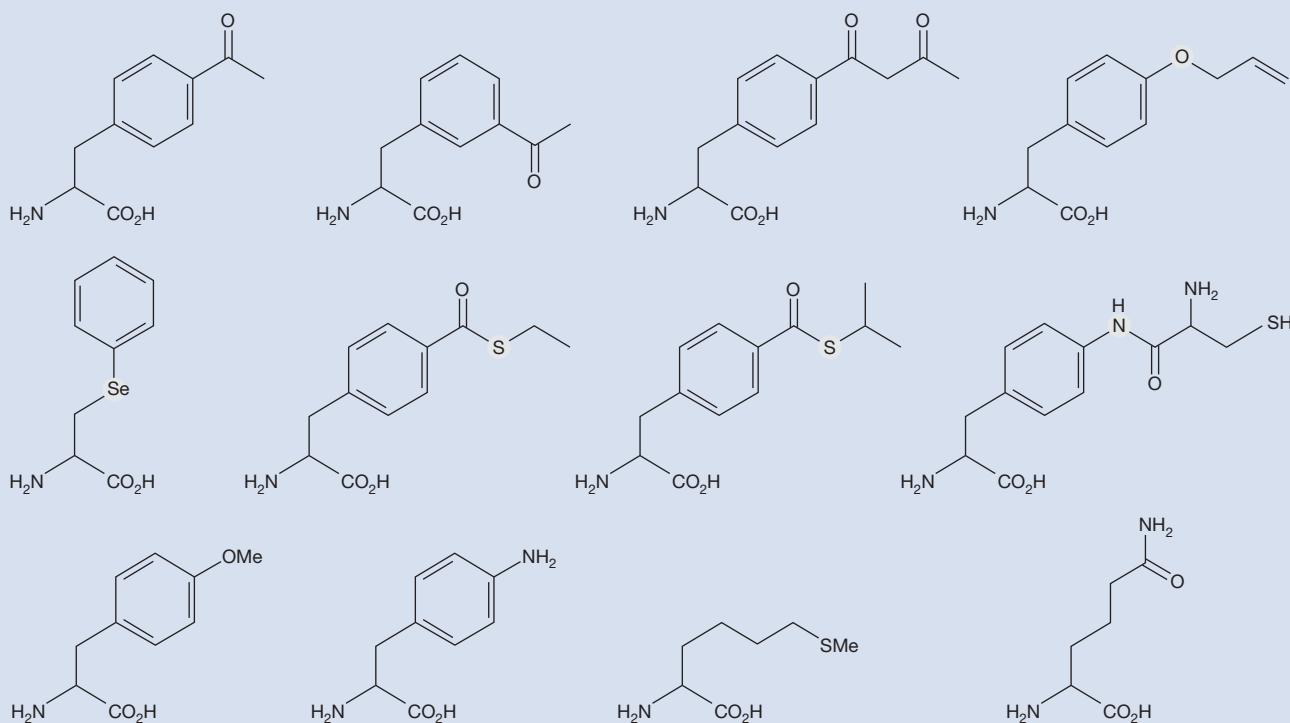
**Box 16-1** Expanding the Genetic Code

As we have seen, 61 of the 64 codons in the genetic code specify the 20 amino acids most commonly found in proteins. Some proteins, however, contain unusual amino acids that are not specified by the code. For example, collagen contains the amino acid hydroxyproline, which is created by hydroxylation of proline after it has been incorporated into the polypeptide chain. And some proteins contain selenocysteine, which is generated on a specialized tRNA that is charged with serine. The serine is converted to selenocysteine by enzymatic incorporation of selenium. The specialized selenocysteine-charged tRNA enters the ribosome at a special UGA codon that is flanked with a sequence in the mRNA that is recognized by a dedicated translation elongation factor. The elongation factor introduces the selenocysteinyl-charged tRNA into the A site where the metal-containing amino acid is incorporated into the growing polypeptide chain. These and other cases of modified amino acids involve specialized mechanisms for introducing altered amino acids into proteins without violating the (near) universality of the genetic code. But what if through genetic engineering we could expand the genetic code to specify unnatural amino acids? Could we do so in a manner that allowed us to incorporate tailor-made amino acids at particular sites in proteins and thereby generate novel proteins and even whole organisms with useful properties?

The convergence of two lines of research has brought these possibilities into the realm of reality. In one, a tRNA has been

created that recognizes UAG, but that is not a substrate for any of the existing amino acyl tRNA synthetases in *E. coli*. Instead, cognate synthetases have been generated by an iterative evolution strategy that recognize the unique tRNA and charge it with a particular artificial amino acid. In this way, a series of synthetases have been evolved that recognize and charge the cognate tRNA with one of a variety of unnatural amino acids with useful features, such as heavy metal binding sites (for facilitating X-ray diffraction studies), fluorescent moieties (for fluorescence microscopy), photo cross-linking sites, or chemically reactive groups (Box 16-1 Fig. 1). *E. coli* strains producing the UAG-recognizing tRNA and one of these novel synthetases are capable of taking up a cognate unnatural amino acid from the growth medium and incorporating it into proteins at UAG sites in the mRNA. For example, if we want to introduce a photo cross-linking site at a particular position in a protein of interest, we introduce a TAG codon into the coding sequence for that protein and express the TAG-containing gene in *E. coli* cells engineered to introduce the unnatural amino acid at UAG codons.

Meanwhile, in a second line of research, an *E. coli* strain is being created in which all 314 TAG (UAG) stop codons in the *E. coli* genome are being replaced with TAA (UAA) stop codons. This engineering feat is being achieved in two stages. In the first stage a multiplex approach known as MAGE (multiplex automated genome engineering) has been used to create 32 strains in which different subsets of TAG codons have been replaced



**BOX 16-1 FIGURE 1** Examples of unnatural amino acids developed for incorporation by an expanded genetic code. (Adapted from <http://schultz.scripps.edu/research.php>, courtesy Peter Schultz.)

**Box 16-1** (Continued)

with the synonymous stop codon. Next, a strategy based on hierarchical genetic conjugation known as CAGE (conjugative assembly genome engineering) is being employed to merge all 314 codon substitutions into a single strain. This strategy promises to culminate in an *E. coli* strain in which the codon TAG has been freed up for the incorporation of tailor-made amino

acids. In the future, strategies such as these could be applied to other microorganisms, such as yeast. Unshackled from the constraint of 20 natural amino acids, *E. coli* and yeast engineered to have a 21st amino acid–specifying codon might have a greater capacity to evolve useful traits under controlled laboratory conditions than their unmodified ancestors.

how these amino acids may arise, through natural or unnatural means, and how they become incorporated into proteins through impressive feats of engineering.

## SUMMARY

In the “universal” genetic code used by every organism from bacteria to humans, 61 codons signify specific amino acids; the remaining three are chain-termination codons. The code is highly degenerate, with several codons (synonyms) usually corresponding to a single amino acid. A given tRNA can sometimes specifically recognize several codons. This ability arises from wobble in the base at the 5' end of the anticodon. The stop codons UAA, UAG, and UGA are read by specific proteins, not specialized tRNA molecules.

The genetic code is subject to three principal rules. Codons are read in a 5' to 3' direction, codons are nonoverlapping and the message contains no gaps, and the message is translated in a fixed reading frame, which is set by the initiation codon.

The genetic code was cracked through the study of protein synthesis in cell-free extracts. Addition of new mRNA to an extract depleted of its original messenger component results in the production of new proteins whose amino acid sequences are determined by the externally added mRNA. The first (and probably most important) step in cracking the genetic code occurred when the synthetic polyribonucleotide poly-U was found to code specifically for polyphenylalanine.

Use of other synthetic polyribonucleotides, both homogeneous (poly-C, and so on) and mixed (poly-AU, and so on), then allowed assignment of codons for the various amino acids. Determination of the exact order of nucleotides in codons subsequently came from a study of specific trinucleotide–tRNA–ribosome interactions and the use of regular copolymers as messengers.

Point mutations that alter the code are missense mutations, which change the codon for one amino acid into the codon for another amino acid; nonsense mutations, which cause protein synthesis to terminate prematurely; and frame-shift mutations, which alter the reading frame of the message. In some cases the effects of missense, nonsense, and frame-shift mutations can be partially suppressed by extragenic suppressors. For example, mutant tRNAs read stop codons generated by nonsense mutations as if they were codons for a specific amino acid.

A slightly different genetic code is utilized in mitochondria and in the principal genomes of certain prokaryotes and protozoa, such as the use of UGA, a stop codon in the “universal code,” as a tryptophan codon.

## BIBLIOGRAPHY

### Books

- Celis J.E. and Smith J.D., eds. 1979. *Nonsense mutations and tRNA suppressors*. Academic Press, New York.
- Clark B. and Petersen H., eds. 1984. Gene expression: The translational step and its control. *Alfred Benzon Symposium*, vol. 19. Copenhagen, Munksgaard.
- Cold Spring Harbor Symposia on Quantitative Biology. 1966. Volume 31: *The genetic code*. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
- Söll D.G., Abelson J.N., and Schimmel P.R., eds. 1980. *Transfer RNA: Biological aspects*. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
- Ycas M. 1969. *The biological code*. Wiley (Interscience), New York.

### Features of the Genetic Code

- Crick F.H.C. 1966. Codon–anticodon pairing: The wobble hypothesis. *J. Mol. Biol.* **19**: 548–555.
- Kohli J. and Grosjean H. 1981. Usage of the three termination codons: Compilation and analysis of the known eukaryotic and prokaryotic translation termination sequences. *Mol. Gen. Genet.* **182**: 430–439.
- Lagerkvist U. 1981. Unorthodox codon reading and the evolution of the genetic code. *Cell* **23**: 305–306.

### How the Code Was Cracked

- Crick F.H.C. 1963. The recent excitement in the coding problem. *Prog. Nucleic Acid Res.* **1**: 164.

- Khorana H.G. 1968. *Polynucleotide synthesis and the genetic code. Harvey Lecture Series 1966–1967.* Vol. 62. Academic Press, New York.
- Nirenberg M. and Leder P. 1964. The effect of trinucleotides upon the binding of sRNA to ribosomes. *Science* **145**: 1399–1407.
- Speyer J.F., Lengyel P., Basilio C., Wahba A.J., Gardner R.S., and Ochoa S. 1963. Synthetic polynucleotides and the amino acid code. *Cold Spring Harbor Symp. Quant. Biol.* **28**: 559–568.

### Three Rules of the Genetic Code

- Brenner S., Stretton A.O.W., and Kaplan S. 1965. Genetic code: The non-sense triplets for chain termination and their suppression. *Nature* **206**: 994–998.
- Crick F.H.C., Barnett L., Brenner S., and Watts-Tobin R.J. 1961. General nature of the genetic code for proteins. *Nature* **192**: 1227–1232.
- Garen A. 1968. Sense and nonsense in the genetic code. *Science* **160**: 149–159.
- Terzaghi E., Okada Y., Streisinger G., Emrich J., Inouye M., and Tsugita A. 1966. Change of a sequence of amino acids in phage T4 lysozyme by acridine-induced mutations. *Proc. Natl. Acad. Sci.* **56**: 500–507.

## QUESTIONS

### MasteringBiology®

For instructor-assigned tutorials and problems, go to *MasteringBiology*.

For answers to even-numbered questions, see Appendix 2: Answers.

**Question 1.** Name two amino acids that do not have codon degeneracy. Explain why.

**Question 2.** Explain the cellular advantage for the codon 5'-AAG-3' to code lysine and the codon 5'-AGG-3' to code for arginine.

**Question 3.** Consider the mRNA codon 5'-ACU-3'. This codon codes for what amino acid? What DNA sequence encodes this codon? Give the sequence of the tRNA anticodon that recognizes this codon.

**Question 4.** Following the 5' → 3' convention of writing nucleotide sequences, indicate (yes or no) whether each of the following mRNA codons can be recognized by the tRNA anticodon ICG.

- \_\_\_\_\_ A. UGC
- \_\_\_\_\_ B. CGA
- \_\_\_\_\_ C. UGA
- \_\_\_\_\_ D. CGU
- \_\_\_\_\_ E. GCG

**Question 5.** Considering the early experiments performed by biochemists to crack the genetic code using RNA sequences like poly-U, what key condition allowed the ribosome to translate the polymer sequences *in vitro*?

**Question 6.** You want to test if the codon 5'-GUG-3' codes for valine. What repeated dinucleotide RNA sequence would you use to obtain support that 5'-GUG-3' codes for valine? When the ribosome translates the repeated dinucleotide sequence,

### Suppression

- Buckingham R.H. and Kurland C.G. 1980. Interactions between UGA-suppressor tRNA<sup>P</sup> and the ribosome: Mechanisms of tRNA selection. In *Transfer R.N.A.: Biological aspects* (ed. D. Söll, et al.), pp. 421–426. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
- Ozeki H., Inokuchi H., Yamao F., Kodaira M., Sakano H., Ikemura T., and Shimura Y. 1980. Genetics of nonsense suppressor of tRNAs in *Escherichia coli*. In *Transfer R.N.A.: Biological aspects* (ed. D. Söll, et al.), pp. 341–349. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
- Steege D.A. and Söll D.G. 1979. Suppression. In *Biological regulation and development I* (ed. R.F. Goldberger), pp. 433–486. Plenum, New York.

### Expanding the Genetic Code

- Isaacs F.J., Carr P.A., Wang H.H., Lajoie M.J., Sterling B., Kraal L., Tolonen A.C., Gianoulis T.A., Goodman D.B., Reppas N.B., et al. 2011. Precise manipulation of chromosomes *in vivo* enables genome-wide codon replacement. *Science* **333**: 348–353.
- Noren C.J., Anthony-Cahill S.J., Griffith M.C., and Schultz P.G. 1989. A general method for site-specific incorporation of unnatural amino acids into proteins. *Science* **192**: 1227–1232.

what other amino acid would you expect to find in the polypeptide sequence?

**Question 7.** You are given the following DNA sequence located in the middle of a gene 5'-ACCGTTTCGGCTAGG-3' from *E. coli*. This strand represents the coding strand. What three rules of the genetic code must you know before you can correctly translate this sequence to a polypeptide?

**Question 8.** You are given the following DNA sequence located in the middle of a gene 5'-ACCGTTTCGGCTAGG-3' from *E. coli*. This strand represents the coding strand. Give the three possible polypeptides that this sequence encodes.

**Question 9.** Shown below is a portion of a wild-type DNA sequence that encodes the last amino acids of a protein that is 270 amino acids long. The first three bolded base pairs indicate the frame and include the coding region.

5'...**GCTAAGTATTGCTCAAGATTAGGATGATAAATAACTGG**3'  
3'...**CGATT**CATAACGAGTTCTAACCTACTATTATTGACC5'

- A. Which strand is the template strand for transcription of this gene? Briefly explain how you know.
- B. An insertion of one base pair causes the protein to decrease in length by seven amino acids. With respect to the sequence given above, where does this insertion occur?
- C. A change of one base pair leads to the protein increasing in length by one amino acid. With respect to the sequence given above, which base pair would you change, and what would you change this base pair to for the protein to increase in length by one amino acid?

**Question 10.** With respect to the wild-type sequence below, identify the mutation in each altered sequence as one of the following types: missense mutation, nonsense mutation, frameshift mutation, reverse mutation, intragenic suppressor mutation, or intergenic suppressor mutation. For each altered sequence, more than one answer is possible. This DNA sequence encodes the last amino acids of a protein that is 270 amino acids long. The first three bolded base pairs indicate the frame and include the coding region. Inserted nucleotides are italicized.

Wild type:

5'...**GCTAAGTATTGCTCAAGATTAGGATGATAAATAACTGG**3'  
3'...CGATT**CATAACGAGTTGGCTAACCTACTATTTATTGACC**5'

A. Altered sequence 1:

5'...**GCTAAGTATTGCTCAACCGATTAGGATGATAAATAACTGG**3'

3'...CGATT**CATAACGAGTTGGCTAACCTACTATTTATTGACC**5'

B. Altered sequence 2 – changing altered sequence 1:

5'...**GCTAAGTATTGCTCACCGATTAGGATGATAAATAACTGG**3'

3'...CGATT**CATAACGAGTGGCTAACCTACTATTTATTGACC**5'

C. Altered sequence 3 – changing altered sequence 1:

5'...**GCTAAGTATTGCTCCGATTAGGATGATAAATAACTGG**3'

3'...CGATT**CATAACGAGGGCTAACCTACTATTTATTGACC**5'

**Question 11.** Researchers found that a single amino acid change of cysteine to tryptophan in a transmembrane protein causes retinal degeneration. Give the relevant sequence of the mRNA that encodes this substitution. Give the wild-type mRNA sequence(s) and the mutated sequence. Is the amino acid substitution caused by a transition or transversion mutation in the DNA?

**Question 12.** Describe the universality of the genetic code. How are there exceptions to the universality?

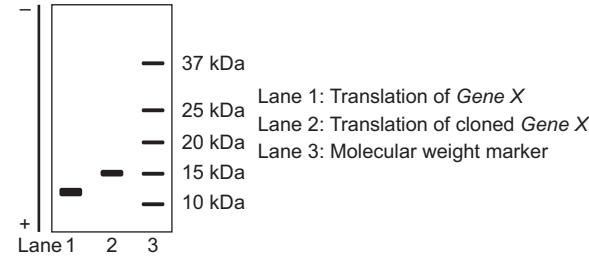
**Question 13.** You are cloning a gene from *Candida albicans* that you want to express in *E. coli* for purification. You want the *E. coli* cells to make a protein with the exact amino acid sequence as in *Candida albicans*. How could you mutate the DNA sequence of your gene of interest to ensure that the 5'-CUG-3' codon codes for serine for instead of leucine?

**Question 14.** You are screening *E. coli* for suppressor mutations for a mutant of your gene of interest. The mutation that currently causes the protein encoded by your gene to not be expressed is a

nonsense mutation. One suppressor mutation is located in the gene encoding a less common tRNA<sup>Leu</sup>.

- A. What type of suppressor mutation is described above?
- B. Why is it more likely that a less common tRNA<sup>Leu</sup> carries a suppressor mutation than a very commonly used tRNA<sup>Leu</sup>?
- C. If you are told that the suppressor mutation only involves one point mutation and can translate the 5'-UAG-3' stop codon, give the sequence of the anticodon in the mutated tRNA<sup>Leu</sup> and the wild-type tRNA<sup>Leu</sup>.
- D. Why is suppression of 5'-UAG-3' codons more efficient than suppression of the other stop codons?

**Question 15.** You are cloning *Gene X* from *S. cerevisiae* into a plasmid. You want to include the sequence 1 kb upstream and downstream from the open reading frame to try to get the gene to express under its own promoter. For this particular region of the genome, there are no other expected AUG codons upstream of the open reading frame for *Gene X*. To ensure that the protein is expressed, you try to complement the deletion strain for *Gene X* with your plasmid version of *Gene X*, but it is not complementing (not returning to the wild-type phenotype). You are not worried about splicing issues for *S. cerevisiae*. You suspect that you have a mutation in your cloned *Gene X*. The sequencing facility is down this week, but you have access to an in vitro translation system in your lab. You translate the wild-type *Gene X* and the cloned gene X using the in vitro translation system. You separate your products using SDS-PAGE and stain with Coomassie Brilliant Blue (data below). You expect a small protein, so you use a low molecular weight maker.



- A. Do you think your cloned *gene X* has a mutation? If so, what type (missense, nonsense, frameshift)? Explain your answer in terms of the data.
- B. Why do you think your cloned gene is not complementing the deletion strain? Explain your answer in terms of the data.

CHAPTER 17



# The Origin and Early Evolution of Life

THE WORKINGS OF THE CELL ARE MEDIATED by the orchestrated action of large numbers of molecular machines. With few exceptions, these machines are principally composed of protein, collectively executing nearly all aspects of the life of the cell. These include many of the processes discussed in this textbook, such as the replication and repair of the genome, control of gene expression, cell division, and chromosome segregation. In addition, metabolism, energy production, and photosynthesis are mediated by proteins, albeit often functioning in conjunction with small-molecule co-factors. Briefly put, we live in a “Protein World” in which the basis for life is largely the concerted action of polymers of amino acids. If so, then it behooves us to ask how this came to be. Indeed, and arguably, no issue is more profound than the question of how protein-based life arose on Earth, if, indeed, it did arise on this planet. By what sequence of chemical and physical events did life arise and evolve into the contemporary protein-based world of living things?

The question of the origin of life immediately poses a conundrum. If the machinery for replicating DNA and translating genetic information into protein is made of protein, then how could the duplication and expression of the genetic material have occurred before there were proteins? Assuming for the moment that a nucleic acid–based genetic material could somehow have arisen in a primordial soup, how could it direct the production of the very proteins it would need to propagate itself before there were proteins? And could this propagation have taken place before there were primitive cells (**protocells**), such as the membrane vesicles we consider later, to house the replication machinery, and how could such protocells have arisen? Beyond the conceptual problems posed by these questions is the formidable challenge of asking historical questions. We cannot go back in time to revisit the events in the origin of life, and we have been left with few or no clues in the fossil record. Nonetheless, and despite these obstacles, molecular biology enables us to pose hypotheses regarding how life might have arisen and the prospect, if not yet the reality, of creating life in a test tube as a proof-of-principle.

What exactly do we mean by life? Life is difficult to define because it is a process, not a thing. Yet, even a child can reliably distinguish an inanimate object from something that is alive. Here, and for the purposes of this chapter, we use the following minimalist definition that best applies to the earliest forms of life: namely, contemporary life arose from a system that was

## O U T L I N E

- When Did Life Arise on Earth?, 594
- 
- What Was the Basis for Prebiotic Organic Chemistry?, 595
- 
- Did Life Evolve from an RNA World?, 599
- 
- Can Self-Replicating Ribozymes Be Created by Directed Evolution?, 599
- 
- Does Darwinian Evolution Require Self-Replicating Protocells?, 603
- 
- Did Life Arise on Earth?, 606
- 
- Visit Web Content for Structural Tutorials and Interactive Animations

capable of self-replication and that was subject to Darwinian evolution. **Self-replication** means that the system relied only on small molecules and energy for propagation. **Darwinian evolution** means that the system was subject to mutation and modification, allowing for the appearance and selection of more complex, variant systems that propagated more efficiently.

Among contemporary life-forms, the simplest, free-living organisms are members of the genus *Mycoplasma*. These bacteria lack a cell wall and have tiny genomes as exemplified by the species *Mycoplasma genitalium*, which has a genome of 580 kilobases (kb), representing 500 protein-coding genes. Even simpler than *Mycoplasma* are viruses, but these small, infectious agents are not considered to be alive because they do not have a cellular structure, do not propagate by cell division, and rely on their host cells for more than small molecules and energy. In addition, simpler than *Mycoplasma* are certain intracellular symbionts of insects, such as the cicada symbiont *Hodgkinia cicadicola*, which has a genome of only 144 kb. Unlike *Mycoplasma*, however, these symbionts are only able to survive inside the cells of their insect hosts, upon which they are obligatorily dependent for their metabolism and propagation.

*Mycoplasma* bacteria consist simply of a cytoplasmic membrane and the machinery for cytokinesis, chromosome replication and segregation, protein synthesis, and metabolism, representing a minimal parts list for protein-based life. (Despite their simplicity, *Mycoplasma* bacteria are not direct descendants of the earliest forms of life. Rather, they are derived by loss of genes and functionalities from more complex bacteria with cell walls and larger genomes.) Interestingly, it is even possible to go a step further in the laboratory and generate wall-less bacteria known as **L-forms** that resemble *Mycoplasma* but lack the machinery for cytokinesis. Instead, these L-forms divide by a spontaneous blebbing-like process not unlike that of the protocells we discuss later in this chapter. Thus, the minimal parts list for contemporary life need not include the machinery for **cytokinesis**.

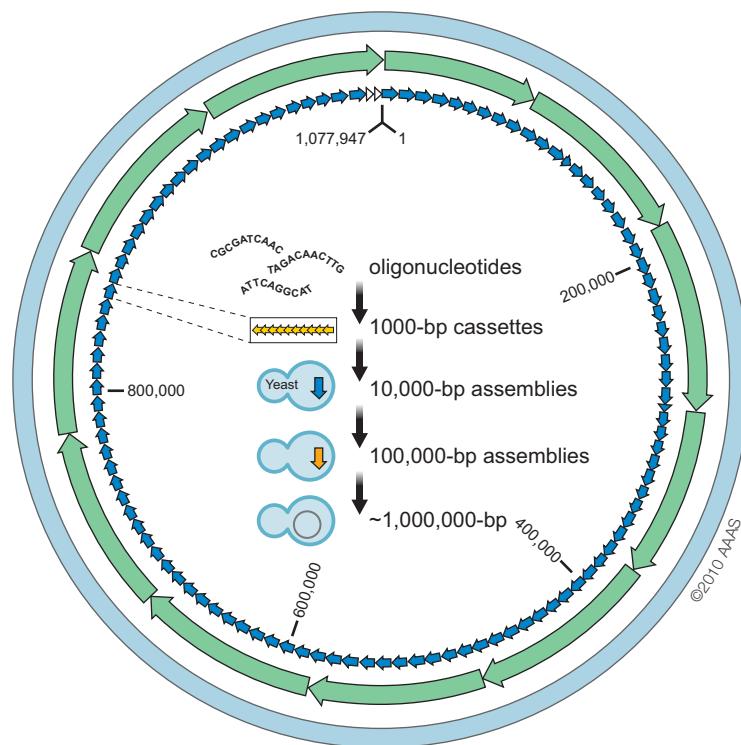
Because *Mycoplasma* bacteria have such small genomes, it has recently become feasible to synthesize the entire genome of the species *Mycoplasma mycoides* and to create a living cell with an artificial genome. The 1 million base pair (Mb) genome of *M. mycoides* was created from chemically synthesized, 1-kb units that were assembled stepwise into a complete genome and propagated in yeast as a surrogate host (Fig. 17-1). In a final step, the synthetic *M. mycoides* genome was recovered from the yeast cells and transplanted into the cytoplasm of another *Mycoplasma* species, thereby displacing that bacterium's genome. This transplantation resulted in a free-living bacterium whose genome was created entirely by synthetic means (Fig. 17-2).

The goal of this chapter is to consider how and under what conditions simple encapsulated molecular systems capable of self-replication might have arisen and how through Darwinian evolution these primitive systems might have given rise to the ancestors of contemporary single-cell organisms. Unlike almost all other topics considered in this textbook, our understanding of the origin of life is largely at the stage of conjecture. We therefore delve into this topic as a series of questions.

## WHEN DID LIFE ARISE ON EARTH?

---

Assuming for now that life arose on Earth (and was not seeded here from elsewhere in the universe), we can presume that it could not have arisen before the appearance of liquid water. Earth formed over the course of tens of millions of years with its moon arising from a collision with a giant projectile.



**FIGURE 17-1** Creating a synthetic genome. The scheme shows the strategy used for the chemical synthesis of the complete genome of *Mycobacterium mycoides*. Cassettes of ~1000 bp (yellow arrows) were assembled stepwise into increasingly longer units and, finally, into a complete artificial chromosome using yeast as a host cell. (Adapted, with permission, from Gibson D.G. et al. 2010. *Science* 329: 52–56. © AAAS.)

This collision marked the end of the main phase of growth of the planet at ~4.5 billion years ago (4.5 Ga, or gigaannum). The energy of the impact would have melted the mantle of the Earth, and hence liquid water would not have existed until enough cooling took place to allow water clouds to form. The geological eon from the formation of the Earth to ~3.8 Ga, corresponding to the age of the oldest, well-preserved sedimentary rocks, is the **Hadean**. It is followed by the **Archaean eon**, which ended at 2.5 Ga, roughly corresponding to the advent of an oxygen-containing atmosphere (Fig. 17-3). Isotopic studies of the mineral zircon indicate that oceans began to form no later than 4.4 Ga, resulting in a stable hydrosphere by ~4.2 Ga in the **Hadean eon**, setting an early boundary for events leading to the origin of life.

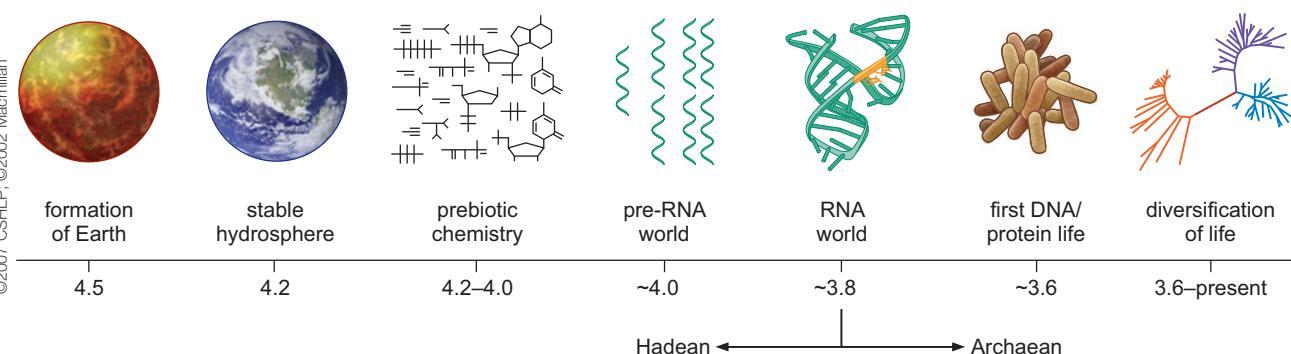
When after the formation of a stable hydrosphere did life appear? Early forms of life are difficult to identify with confidence. Nonetheless, stromatolites (layered structures) in the fossil record formed from the trapping of microbial communities in physical sediments, and isotopic evidence dating the appearance of biological carbon and sulfur cycles indicate that life existed as early as 3.5 Ga in the Archaean eon. Evidently, then, life arose during the Hadean eon or early in the Archaean eon but precisely when remains uncertain. A complication in dating the origin of life is that life might have arisen independently multiple times. Yet only one such life-form became what is referred to as the **Last Universal Common Ancestor** to contemporary life. This Last Universal Common Ancestor might have appeared earlier or conceivably later than the earliest fossils.



**FIGURE 17-2** Living bacteria with a synthetic genome. The complete, artificial chromosome, constructed as described in Figure 17-1, was removed from yeast cells and used to displace the natural chromosome from another *Mycoplasma* species. Shown here is a scanning electron micrograph of a group of *Mycoplasma mycoides* JCVI-Syn1.0 bacteria, each carrying the synthetic genome. Each bacterium is ~500 nm in diameter. (Image by Thomas Deerinck and Mark Ellisman, NCMIR, UCSD, and John Glass, JCVI.)

## WHAT WAS THE BASIS FOR PREBIOTIC ORGANIC CHEMISTRY?

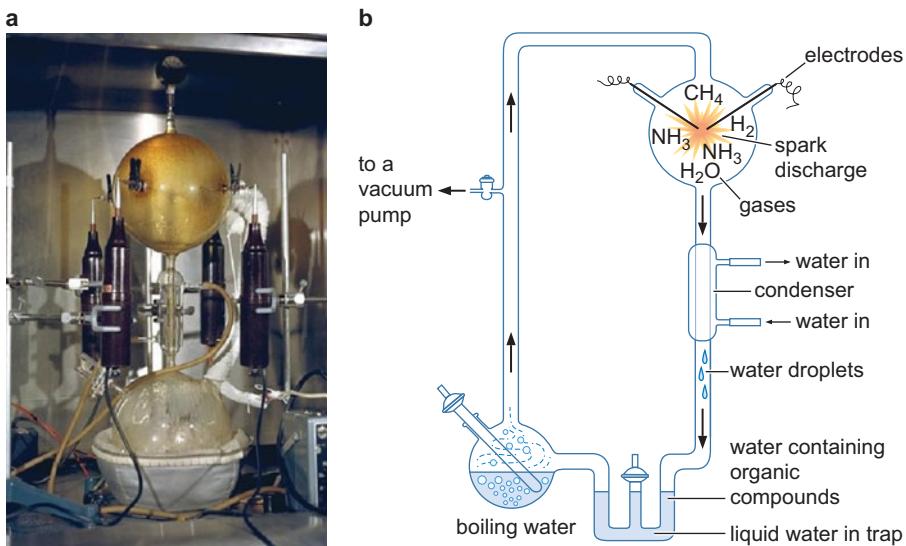
The building blocks of life are **organic molecules**. If life arose spontaneously on Earth, then it must have done so from organic molecules that resemble the



**FIGURE 17-3** Geological timeline for early Earth and the appearance of life. Shown is the sequence of events from the formation of Earth through the Hadean eon and into the Archaean eon and the corresponding events in the origin of life according to the RNAWorld hypothesis. (Modified, with permission, from Barton N.H. et al. 2007. *Evolution*, Fig. 4.4, p. 91. © Cold Spring Harbor Laboratory Press. Originally, with permission, from Joyce G.F. 2002. *Nature* **418**: 214–221. © Macmillan.)

constituents of contemporary life. In our Protein World, amino acids, sugars, nucleotides, and lipids are produced by enzymes in complex, multistep biosynthetic pathways. Where did the organic building blocks of life come from before there was life?

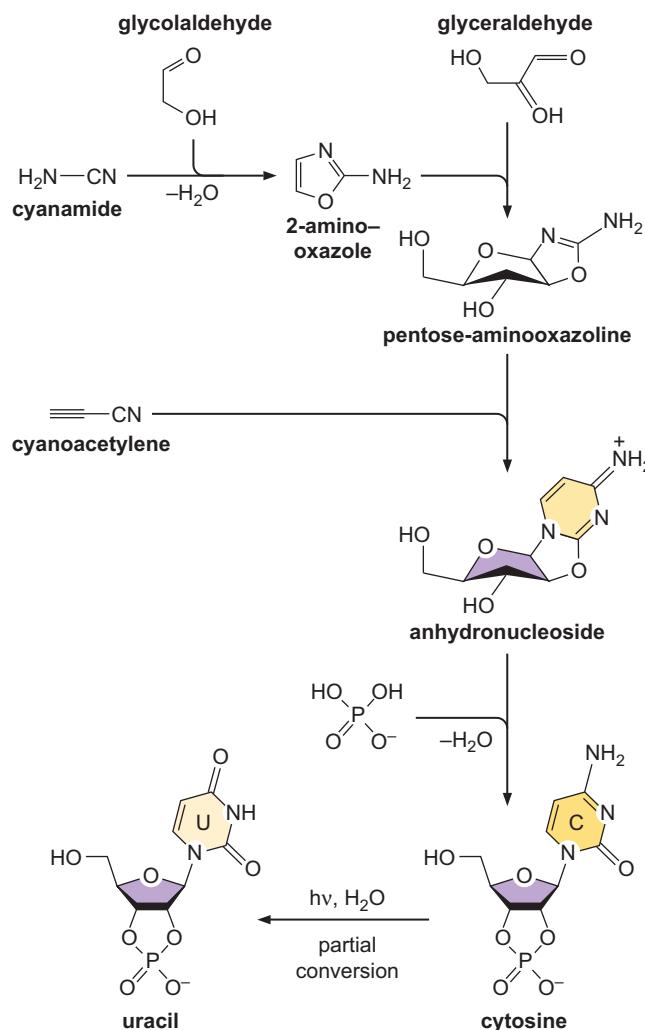
The year 1953 is recognized in the history of molecular biology for the revolutionary discovery of the structure of DNA by James D. Watson and Francis H. Crick. But 1953 is also celebrated for the classic experiment of Stanley Miller and Harold Urey of the University of Chicago that addressed the question of whether the conditions of primitive Earth could have sustained chemical reactions capable of creating organic molecules from inorganic precursors. Miller and Urey subjected water, methane, ammonia, and hydrogen to electrical discharges in sealed flasks (Fig. 17-4). Within a week,



**FIGURE 17-4** The Miller–Urey experiment. The experiment showed that amino acids could be created from water, methane, ammonia, and hydrogen. (a) Shown here is a recreation of the original apparatus for the experiment. (Photo courtesy of NASA.) (b) The diagram of the apparatus shows the sequence of steps. Water was boiled in a 500-cc flask, driving steam and gases upward through an electrical discharge created by tungsten electrodes in a 5-L flask. Steam from the large flask was cooled in the condenser back into water, which collected in the U-shaped trap at the bottom with chemical products generated by the discharge.

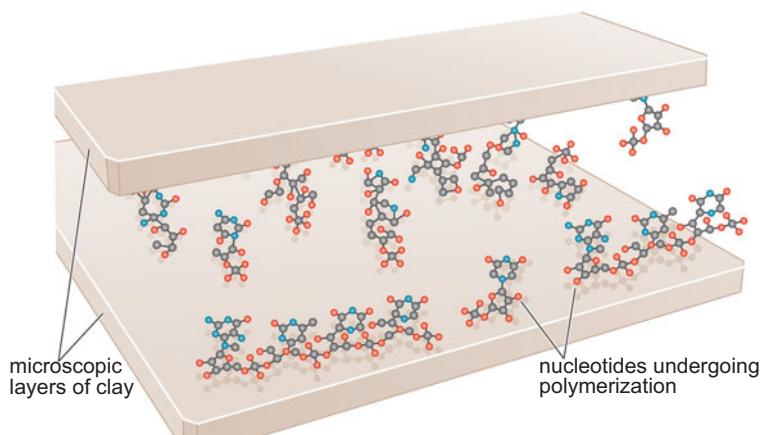
a significant portion of the methane was converted into organic molecules including a racemic mixture of 11 amino acids. Analyses of the contents of sealed vials from some of Miller's later experiments in which he included hydrogen sulfide revealed the presence of all 20 amino acids found in contemporary proteins as well as many other amino acids. Remarkably, these analyses were performed by other scientists more than half a century after Miller had performed the original experiments and sealed the vials. Miller's conditions were believed at the time to simulate the atmosphere of early Earth, which was assumed to be reducing and rich in methane. But current thinking is that methane was of low abundance in the early atmosphere (except perhaps for brief periods), with carbon largely in the form of carbon dioxide. Nonetheless, organic molecules are also generated when carbon dioxide, nitrogen, and water are subjected to electrical discharge, ionizing radiation, and ultraviolet light.

Despite the success of the iconic experiments of Miller and Urey, efforts to mimic the conditions of **prebiotic chemistry** have not until recently succeeded in generating the building blocks of polynucleotides. Earlier and unsuccessful approaches were based on reacting phosphate, ribose, and nucleobases in an effort to generate nucleotides. However, nucleotides have now been created via a new approach involving four simple organic molecules (cyanamide, cyanoacetylene, glycolaldehyde, and glyceraldehyde) that are readily produced under plausible prebiotic conditions (Fig. 17-5).



**FIGURE 17-5** Generating pyrimidine nucleotides from simple organic molecules. Shown is a recently discovered route for the synthesis of a pyrimidine nucleotide from simple organic molecules. The building blocks glycolaldehyde and cyanamide react to create 2-amino-oxazole, which can be thought of as part of a pentose sugar and part of a pyrimidine ring. The addition of glyceraldehyde completes the pentose in pentose-aminooxazoline, and the addition of cyanoacetylene completes the pyrimidine, creating an anhydronucleoside. Rearrangement and reaction with phosphate convert the anhydronucleoside to a mixture of cytosine (C) and uracil nucleotides (U). (Adapted, with permission, from Sutherland J.D. 2010. Ribonucleotides. In *The origins of life* (ed. Deamer D. and Szostak J.W.), pp. 109–121, Fig. 2, p. 114. © Cold Spring Harbor Laboratory Press.)

**FIGURE 17-6** Postulated role for clay in polyribonucleotide synthesis. Clay minerals have been shown to promote phosphodiester-bond formation by binding and concentrating nucleotides. Microscopic layers of clay mineral may have played a similar role in the origin of life in the formation of the first polyribonucleotides. (Adapted, with permission, from Ricardo A. and Szostak J.W. 2009. *Sci. Am.* 301: 54–61. © Andrew Swift MS CM1.)



For example, cyanoacetylene is a major product of the reaction of methane and nitrogen when subjected to an electrical discharge. The key new finding is that glycolaldehyde, cyanamide, and phosphate (acting as a buffer and as an acid–base catalyst) react to form an intermediate (2-amino-oxazole) that is part of a pyrimidine ring and part of a pentose sugar (Fig. 17-5). Further reactions involving cyanoacetylene, glyceraldehyde, and phosphate convert 2-amino-oxazole to pyrimidine nucleotides (in the form of cytidine and uridine 2',3'-cyclic monophosphates). Thus, rather than attempting to join phosphate, ribose, and nucleobases, the new strategy creates nucleotides via an intermediate that is neither a ribose nor a nucleobase.

Yet other work has shown that nucleotides can react to form oligonucleotides on clay. Microscopic layers of clays are found to adsorb and concentrate nucleotides, allowing them to react with each other to form ribonucleotide chains (Fig. 17-6). *In toto*, these recent successes potentially fill an important gap in our understanding of the origin of life, because the most compelling ideas on how life got its start posit the spontaneous appearance in prebiotic Earth of self-replicating RNA molecules (as we explain later).

Finally, we note that prebiotic organic molecules need not have arisen solely from reactions involving atmospheric carbon. Another potential source of organic molecules was from comets and meteorites after the Moon-forming impact. Certain meteorites known as chondrites (meaning that they had not undergone melting during their formation) are sometimes found to be carbon-containing (carbonaceous chondrites). Chemical analyses have revealed the presence in these carbonaceous chondrites of racemic mixtures (containing both L- and D-chirality molecules) of numerous amino acids, including those found in contemporary proteins. One large and well-studied carbonaceous chondrite, the **Murchison meteorite**, which landed in Murchison, Australia in 1969, is particularly rich in organic molecules (Fig. 17-7). Intriguingly, among the amino acids identified in the Murchison meteorite and other carbonaceous chondrites are some, such as isovaline, that are recovered with an enantiomeric excess of L to D chirality. Indeed, only L-amino (and not D-amino) excesses have been found in material from carbonaceous chondrites. Conceivably, L-amino acid homochirality (arising perhaps by amplification of an excess of L to D) was a feature of the origin of life from its very beginnings. Another provocative discovery from the Murchison meteorite was the presence of certain nucleobases, including uracil. Thus, the transport to Earth of amino acids and nucleobases on carbonaceous chondrites may have contributed to the chemical soup that spawned the earliest living systems.



**FIGURE 17-7** The Murchison meteorite. This meteorite can be found at the National Museum of Natural History in Washington, D.C. The total known weight of the meteorite is 100 kg. ([http://en.wikipedia.org/wiki/File:Murchison\\_crop.jpg](http://en.wikipedia.org/wiki/File:Murchison_crop.jpg))

## DID LIFE EVOLVE FROM AN RNA WORLD?

---

As we saw in Chapter 5, RNA is capable of great diversity in structure and function, including its capacity to be an enzyme. The discovery of RNA enzymes—**ribozymes**—provided a potential solution to the conundrum posed at the beginning of this chapter. Thus, if RNA could serve both as an information carrier and simultaneously as a self-replicase, then life could, in principle, have started without the need for protein. In this view, mutation and natural selection would have eventually allowed systems to evolve that had the capacity to catalyze peptide-bond formation, in a manner directed by nucleic acid sequence information. The idea that contemporary, protein-based life arose from earlier life-forms based on RNA or related molecules is known as the **RNA World hypothesis**. According to this hypothesis, our Protein World evolved from a primordial RNA World (Fig. 17-3).

If so, what evidence do we have that early life was based on RNA? Much of the evidence comes from the existence of possible relics of an RNA World. These include self-splicing introns, ribozymes, and riboswitches (see Chapter 20) and nucleotide-containing enzyme co-factors, such as coenzyme A, flavin adenine dinucleotide, and nicotinamide adenine dinucleotide. In addition, the fact that deoxyribonucleotides are biosynthetically derived from ribonucleotides is consistent with the notion that RNA preceded DNA. Although most of this evidence is circumstantial, we do have one dramatic example of an apparent remnant of early life: catalysis of peptide bond formation by an RNA component of the ribosome.

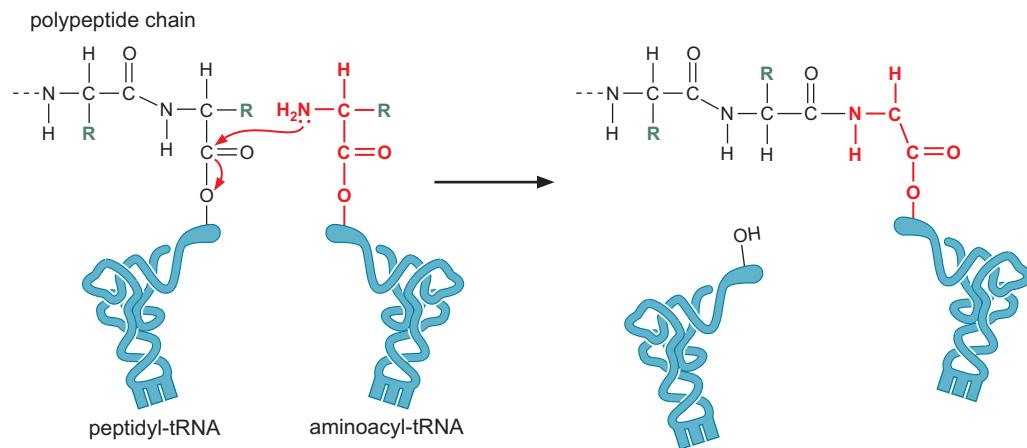
As noted in Chapter 5 and discussed in detail in Chapter 15, at the heart of the ribosome is a **ribozyme**, the large RNA component (23S in bacteria) of the large subunit, which is responsible for catalyzing peptide-bond formation by transferring the growing polypeptide chain to the amino acid moiety of the incoming charged tRNA on the ribosome. The structural evidence presented in Chapter 15 shows that this reaction is catalyzed entirely by RNA without the apparent participation of protein components of the ribosome whose side chains do not extend into the catalytic center. Unlike other naturally occurring ribozymes, which act on phosphorous centers, the ribosome ribozyme acts on a carbon center to create the peptide bond (Fig. 17-8). Thus, the most fundamental chemical reaction in the Protein World is catalyzed by an RNA molecule. It is tempting to believe therefore that the ribosome ribozyme is a molecular fossil from an earlier life-form when many or all macromolecular transactions were executed by RNAs.

Taking poetic license, we might say that with RNA we throw away many of the rules! If DNA is straightlaced and uniform, RNA is freewheeling and audacious. RNA is the nonconformist. It has ceded primacy as the repository of genetic information to DNA, but it has gained versatility. It is a master architect, forming complex, three-dimensional (3D) structures, and it can perform catalysis, a trick it learned long before proteins knew how to be enzymes. In short, life probably evolved from an RNA World.

## CAN SELF-REPLICATING RIBOZYMES BE CREATED BY DIRECTED EVOLUTION?

---

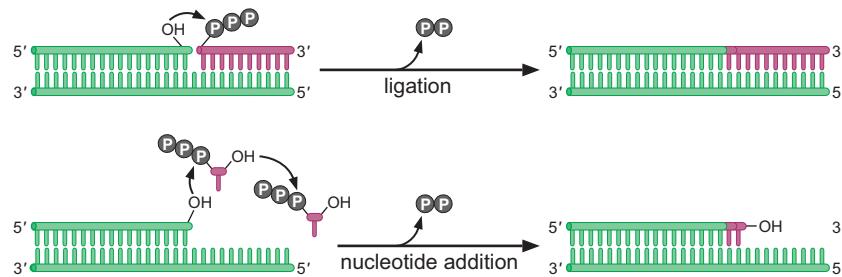
Although it is not possible to go back in time to identify the hypothesized RNA polymerase ribozymes of primordial life, we might be able to create such a self-replicating ribozyme in the laboratory by **directed evolution**. If so, the creation of an RNA polymerase ribozyme would show that life could,



**FIGURE 17-8** The peptidyl transferase acts on a carbon center to create the peptide bond. Peptide-bond formation occurs by nucleophilic attack of the amino group of an incoming amino acid (carried on a charged tRNA) on the carbon center of the carboxyl group at the end of the growing polypeptide chain. Nucleophilic attack results in transfer of the growing polypeptide chain to an incoming amino acid. The reaction is catalyzed by the large RNA component of the large subunit of the ribosome (see Chapter 15).

in principle, have gotten its start with a non-protein replicase. We usually think of RNA as being single-stranded but such a self-replicating ribozyme would generate double-stranded RNA by polymerizing ribonucleotides on a complementary RNA template. It has not yet been possible to create an RNA polymerase ribozyme that is capable of fully replicating itself. Nonetheless, RNA polymerase ribozymes that are capable of accurately polymerizing ribonucleotides on RNA templates have been created in the laboratory.

As we discussed in Chapter 5, it is possible to generate novel RNA species that have specific desirable properties by synthesizing RNA molecules with randomized sequences, followed by rounds of selection and sequence diversification until the desired property is obtained (see the discussion on Directed Evolution and Fig. 5-8). This strategy, known as Systematic Evolution of Ligands by Exponential Enrichment (SELEX), takes advantage of the enormous sequence diversity that is generated by randomizing RNA sequences. In the case of RNA polymerase ribozymes, directed evolution was achieved in two stages (Fig. 17-9). In the first stage, ribozymes were selected that were capable of joining (ligating) the 3'-hydroxyl of one RNA

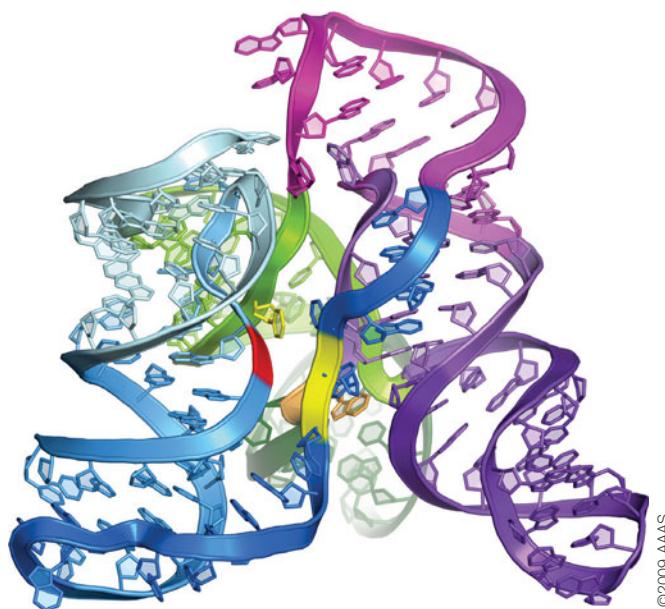


**FIGURE 17-9** A comparison of RNA ligation and RNA nucleotide addition. In ligation, the 3'-OH of one RNA is joined to the 5' end of another RNA via phosphodiester bond formation and the release of pyrophosphate. Nucleotide addition involves the same chemical reaction but the 3'-OH is joined to a nucleotide instead of to an RNA.

molecule to the 5'-triphosphate of another RNA on an RNA template that was complementary to both RNAs. This ligation strategy was used in the first stage because the chemistry of phosphodiester bond formation is the same as in the addition of a nucleoside triphosphate to the 3'-hydroxyl of an RNA molecule. Yet, the molecules to be ligated could easily be aligned with each other by annealing to a complementary RNA template. In the second stage, one such ligase ribozyme was subjected to rounds of sequence diversification, but this time the selection was for RNA molecules that could attach nucleotides to their 3' termini. Selection took advantage of the fact that elongation at the 3' terminus would cause the ribozyme to grow larger. Molecules were selected in which the addition of nucleotides occurred in a manner that was dependent on an RNA template. Successive rounds of selection yielded polymerases of progressively greater efficiency, and the use of a variety of templates ensured that the sequence of incorporated nucleotides was determined by the sequence of the template, rather than by an intrinsic property of the ribozyme.

This strategy culminated in the creation of an ~200-nucleotide-long RNA polymerase ribozyme that is capable of extending its 3' terminus in an accurate, template-dependent manner by at least 20 nucleotides, representing almost two turns of a helix. And very recently, further selection strategies yielded an RNA polymerase ribozyme that is capable of generating RNAs as long as 95 nucleotides (although with low efficiency). In this recent example, the RNA polymerase ribozyme is capable of extending a primer bound to a separate RNA template. Strikingly, it is capable of copying in its entirety a template for the hammerhead ribozyme discussed in Chapter 5, resulting in an RNA product that is itself enzymatically active! The ribozyme-generated hammerhead ribozyme was able to accurately cleave a separate substrate RNA.

What do these RNA polymerase ribozymes look like and how do they work? We do not yet know, but a clue comes from the structure of a 120-nucleotide-long ligase ribozyme similar to that used as a starting point for the directed evolution of RNA polymerase ribozymes. The tertiary structure of the ribozyme consists of three helical domains that resemble a tripod with the catalytic center identified in yellow (Fig. 17-10). Nucleotides important in catalysis were identified in experiments involving chemical

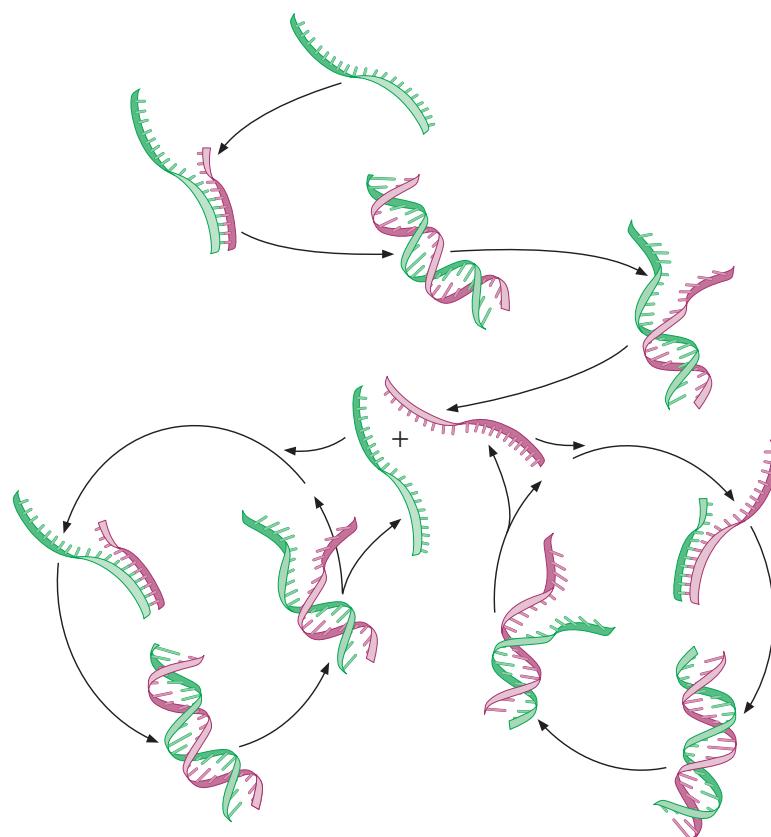


**FIGURE 17-10 Structure of an evolved RNA ligase ribozyme.** The tripod-like crystal structure shows the three domains in blue, pink, and purple with catalytic center (yellow) and ligation site (red). (Adapted, with permission, from Shechner D.M. et al. 2009. *Science* **326**: 1271–1275, Fig. 1C. © AAAS; kindly provided to us by D.M. Shechner.)

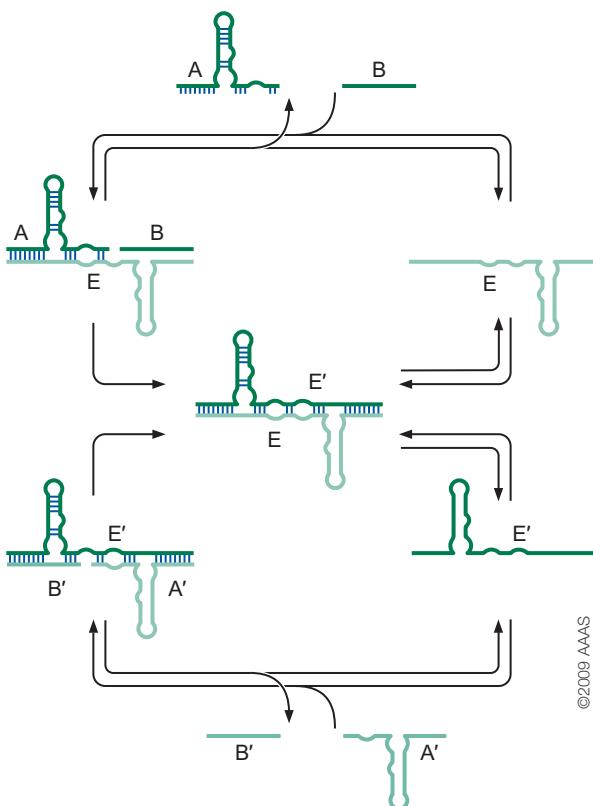
modification and substitution of specific nucleotides. This evidence assigns a critical role to a particular cytosine in the active site, leading to the model in which the exocyclic amine of the base stabilizes the developing negative charge on the pyrophosphate leaving group during phosphodiester-bond formation. (See Chapter 9 for a discussion of the mechanism of phosphodiester-bond formation by proteinaceous polymerases.)

Despite these successes, we are still a long way from a ribozyme that is capable of copying itself entirely. What is needed is a ribozyme that produces a complete and separate copy of itself. This leads us to another conceptual problem. It seems difficult to imagine that a single RNA molecule could simultaneously serve as a template for its own duplication and as the polymerase that is doing the copying. Instead, it is likely that the primordial **replicase ribozyme** would have had to make a copy of a sister molecule. This would have temporarily generated an RNA–RNA duplex with two complementary strands, one being the replicase and the other being the complement of the replicase (Fig. 17-11). Additional rounds of copying of the complement would generate more replicases, whereas additional rounds of copying of the replicase strand would generate more copies of the complement, which would, in turn, serve as templates for generating additional replicases.

Although we still lack a fully self-replicating RNA polymerase ribozyme, self-sustained replication of a ribozyme that ligates RNA molecules together has been achieved. This self-replicating system consists of complementary ligase ribozymes that join pairs of complementary RNA substrates to each other (Fig. 17-12). In this system, a ligase ribozyme designated E (light green in figure) catalyzes the template-mediated joining of two RNA molecules (A and B) that are complementary to the ribozyme. This produces a complementary RNA designated E' (dark green in the center of the figure) that is



**FIGURE 17-11** The primordial replicase ribozyme would have copied its complement. Because the product of the replicase (original replicase shown in green) would be a complement of itself, the newly synthesized strand (shown in purple) would not be a replicase. Instead, it would be the complement of the replicase. This complement could, in turn, serve as a template for the synthesis of additional replicases. Thus, the primordial replicase ribozyme would likely have generated a double-stranded product that would, in a subsequent step, have dissociated into replicase and complement strands. Copying of the replicase strand would generate additional template strands, and copying of the complement strand would generate additional replicases. Throughout the figure the replicase strands are shown in green and the complement strands are shown in purple. (Adapted, with permission, from Ricardo A. and Szostak J.W. 2009. *Sci. Am.* **301**: 54–61. © AAAS.)



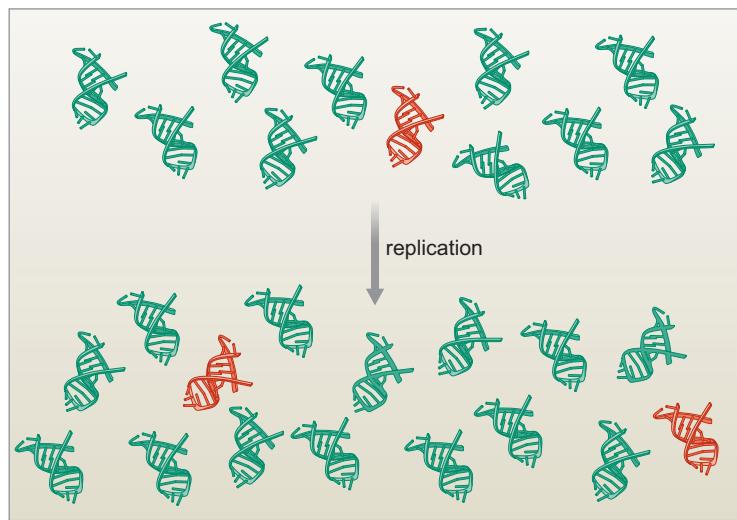
**FIGURE 17-12** A self-sustaining replicating ribozyme. The ligase ribozyme (E) joins two RNA molecules (A and B), each complementary to a part of E, thereby producing the complementary ribozyme E'. See text for further details. (Adapted, with permission, from Lincoln T.A. and Joyce G.F. 2009. *Science* **323**: 1229–1232. Fig. 1A, p. 7. © AAAS.)

itself a ligase ribozyme for two RNAs that are complementary to it (A' and B' in the bottom of the figure). This round of ligation creates a new molecule of E (light green in center of the figure). Thus, replication consists of two interconnected cycles: one catalyzed by E and shown in the top half of the figure that generates E' from A and B and one catalyzed by E' and shown in the bottom half of the figure that generates E from A' and B'. Remarkably, E and E' catalyze multiple rounds of ligation from a common pool of four substrate RNAs (two—A and B—complementary to E and two—A' and B'—complementary to E'), culminating in the amplification of E and E' and the depletion of the substrates. Interestingly, different ligase ribozymes were shown to be capable of competing with each other for a limited pool of substrates, representing a primitive example of genetic selection. Thus, a ribozyme can, indeed, be fully self-replicating, but so far only with RNAs rather than with ribonucleotides as substrates.

### DOES DARWINIAN EVOLUTION REQUIRE SELF-REPLICATING PROTOCELLS?

We are left with yet another conundrum! Even if we imagine an RNA World in which the same RNA molecules served both as the genetic material and as replicase ribozymes, such a system would have arguably been limited in its capacity to undergo Darwinian evolution. Consider a primordial soup populated with replicase ribozymes, and imagine that a mutation arises in one of the replicases that enables it to replicate more efficiently than its siblings. If, as argued above, each replicase must copy another replicase, then the improved replicase would be constrained to merely making copies of unimproved sister molecules in the primordial soup, and its own

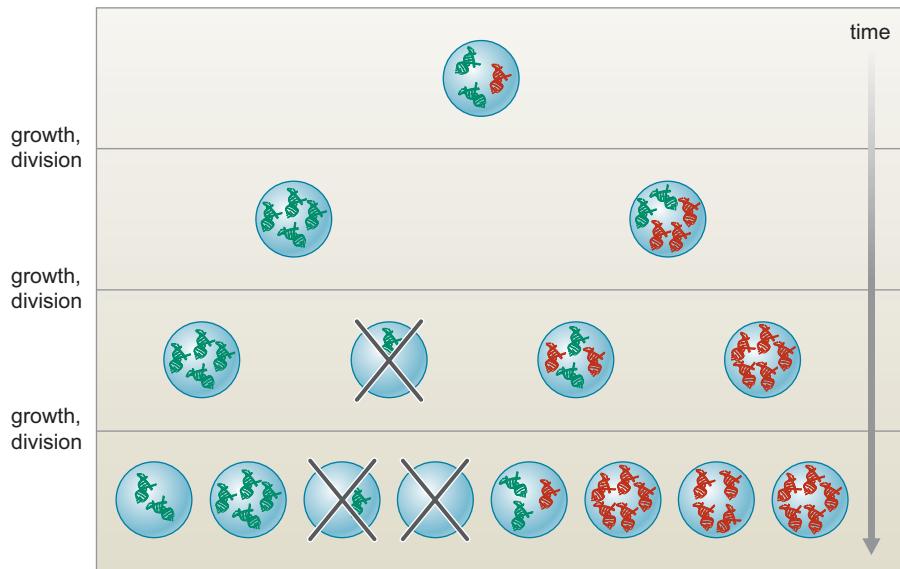
**FIGURE 17-13** Improved mutant replicase ribozymes would have no selective advantage in the primordial soup. Consider a primordial soup without compartments containing a population of replicase ribozymes (green) and a rare mutant replicase (red) with improved replication efficiency. After a round of replication in which the replicase ribozymes randomly copy each other, there is no enrichment for the improved replicase ribozyme relative to the unimproved ones for the following reason: unimproved replicases are just as likely to copy each other as the improved replicase. And after an improved replicase has been duplicated, because of diffusion the daughter replicases are much more likely to encounter and copy unimproved replicases than a sibling.



propagation would depend on the action of unimproved siblings. (The probability of two improved replicases being near enough to each other to copy each other would, presumably, have been vanishingly small [Fig. 17-13].) Thus, in this scenario, an improved mutant replicase would have no opportunity to amplify itself more rapidly than its sisters. These considerations have led to the idea that the early life not only required replicase ribozymes, but also, as we now explain, protocells to house the replicases.

Imagine that the replicases were encapsulated in membrane protocells that could grow and divide by uptake of lipids. Instead of a primordial soup with vast numbers of largely identical replicases, small numbers of replicases might be grouped (binned) together in protocells, isolated from other replicases (Fig. 17-14). Now imagine that in one of these protocells a rare mutant replicase arises that is superior to its siblings. Because there are only a small number of replicases in the protocells, the improved replicase would have a reasonable chance of being copied by an unimproved sibling, resulting in more than one copy of the improved mutant replicase in the same compartment. (For the sake of simplicity, we ignore the complication introduced above that replication involves going through an intermediate

**FIGURE 17-14** Compartmentalization permits Darwinian selection. (Bottom) The consequence of binning replicases in protocells with small numbers of replicases. Owing to chance, a protocell is likely to arise that has inherited two or more copies of an improved replicase (shown in red). Now the probability of improved replicases copying each other is high. Moreover, rapid replication will drive growth and division of the protocell, allowing for selection for protocells harboring improved replicases. Also owing to chance, some protocells might inherit only one or no replicases, in which case no further propagation will be possible, as indicated by the large Xs. (Courtesy Jack W. Szostak.)

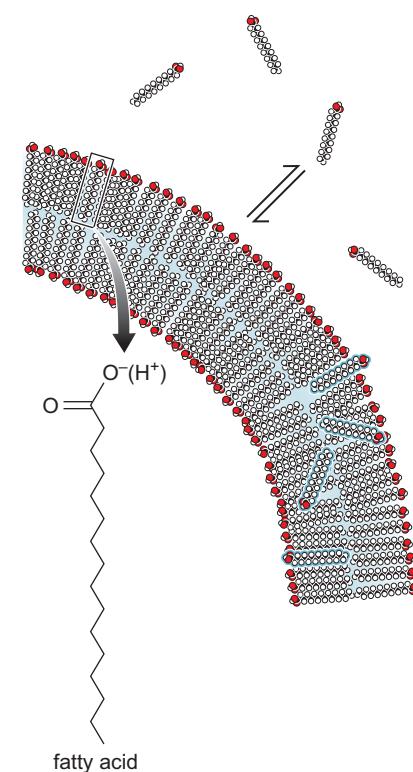


step involving a complementary copy of the ribozyme.) Now when the protocell divides, one of the daughter cells might *per chance* “inherit” two copies of the mutant replicase. If so, then in that daughter cell, improved mutant replicases might be able to make copies of each other, and hence in that protocell replication will take place more rapidly than in other protocells.

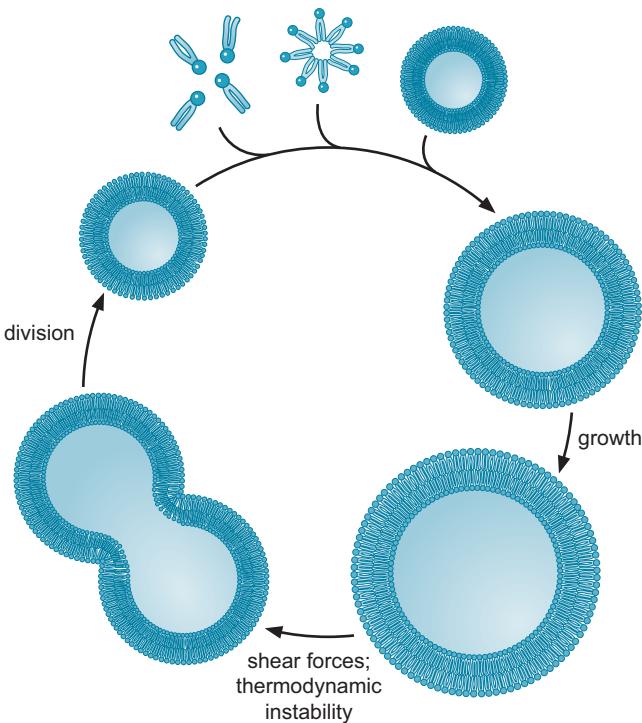
Superior replicase ribozymes could confer a growth advantage to the protocell via the following osmotic mechanism. Replicase ribozymes would replicate by the polymerization of nucleotides that would enter the protocell by diffusion across the membrane. This ribozyme-driven incorporation of nucleotides into nondiffusible RNA would be expected to increase the osmolarity of the protocell as water diffuses in to equalize its concentration inside and outside. The influx of water would create tension on the membrane, causing the protocell to swell by the uptake of lipids from the surrounding primordial soup. Growth of the protocell would, in turn, render it unstable, eventually causing it to divide. Therefore, the protocell with the enhanced mutant replicases would have a selective growth and division advantage over other protocells, resulting in disproportionate amplification of the new and improved replicase. In other words, using membrane compartments to partition replicase ribozymes allows genetic innovation to be rewarded by Darwinian selection.

Remarkably, cell-like compartments with the properties invoked for early life are relatively easy to generate in the laboratory. Simple lipids such as fatty acids (and other amphiphiles), which are components of the phospholipids found in the membranes of contemporary cells, have the capacity to spontaneously form tiny, topologically closed sacs known as vesicles in aqueous environments (Fig. 17-15). Briefly put, this is an example of the hydrophobic effect in which vesicle formation is driven by the favorable free energy of stacking up the greasy tails of the lipids against each other in sheets. The sheets, in turn, spontaneously close into sacs to minimize interaction with water molecules at their edges. Moreover, such vesicles are able to grow by accretion of additional fatty acid molecules. With growth, the vesicles become unstable and spontaneously fragment into daughter vesicles under mild agitation (Fig. 17-16). Thus, unlike polynucleotides, which might have relied on clay minerals to catalyze their earliest appearance, cell-like compartments to house polyribonucleotides could have formed spontaneously from the self-assembly of fatty acids. Finally, and pertinent to the osmotic-driven growth and division of protocells postulated above, vesicles composed of simple lipids such as fatty acids are not impermeable to nucleotides. Rather, they have been shown to allow nucleotides to diffuse across their membranes. This diffusion probably occurs because fatty acids (as compared to phospholipids) are relatively disordered in the membrane bilayer, which renders the membrane permeable to small molecules, such as nucleotides (Fig. 17-15). Thus, replicase-containing protocells composed of fatty acids could, in principle, obtain nucleotide substrates from the outside via simple diffusion.

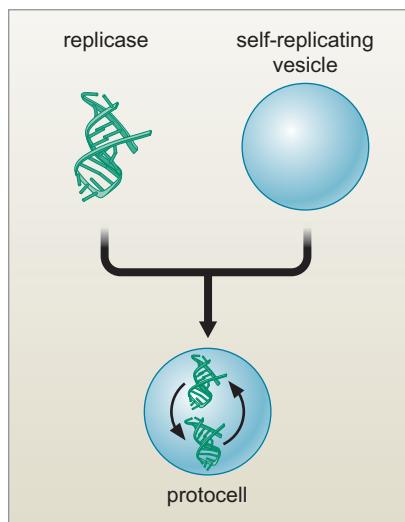
At last, we have the ingredients needed for the primordial cell that conforms to the definition of early life: relying on small molecules for reproduction (nucleotides and fatty acids) and being capable of Darwinian evolution (Fig. 17-17). According to the RNA World hypothesis, the primordial cell had an RNA genome that was also a replicase ribozyme, and it propagated itself in vesicle-like protocells. Osmotic forces from the capture of nucleotide substrates in polynucleotides would have driven the protocells to grow and then spontaneously split into daughter protocells. Moreover, because small groups of replicase ribozymes were partitioned from each other in compartments, rare mutants with enhanced capacity to replicate would have had a selective advantage, allowing for Darwinian evolution.



**FIGURE 17-15** Fatty acids spontaneously form bilayer vesicles. Fatty acids are simple lipids with a polar side chain and a carboxylate head group that spontaneously assemble into vesicles. Because the carboxylate head groups are easily neutralized, fatty acids are thought to be disordered in the bilayer, allowing the membranes to be permeable to small molecules such as nucleotides. (Redrawn from Budin I. and Szostak J.W. 2010. *Annu. Rev. Biophys.* **39**: 245–263, Fig. 4. © Annual Reviews, Inc.)



**FIGURE 17-16** Growth and division of lipid vesicles. (Courtesy Jack W. Szostak.)



**FIGURE 17-17** A hypothetical primordial cell that combines self-replicating ribozymes with self-replicating vesicles. (Courtesy Jack W. Szostak.)

In time, ribozymes would have evolved the further trick of catalyzing metabolic reactions (indeed, this may have preceded the appearance of the replicase ribozyme) and ultimately peptide-bond formation (immortalized in the contemporary peptidyl transferase ribozyme), giving rise to the beginnings of the Protein World. Of course, the journey from the replicase ribozyme to the ribosome is mysterious and must have involved many steps. These include the evolution of progenitors to transfer RNAs (tRNAs), the capacity to synthesize peptides first in a non-coded manner and then a coded manner with the invention of messenger RNAs (mRNAs), the two-subunit ribosome with its RNA and protein components, and eventually the capacity to synthesize long peptides that could adopt complex tertiary structures.

### DID LIFE ARISE ON EARTH?

The preceding narration has of necessity been highly speculative. We do not know where and how life arose, and we are a very long way from creating life in the test tube. Still, the twin concepts that RNA was both the genetic material and the replicase and that compartmentalization would permit Darwinian evolution represent important advances in understanding how life might have arisen. Nonetheless, some investigators have considered alternative chemistries to RNA and alternative views of the origin of life based on metabolic evolution rather than genetic evolution. Indeed, some scientists consider the hurdles involved in the generation of life on Earth to be so formidable as to doubt the very premise that life arose on this planet. Famously among the skeptics was Francis Crick (although, in fairness, before RNA polymerase ribozymes had been invented). Crick and others subscribed to the view that life was seeded on Earth from another planet, perhaps carried here on a meteor.

Of course, even if the (unlikely) theory that life was seeded on Earth from elsewhere is true, it begs the question of how life arose, wherever it arose, bringing us back to the very same issues with which we began this chapter.

## SUMMARY

Life arose between 4.4 Ga, when liquid water appeared on Earth, and 3.5 Ga, when life was already present as judged from isotopic and fossil evidence. Life in its simplest form is a system that is capable of self-replication and that is subject to Darwinian evolution. The discovery that some RNAs (ribozymes) catalyze enzymatic reactions led to the RNA World hypothesis, which holds that protein-based life arose from a primordial life-form in which RNA served both as an information carrier and an RNA polymerase ribozyme that was capable of self-replication. In time, self-replicating RNA molecules would have evolved the capacity to produce proteins, giving rise to the contemporary Protein World. The contemporary peptidyl transferase ribozyme, which catalyzes peptide-bond formation in the ribosome, may be a molecular fossil from the RNA World.

As a proof-of-principle of the RNA World hypothesis, researchers have attempted to create ribozymes that are

capable of self-replication using methods of directed evolution. So far, these efforts have culminated in the creation of RNA polymerase ribozymes that can synthesize RNA chains as long as 95 nucleotides in a template-dependent manner. However, the creation of a replicase ribozyme that is capable of duplicating itself has not been achieved.

A self-replicating system that is also subject to Darwinian evolution may have required cell-like compartments that were capable of growth and division. Such protocells may have arisen from lipid vesicles composed of fatty acids or other amphiphiles, which can grow by accretion of fatty acids and can fragment into daughter protocells. Thus, the earliest form of life may have been vesicle-life compartments that encapsulated replicase ribozymes and that were capable of evolving through mutation and natural selection for more rapidly propagating protocells.

## BIBLIOGRAPHY

- Deamer D. and Szostak J.W. eds. 2010. *The origins of life*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
- Johnston W.K., Unrau P.J., Lawrence M.S., Glasner M.E., and Bartel D.P. 2001. RNA-catalyzed RNA polymerization: Accurate and general. *Science* **292**: 1319–1325.
- Lincoln T.A. and Joyce G.F. 2009. Self-sustained replication of an RNA enzyme. *Science* **323**: 1229–1232.
- Mansy S.S. and Szostak J.W. 2009. Reconstructing the emergence of cellular life through the synthesis of model protocells. *Cold Spring Harb. Symp. Quant. Biol.* **74**: 47–54.
- Powner M.W., Gerland B., and Sutherland J.D. 2009. Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. *Nature* **459**: 239–242.
- Ricardo A. and Szostak J.W. 2009. Origin of life on earth. *Sci. Am.* **301**: 54–61.
- Shechner D.M., Grant R.A., Bagby S.C., Koldobskaya Y., Piccirilli J.A., and Bartel D.P. 2009. Crystal structure of the catalytic core of an RNA-polymerase ribozyme. *Science* **326**: 1271–1275.
- Wochner A., Attwater J., Coulson A., and Holliger P. 2011. Ribozyme-catalyzed transcription of an active ribozyme. *Science* **332**: 209–212.
- Zaher H.S. and Unrau P.J. 2007. Selection of an improved RNA polymerase ribozyme with superior extension and fidelity. *RNA* **13**: 1017–1026.
- Zhu T.F., Schrum J.P., and Szostak J.W. 2010. The origins of cellular life. *Cold Spring Harb. Perspect. Biol.* **2**: a002212.

## QUESTIONS

### MasteringBiology®

*For instructor-assigned tutorials and problems, go to MasteringBiology.*

For answers to even-numbered questions, see Appendix 2: Answers.

**Question 1.** Name two key characteristics that define life in the context of the origin of life.

**Question 2.** Explain why bacteria like *Mycoplasma genitalium* are considered alive unlike viruses or the symbiont *Hodgkinia cicadicola*.

**Question 3.** Think about the Miller and Urey experiment. What impact did the results have on the study of the origin of life? What is a drawback to the results?

**Question 4.** Give a specific example of a catalytic RNA critical to cells today. How does this example help support the RNA World hypothesis for the origin of life?

**Question 5.** The favored hypothesis is that life evolved from an RNA World. Provide some reasons why a protein-centered

hypothesis is not favored. Provide some reasons why a DNA-centered hypothesis is not favored.

**Question 6.** For each function, name the enzyme or ribozyme that performs the function.

- i. Transcribes RNA from a DNA template
- ii. Synthesizes a complementary DNA strand from an RNA template
- iii. Replicates DNA from a DNA template
- iv. Replicates RNA from an RNA template

**Question 7.** List the key steps necessary to select a ribozyme with ligase activity through Systematic Evolution of Ligands by Exponential Enrichment (SELEX). Assume that you will complete multiple rounds of selection and that you want to diversify the pool each round. You add an RNA molecule that includes a sequence tag to the mixture of your potential ribozymes. If a ribozyme ligates the tag to its own 5' end, the ligating ribozyme can be purified out from the mix.

**Question 8.** Explain how compartmentalization in a protocell enhances the propagation of a more efficient, mutant RNA replicase over the other RNA replicases.

**Question 9.** Describe the membrane of a laboratory model protocell. How does this model protocell grow and divide?

**Question 10.** Do scientists think the building blocks of polynucleotides arose from a reaction between phosphate, ribose, and a nucleobase or from a reaction between intermediates derived from reactions between organic molecules present on primitive Earth? Explain your answer.

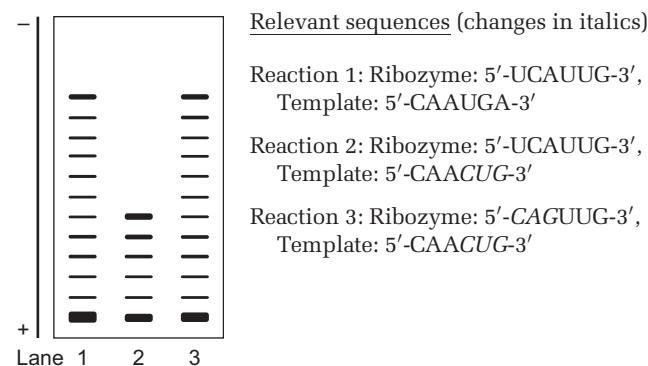
**Question 11.** Explain how pyrophosphate is stabilized in the mechanism of protein-based phosphodiester bond formation versus the model for ribozyme-based phosphodiester bond formation.

**Question 12.** List the reaction components and ribozyme function required for self-replication by a ribozyme.

**Question 13.** Describe the difference between ligation of RNA and RNA polymerization.

**Question 14.** You are studying the properties of an RNA polymerase ribozyme using a primer extension assay. This particular ribozyme has enhanced activity compared to a previous version of the ribozyme. The difference between the two ribozymes is that the enhanced ribozyme includes an additional domain at

the 5' end. In each reaction of your primer extension, you include your ribozyme, your 5'-radiolabeled RNA primer bound to an RNA template, and rNTPs in the appropriate buffer. In reactions 2, you altered the sequence of the RNA template. In reaction 3, you altered the sequence of the RNA template and the sequence within the new domain of the ribozyme. You separate your products by denaturing polyacrylamide gel electrophoresis and visualize the bands on an autoradiograph (shown below).



- A. Based on the data, describe the relative differences in ribozyme replicase activity between the three reactions.
- B. Hypothesize why the products in lane 3 have a similar migration pattern as the products in lane 1.

Data adapted from Wochner et al. (2011. *Science* **332**: 209–212).

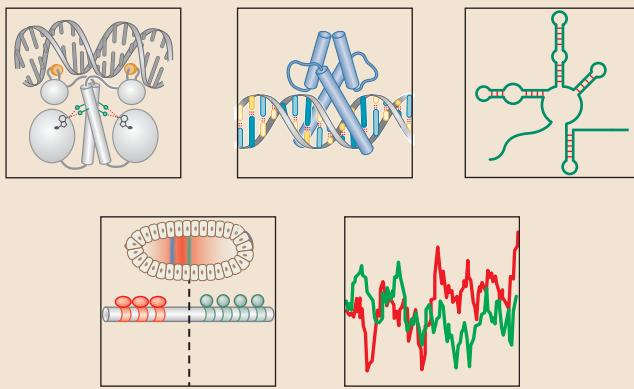
**Question 15.** Researchers studied the permeability of fatty acid vesicles as model protocells.

- A. Explain why membrane permeability properties are important when considering the RNA World hypothesis.
- B. In one experiment, the researchers encapsulated nucleotides within fatty acid vesicles. Instead of measuring permeability by detecting nucleotides entering the vesicle, they measured the percent of nucleotides from within the vesicle that left the vesicle. They found negligible loss of AMP, ADP, and ATP from the vesicles over a 24-hour period. In the presence of Mg<sup>2+</sup>, they found that AMP and ADP slowly leaked from the vesicles, but not ATP. Propose why AMP and ADP are able to cross the membrane in the presence of Mg<sup>2+</sup>. What does the failure of ATP to cross the membrane in the presence of Mg<sup>2+</sup> suggest about the permeability of early protocells?

Data adapted from Mansy et al. (2008. *Nature* **454**: 122–125).

P A R T  
**5**

# REGULATION



## O U T L I N E

---

- CHAPTER 18**  
Transcriptional Regulation in Prokaryotes, 615
- 
- CHAPTER 19**  
Transcriptional Regulation in Eukaryotes, 657
- 
- CHAPTER 20**  
Regulatory RNAs, 701
- 
- CHAPTER 21**  
Gene Regulation in Development and Evolution, 733
- 
- CHAPTER 22**  
Systems Biology, 775

**I**N PART 4 WE CONSIDERED HOW THE GENETIC information encoded in the DNA is expressed. This involves the transcription of DNA sequences into an RNA form, which is then used as a template for translation into protein.

But not all genes are expressed in all cells all the time. Indeed, much of life depends on the ability of cells to express their genes in different combinations at different times and in different places. Even a lowly bacterium expresses only some of its genes at any given time, thus ensuring it can, for example, make the enzymes needed to metabolize the nutrients it encounters while not making enzymes for other nutrients that are not available at that time. Development of multicellular organisms offers an even more striking example of this so-called “differential gene expression.” Essentially all the cells in a human contain the same genes, but the set of genes expressed in forming one cell type is different from that expressed in forming another. Thus, a muscle cell expresses a set of genes different (at least in part) from that expressed by a neuron, a skin cell, and so on. By and large, these differences occur at the level of transcription—most commonly, the initiation of transcription.

In the following chapters, we look mainly at how transcription is regulated. We start in Chapter 18 with how this is done in bacteria. It is here that the basic mechanisms can most readily be appreciated. Thus, we deal with simple cases that illustrate different mechanisms of transcriptional regulation. These include the case of the *lac* operon, which is a group of genes that encode proteins needed for metabolism of the sugar lactose—genes that are transcribed only when that sugar is available in the growth medium. In this case we learn how genes can be activated (switched on) and repressed (switched off) in response to different signals. We then look at other examples: some where regulation is similar to the *lac* genes and some that illustrate rather different mechanisms of transcriptional regulation. Finally in this chapter, we describe how transcriptional regulation of alternative sets of genes in phage  $\lambda$  underpins the ability of that virus to choose between alternative development pathways upon infection of a bacterial cell.

In Chapter 19, we consider basic mechanisms of transcriptional regulation in eukaryotes, from yeast to higher eukaryotes. Mechanisms of transcriptional activation and repression are compared to those in bacteria, and we see where mechanisms are conserved and where there are additional features—most notably the effects of nucleosome positioning, remodeling, and modification as discussed in Chapter 8. We also discuss the meaning and mechanisms of so-called epigenetic gene regulation.

Up to this point, the regulation we have discussed is driven by protein regulators—activators and repressors, and proteins they recruit to genes. In Chapter 20, we look at regulatory RNAs. Here we describe how RNA molecules can activate, or more commonly repress, expression of genes in bacteria and eukaryotes. This includes long-understood mechanisms, such as attenuation of the tryptophan operon, and also more recently uncovered mechanisms, such as RNA interference and the role of microRNAs in higher eukaryotes.

In Chapter 21, we consider gene regulation in the context of developmental biology and evolution. We look at how genes are regulated to bestow cell type specificity (differentiation) and pattern formation (morphogenesis) on a group of genetically identical cells—for example, those found in a developing embryo. We also discuss diversity among closely related organisms and see how, in many of these, the differences in morphology or behavior result not from changes in the genes but from differences in where and when those genes are expressed within each organism during development. The most striking finding to come from whole-genome sequences is that

most animals (for example) have essentially the same genes—be they mice, men, or even flies. This observation again underscores the general role of gene regulation—most of it transcriptional regulation—in defining what each genome produces.

Consideration of gene regulatory networks in development leads us to the last chapter in this section of the book—Systems Biology. The field remains rather ill-defined and seems to embrace a range of different areas, but in the current context we focus on gene regulatory networks. Thus we present the nomenclature and basic ideas behind newly defined ways of thinking about how networks of genes are regulated. A new generation of molecular biologists—many with backgrounds in computing or physics—are describing such networks, using their own representations, in terms of the logic of information flow rather than molecular mechanisms that underlie their operation.

## PHOTOS FROM THE COLD SPRING HARBOR LABORATORY ARCHIVES

---



**Mark Ptashne and Joseph Goldstein, 1988 Symposium on Molecular Biology of Signal Transduction.** Ptashne was instrumental in taking the early ideas of Jacob and Monod about how gene expression is regulated, and describing how these work at a molecular level, first in phage  $\lambda$ , and then in yeast (Chapters 18 and 19). Goldstein, with his long-time collaborator Michael S. Brown, worked out the signal transduction pathways (Chapter 19) that control expression of genes involved in cholesterol metabolism, for which they won the 1985 Nobel Prize in Physiology or Medicine.



**Scott Emmons, Gary Ruvkun, and Barbara Meyer, 2004 Symposium on Epigenetics.** While studying the genetics of development in worms, Victor Ambros and Ruvkun identified the first miRNA and target gene (Chapter 20). The NASA T-shirt is a clue to another of Ruvkun's many research interests: the quest for life on Mars. Emmons studies behavior in worms, at all levels from gene expression to the neurobiology, and Meyer, who as a graduate student contributed much to elucidating the phage  $\lambda$  genetic switch (Chapter 18), now works on sex determination and dosage compensation in the worm (Chapter 20).



**John Gurdon and Ann McLaren, 1985 Symposium on Molecular Biology of Development.** Gurdon performed the first animal cloning experiment in 1962 when he transplanted the nucleus of an adult frog cell into an enucleated egg, from which arose a fully functional tadpole (Chapter 21). For this work he shared, with Shinya Yamanaka, the 2012 Nobel Prize in Physiology or Medicine. McLaren was an expert in mammalian genetics and reproductive biology, her research laying essential groundwork for the later development of in vitro fertilization, among other things. Her expertise in reproductive biology led to roles in policy matters as well, including as a member of the hugely influential Warnock Committee in the United Kingdom.



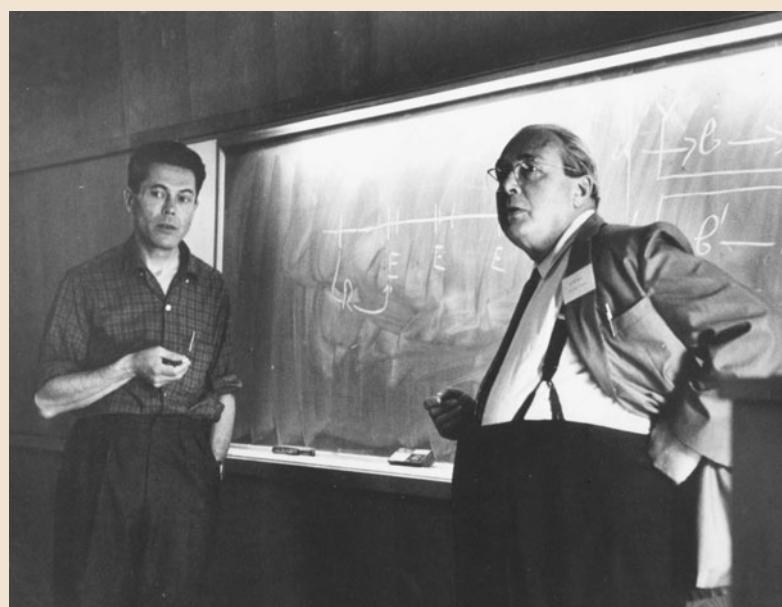
**Christiane Nüsslein-Volhard, 1996 Meeting on Zebrafish Development and Genetics.** Mutant screens carried out in fruit flies by Nüsslein-Volhard and her colleague Eric F. Wieschaus identified many genes critical to the early embryonic development of that organism, and probably all animals (Chapter 21). For this the two of them shared in the 1995 Nobel Prize in Physiology or Medicine with Edward B. Lewis.



**Mrs. I.H. Herskowitz with sons, Ira and Joel, 1947 Symposium on Nucleic Acids and Nucleoproteins.** Ira Herskowitz pioneered the use of the yeast *Saccharomyces cerevisiae* as a model organism for molecular biology (Appendix 1) and made major contributions to ideas about gene regulation in this organism as he had, earlier, in bacteriophage  $\lambda$  (Chapters 18 and 19). His father, Irwin, later the author of a genetics textbook, was attending the symposium that year.



**Richard Jorgensen and David Baulcombe, 2006 Symposium on Regulatory RNAs.** Jorgensen found that overexpression of the petunia pigment gene could generate flowers that had white rather than dark purple flowers (Chapter 20). Although unknown at the time, this effect was caused by RNAi. The small interfering RNAs—the critical intermediates in this process—were later identified by Baulcombe (Chapter 20).



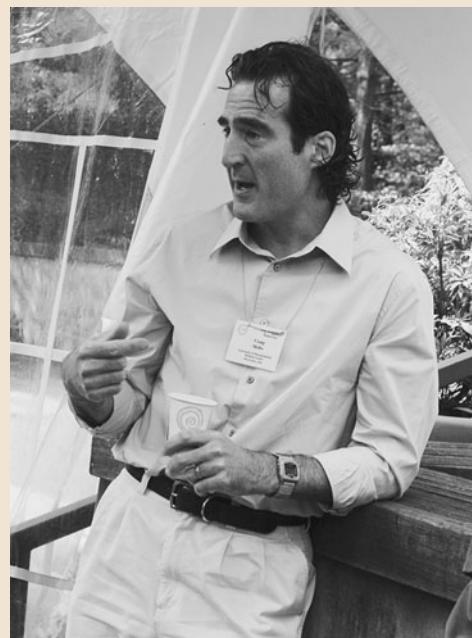
**Jacques Monod and Leo Szilard, 1961 CSH Laboratory.** Monod, together with Françoise Jacob, formulated the operon model for the regulation of gene expression (Chapter 18). The two of them, together with their colleague André Lwoff, shared the 1965 Nobel Prize in Physiology or Medicine for this achievement. Leo Szilard was a wartime nuclear physicist who turned to molecular biology after taking the phage course at Cold Spring Harbor in 1947. He ran a lab with Aaron Novick in Chicago. (Courtesy of Esther Bubley.)



**Shinya Yamanaka, 1994 CSHL course on Advanced In Situ Hybridization and Immunocytochemistry.** Yamanaka (third from left) attended this course as a student and is pictured with the other students and their instructors. With John Gurdon, Yamanaka won the 2012 Nobel Prize in Physiology or Medicine for the creation of iPS cells. He showed that expressing just four specific DNA-binding transcription factors was sufficient to drive differentiated cells into a dedifferentiated, pluripotent state. This striking experiment is described in Chapter 21.



**Jeffrey W. Roberts and Ann B. Burgess, 1970 Symposium on Transcription of Genetic Material.** Roberts' research has focused on regulators of gene expression in bacteria and phage, particularly antiterminators in phage  $\lambda$  (Chapter 18). Burgess became a biology educator and is involved in national efforts to improve science education. Roberts was an author of the fourth edition of this book, and Burgess has a cousin among the current authors (TB).

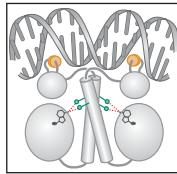


**Craig Mello, 2004 Symposium on Epigenetics.** Together with Andrew Fire, Mello found that by simply introducing dsRNAs into cells, genes with homology to that RNA can be silenced. From this observation, which they called RNA interference, the whole field of RNAi exploded (Chapter 20). They shared the 2006 Nobel Prize in Physiology or Medicine for their work.



**Edward B. Lewis, Carl C. Lindegren, Alfred D. Hershey, and Joshua Lederberg, 1951 Symposium on Genes and Mutations.** Lewis instigated the genetic analysis of development, using the fruit fly as his model (Chapter 21). He shared, with Eric F. Wieschaus and Christiane Nüsslein-Volhard, the 1995 Nobel Prize in Physiology or Medicine for his work. Lindegren was a pioneer of yeast genetics (Appendix 1). Hershey was, together with Max Delbrück and Salvador E. Luria, the leader of the group that used phage as their model system in the early days of molecular biology (Appendix 1); the three of them shared the 1969 Nobel Prize in Physiology or Medicine. Lederberg discovered that DNA could pass between bacteria by a mating process called conjugation (Appendix 1), for which he shared, with George Beadle and Edward Tatum, in the 1958 Nobel Prize in Physiology or Medicine.

CHAPTER 18



# Transcriptional Regulation in Prokaryotes

## OUTLINE

**I**N CHAPTER 13, WE SAW HOW DNA IS TRANSCRIBED into RNA by the enzyme RNA polymerase. We also described the sequence elements that constitute a promoter—the region at the start of a gene where the enzyme binds and initiates transcription. In bacteria, the most common form of RNA polymerase (that bearing  $\sigma^{70}$ ) recognizes promoters formed from various sequence elements—the three major ones being “−10,” “−35,” and “UP” elements—and we saw that the strength of any given promoter is determined by which elements it possesses and how well they match optimum “consensus” sequences. In the absence of regulatory proteins, these elements determine the efficiency with which polymerase binds to the promoter and, once bound, how readily it initiates transcription.

Now we turn to the mechanisms that regulate expression—that is, those mechanisms that increase or decrease expression of a given gene as the requirement for its product varies. There are various stages at which expression of a gene can be regulated. The most common is transcription initiation, and the bulk of this chapter focuses on the regulation of that step in bacteria. We start with an overview of general mechanisms and principles and proceed to some well-studied examples that demonstrate how the basic mechanisms are used in various combinations to control genes in specific biological contexts. We also consider mechanisms of transcriptional regulation that operate at steps after initiation, specifically during elongation and termination. Other examples of transcriptional regulation in prokaryotes—those mediated by RNA—are considered in Chapter 20, Regulatory RNAs. An example of prokaryotic gene regulation at the level of translation was discussed in Chapter 15.

### Principles of Transcriptional Regulation, 615

### Regulation of Transcription Initiation: Examples from Prokaryotes, 620

### The Case of Bacteriophage λ: Layers of Regulation, 636

### Visit Web Content for Structural Tutorials and Interactive Animations

## PRINCIPLES OF TRANSCRIPTIONAL REGULATION

### Gene Expression Is Controlled by Regulatory Proteins

Genes are very often controlled by extracellular signals; in the case of bacteria, this typically means molecules present in the growth medium. These

signals are communicated to genes by regulatory proteins, which come in two types: positive regulators, or **activators**, and negative regulators, or **repressors**. Typically, these regulators are DNA-binding proteins that recognize specific sites at or near the genes they control. An activator increases transcription of the regulated gene, and repressors decrease or eliminate that transcription.

How do these regulators work? Recall the steps in transcription initiation described in Chapter 13 (see Fig. 13-3). First, RNA polymerase binds to the promoter in a closed complex (in which the DNA strands remain together). The polymerase–promoter complex then undergoes a transition to an open complex in which the DNA at the start site of transcription is unwound and the polymerase is positioned to initiate transcription. This is followed by promoter escape, the step in which polymerase leaves the promoter and starts transcribing. Polymerase then proceeds through the elongation phase before finally terminating. Which steps are stimulated by activators and inhibited by repressors depends on the promoter and regulators in question.

### Most Activators and Repressors Act at the Level of Transcription Initiation

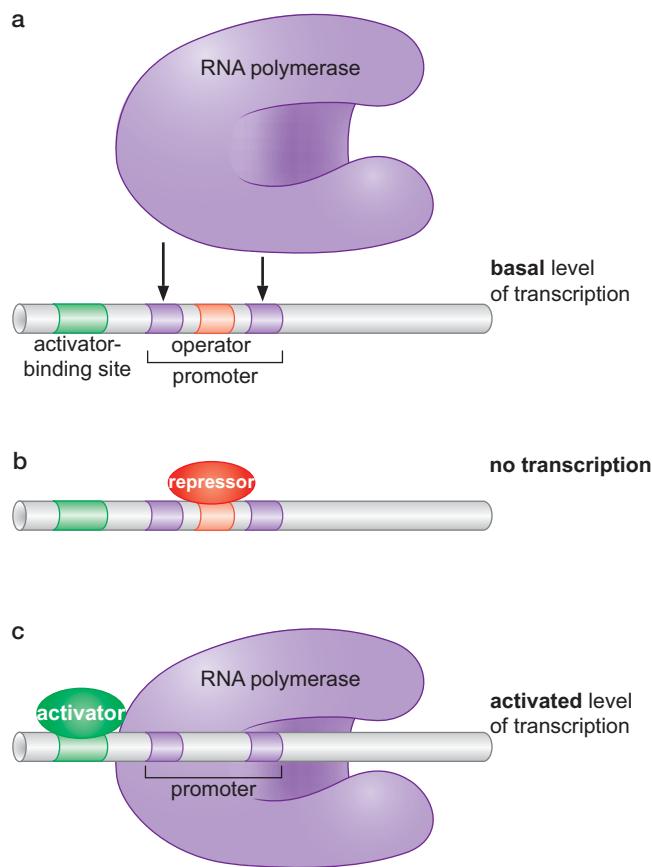
Although we shall see cases where gene expression is regulated at essentially every step from the gene to its product, the most common step at which regulation impinges is the initiation of transcription—the focus of this chapter. There are two reasons why this might make sense. First, transcription initiation is the most energetically efficient step to regulate. By this we mean that deciding whether or not to express a gene at the first step ensures that no energy or resources are wasted making, for example, part or all of an mRNA that will not then be used (e.g., be translated). Second, regulation at this first step is easier to do well. There is only a single copy of each gene (in a haploid genome), and so typically only a single promoter on a single DNA molecule must be regulated to control expression of a given gene. In contrast, to regulate that gene at the point of translation, for example, each of several mRNA molecules must be acted on.

Why then is not all regulation focused on the step of transcription initiation? Regulating later steps can have two advantages. First, it allows for more inputs: if a gene is regulated at more than one step, more signals can modulate its expression, or the same signals can do so even more effectively. Second, regulation at steps later than transcription initiation can reduce the response time. Thus, consider again the example of translational regulation (see Fig. 15-44 for an example). If a signal relieves repression of this step, the protein product encoded by the gene will be produced immediately upon receipt of that signal. This reduced response time might obviously be advantageous in some situations. But, as we have said, it is the initiation of transcription that is most often regulated, and we now consider, in general terms, how activators and repressors regulate transcription initiation (see Interactive Animation 18-1).



### Many Promoters Are Regulated by Activators That Help RNA Polymerase Bind DNA and by Repressors That Block That Binding

At many promoters, in the absence of regulatory proteins, RNA polymerase binds only weakly. This is because one or more of the promoter elements



**FIGURE 18-1** Activation by recruitment of RNA polymerase. (a) In the absence of both activator and repressor, RNA polymerase occasionally binds the promoter spontaneously and initiates a low level (basal level) of transcription. (b) Binding of the repressor to the operator sequence blocks binding of RNA polymerase and so inhibits transcription. (c) Recruitment of RNA polymerase by the activator gives high levels of transcription. RNA polymerase is shown recruited in the closed complex (see Fig. 13-3). It then spontaneously isomerizes to the open complex and initiates transcription. If both the repressor and activator are present and functional, the action of the repressor typically overcomes that of the activator. (This case is not shown in the figure.)

discussed above is absent or imperfect. When polymerase does occasionally bind, however, it spontaneously undergoes a transition to the open complex and initiates transcription. This gives a low level of **constitutive** expression called the **basal** level. Binding of RNA polymerase is the rate-limiting step in this case (Fig. 18-1a).

To control expression from such a promoter, a repressor need only bind to a site overlapping the region bound by polymerase. In that way, the repressor blocks polymerase binding to the promoter, thereby preventing transcription (Fig. 18-1b), although it is important to note that repression can work in other ways as well. The site on DNA where a repressor binds is called an **operator**.

To activate transcription from this promoter, an activator can just help the polymerase bind the promoter. Typically, this is achieved as follows: the activator uses one surface to bind to a site on the DNA near the promoter; with another surface, the activator simultaneously interacts with RNA polymerase, bringing the enzyme to the promoter (Fig. 18-1c). This mechanism, often called **recruitment**, is an example of **cooperative binding** of proteins to DNA (a process we describe in more detail later, particularly in Box 18-4). The interactions between the activator and polymerase, and between activator and DNA, serve merely “adhesive” roles: the enzyme is active and the activator simply brings it to the nearby promoter. Once there, it spontaneously isomerizes to the open complex and initiates transcription.

The *lac* genes of *Escherichia coli* are transcribed from a promoter that is regulated by an activator and a repressor working in the simple way outlined above. We describe this case in detail later in the chapter.

## Some Activators and Repressors Work by Allostery and Regulate Steps in Transcriptional Initiation after RNA Polymerase Binding

Not all promoters are limited in the same way. Thus, consider another class of promoter in which RNA polymerase binds efficiently unaided and forms a stable closed complex. But that closed complex does not spontaneously undergo transition to the open complex (Fig. 18-2a). At this promoter, an activator must stimulate the transition from a closed to open complex, since that transition is the rate-limiting step.

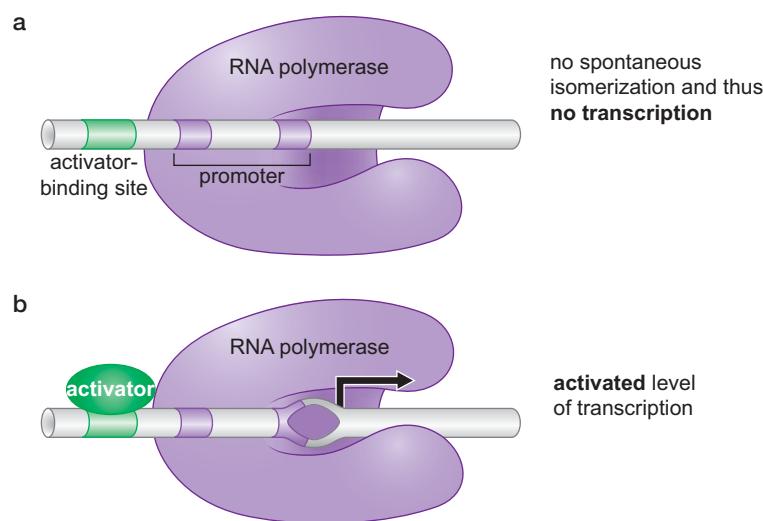
Activators that stimulate this kind of promoter work by triggering a conformational change in either RNA polymerase or DNA; that is, they interact with the stable closed complex and induce a conformational change that causes transition to the open complex (Fig. 18-2b). This mechanism is an example of **allostery**.

In Chapter 6, we encountered allostery as a general mechanism for controlling the activities of proteins. In this chapter, we shall see two examples of transcriptional activators working by allostery. In one case (at the *glnA* promoter), the activator (NtrC) interacts with the RNA polymerase bound in a closed complex at the promoter, stimulating transition to the open complex. In the other example (at the *merT* promoter), the activator (MerR) achieves the same effect but does so by inducing a conformational change in the promoter DNA. In still another class of promoter, transcription initiation is limited at the step of promoter escape (see Fig. 12-3). One example of such a promoter directs expression of the *malT* gene. In the absence of an activator, it undergoes abortive initiation, and only in the presence of an activator will it efficiently escape into elongation.

In a similar vein, repressors can work in ways other than just blocking the binding of RNA polymerase. For example, some repressors interact with polymerase at the promoter and inhibit transition to the open complex, or promoter escape. We see examples of these later in the chapter (e.g., the Gal repressor).

### Action at a Distance and DNA Looping

Thus far we have tacitly assumed that DNA-binding proteins that interact with each other bind to adjacent sites (e.g., RNA polymerase and activator in Figs. 18-1 and 18-2). This is often the case. But some proteins interact



**FIGURE 18-2** Allosteric activation of RNA polymerase. (a) Binding of RNA polymerase to the promoter in a stable closed complex. (b) The activator interacts with polymerase to trigger transition to the open complex and high levels of transcription. The representations of the closed and open complexes are shown diagrammatically; for a more complete description of those states, see Chapter 13, Figure 13-3.

with each other even when bound to sites well separated on the DNA. To accommodate this interaction, the DNA between the sites loops out, bringing the sites into proximity with one another (Fig. 18-3).

We will encounter examples of this kind of interaction in bacteria. We will consider repressors that interact to form DNA loops of up to 3 kb. In the next chapter—on eukaryotic transcriptional regulation—we are faced with more numerous and more dramatic examples of this “action at a distance.”

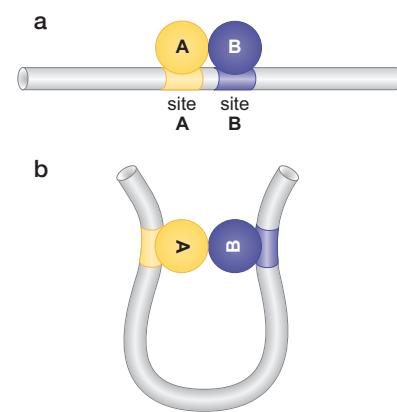
Distant DNA sites can be brought closer together to help loop formation. In bacteria, for example, there are cases in which a protein binds between an activator-binding site and the promoter and helps the activator interact with polymerase by bending the DNA in a favorable direction (Fig. 18-4). There are also cases where such a protein hinders loop formation and activation by bending the DNA in an unfavorable direction. Such “architectural” proteins facilitate (or hinder) interactions between proteins in other processes as well (e.g., site-specific recombination; see Chapter 12).

### Cooperative Binding and Allostery Have Many Roles in Gene Regulation

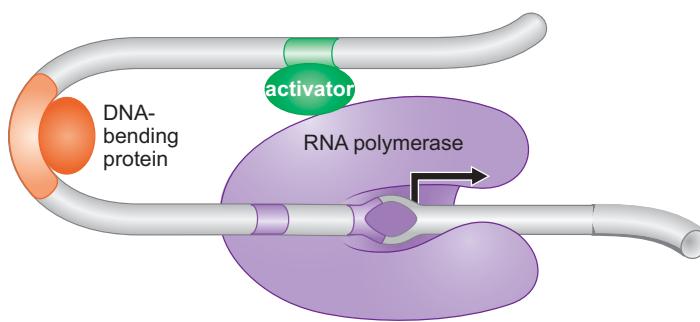
We have already pointed out that gene activation can be mediated by simple cooperative binding: the activator interacts simultaneously with DNA and with polymerase and so recruits the enzyme to the promoter. We have also described how activation can, in other cases, be mediated by allosteric events: an activator interacts with polymerase already bound to the promoter and, by inducing a conformational change in the enzyme or the promoter, stimulates transcription initiation. Both cooperative binding and allostery have additional roles in gene regulation.

For example, groups of regulators often bind DNA cooperatively: two or more activators and/or repressors interact with each other and with DNA and thereby help each other bind near a gene they all regulate. As we shall see, this kind of interaction can produce sensitive switches that allow a gene to go from completely off to fully on in response to only small changes in conditions. Cooperative binding of activators can also serve to integrate signals: some genes are activated only when multiple signals (and thus multiple regulators) are simultaneously present. A particularly striking and well-understood example of cooperativity in gene regulation is provided by bacteriophage  $\lambda$ , discussed in detail later in the chapter. The basic mechanism and consequences of cooperative binding are considered in more detail when we discuss that example later in the chapter and also in Box 18-3.

Allostery, for its part, is not only a mechanism of gene activation, but also often the way regulators are controlled by their specific signals. Thus, a



**FIGURE 18-3** Interactions between proteins bound to DNA. (a) Cooperative binding of proteins to adjacent sites. (b) Cooperative binding of proteins to separated sites.



**FIGURE 18-4** A DNA-bending protein can facilitate interaction between distantly bound DNA-binding proteins. A protein that bends DNA binds to a site between the activator-binding site and the promoter. If the direction of the bend is favorable, this action brings the two sites closer together in space and thereby helps the interaction between the DNA-bound activator and polymerase. If the bend is unfavorable, it has the opposite effect.

typical bacterial regulator can adopt two conformations: in one, it can bind DNA; in the other, it cannot. Binding of a signal molecule locks the regulatory protein in one or another conformation, thereby determining whether or not it can act. An example of this was seen in Chapter 6 (Fig. 6-19), where we also considered the basic mechanism of allostery in some detail. In this and the next chapter, we will see several examples of allosteric control of regulators by their signals.

### Antitermination and Beyond: Not All of Gene Regulation Targets Transcription Initiation

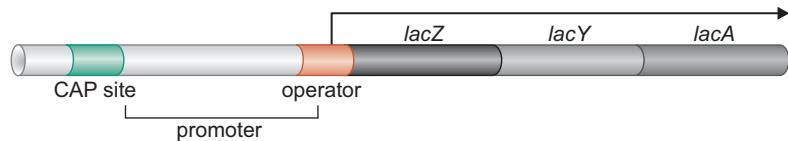
As stated at the beginning of this chapter, the bulk of gene regulation takes place at the initiation of transcription. This is true in eukaryotes just as it is in bacteria. But regulation is certainly not restricted to that step in either class of organism. In this chapter, we will see examples in bacteria of gene regulation at the level of transcriptional elongation and termination. Other examples of gene regulation in bacteria are found in Chapter 15, where we discuss an example of the regulation of translation of ribosomal protein genes, and in Chapter 20, where we consider cases involving regulation by RNAs (e.g., attenuation, riboswitches, and small RNAs). Some of these RNA cases involve regulation of transcription and others involve regulation of translation.

## REGULATION OF TRANSCRIPTION INITIATION: EXAMPLES FROM PROKARYOTES

Having outlined basic principles of transcriptional regulation, we turn to some examples that show these principles in action. First, we consider the genes involved in lactose metabolism in *E. coli*. Here, we see how an activator and a repressor regulate expression in response to two signals. We also describe some of the experiments that reveal how these regulators work.

### An Activator and a Repressor Together Control the *lac* Genes

The three *lac* genes—*lacZ*, *lacY*, and *lacA*—are arranged adjacently on the *E. coli* genome and are together called the ***lac operon*** (Fig. 18-5). The *lac* promoter, located at the 5' end of *lacZ*, directs transcription of all three genes as a single mRNA (called a polycistronic message because it includes more than one gene); this mRNA is translated to give the three

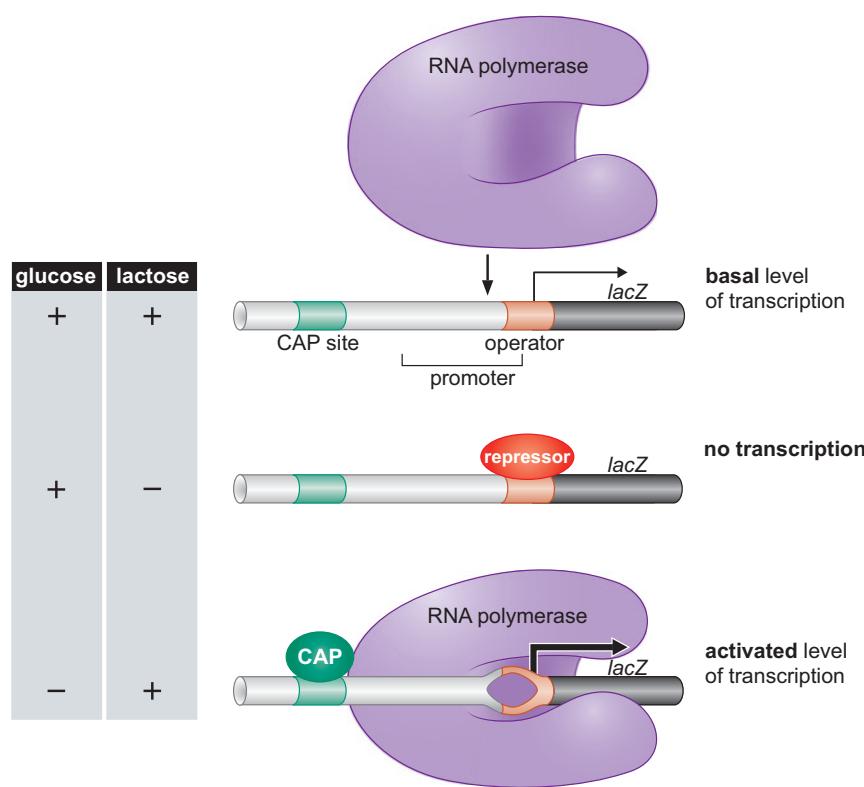


**FIGURE 18-5** The *lac operon*. The three genes (*lacZ*, *lacY*, and *lacA*) are transcribed as a single mRNA from the promoter (as indicated by the arrow). The CAP site and the operator (the site bound by Lac repressor) are each about 20 bp. The operator lies within the region bound by RNA polymerase at the promoter, and the CAP site lies just upstream of the promoter (see Fig. 18-8 for more details of the relative arrangements of these binding sites and the text for a description of the proteins that bind to them). The picture is simplified in that there are two additional, weaker, *lac* operators located nearby (see Fig. 18-12), but we do not need to consider those at present.

protein products. The *lacZ* gene encodes the enzyme  $\beta$ -galactosidase, which cleaves the sugar lactose into galactose and glucose, both of which are used by the cell as energy sources. The *lacY* gene encodes the lactose permease, a protein that inserts into the cell membrane and transports lactose into the cell. The *lacA* gene encodes thiogalactoside transacetylase, which rids the cell of toxic thiogalactosides that also get transported in by *lacY*.

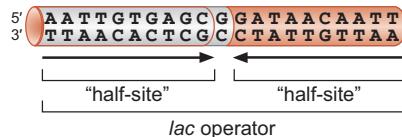
These genes are expressed at high levels only when lactose is available, and glucose—the preferred energy source—is not. Two regulatory proteins are involved: one is an activator called **CAP**, and the other is a repressor called the **Lac repressor**. The Lac repressor is encoded by the *lacI* gene, which is located near the other *lac* genes, but transcribed from its own (constitutively expressed) promoter. The name CAP stands for catabolite activator protein, but this activator is also known as **CRP** (for cAMP receptor protein, for reasons that will be explained later). The gene encoding CAP is located elsewhere on the bacterial chromosome, not linked to the *lac* genes. Both CAP and the Lac repressor are DNA-binding proteins and each binds to a specific site on DNA at or near the *lac* promoter (the CAP site and the operator, respectively; see Fig. 18-5).

Each of these regulatory proteins responds to one environmental signal and communicates it to the *lac* genes. Thus, CAP mediates the effect of glucose, whereas Lac repressor mediates the lactose signal. This regulatory system works in the following way (and as shown in Fig. 18-6). Lac repressor can bind DNA and repress transcription only in the absence of lactose. In the presence of that sugar, the repressor is inactive and the genes derepressed (expressed). CAP can bind DNA and activate the *lac* genes only in the *absence* of glucose. Thus, the combined effect of these two regulators ensures that the genes are expressed at significant levels only when lactose is present and glucose absent.



**FIGURE 18-6 Expression of the *lac* genes.** The presence or absence of the sugars lactose and glucose control the level of expression of the *lac* genes. High levels of expression require the presence of lactose (and hence the absence of functional Lac repressor) and absence of the preferred energy source, glucose (and hence presence of the activator CAP). When bound to the operator, Lac repressor excludes polymerase whether or not active CAP is present. CAP and Lac repressor are shown as single units, but CAP actually binds DNA as a dimer, and Lac repressor binds as a tetramer (see Fig. 18-12). CAP recruits polymerase to the *lac* promoter where it spontaneously undergoes isomerization to the open complex (the state shown in the bottom line).

### CAP and Lac Repressor Have Opposing Effects on RNA Polymerase Binding to the *lac* Promoter



**FIGURE 18-7** The symmetric half-sites of the *lac* operator.

As we have seen, the site bound by the Lac repressor is called the *lac operator*. This 21-bp sequence is twofold symmetric and is recognized by two subunits of Lac repressor, one binding to each half-site (see Fig. 18-7). We discuss that binding in more detail later in this chapter, in the section CAP and Lac Repressor Bind DNA Using a Common Structural Motif. How does the repressor, when bound to the operator, repress transcription?

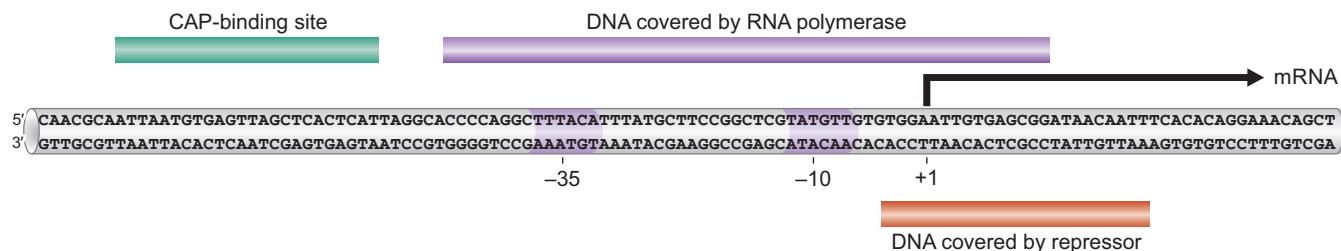
The *lac* operator overlaps the promoter, and so the repressor bound to the operator physically prevents RNA polymerase from binding to the promoter and thus initiating RNA synthesis (see Fig. 18-8). Protein-binding sites in DNA can be identified, and their location mapped, using DNA-footprinting and gel-mobility assays as described in Chapter 7.

As we have seen, RNA polymerase binds the *lac* promoter poorly in the absence of CAP, even when there is no functional repressor present. This is because the sequence of the  $-35$  region of the *lac* promoter is not optimal for its binding, and the promoter lacks an UP-element (see Fig. 13-5, Box 13-1, and Fig. 18-8). This is typical of promoters that are controlled by activators.

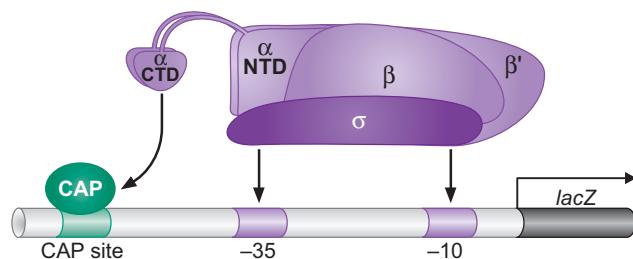
CAP binds as a dimer to a site similar in length to that of the *lac* operator, but different in sequence. This site is located some 60 bp upstream of the start site of transcription (see Fig. 18-8). When CAP binds to that site, the activator helps polymerase bind to the promoter by interacting with the enzyme and recruiting it to the promoter (see Fig. 18-6). This cooperative binding stabilizes the binding of polymerase to the promoter. We now look at CAP-mediated activation in more detail.

### CAP Has Separate Activating and DNA-Binding Surfaces

Various experiments support the view that CAP activates the *lac* genes by simple recruitment of RNA polymerase. Mutant versions of CAP have been isolated that bind DNA but do not activate transcription. The existence of these so-called **positive control (pc)** mutants demonstrates that to activate transcription, the activator must do more than simply bind DNA near the promoter. Thus, activation is not caused by, for example, the activator changing local DNA structure. The amino acid substitutions in the positive control mutants identify the region of CAP that touches polymerase, called the **activating region**.



**FIGURE 18-8** The control region of the *lac* operon. The nucleotide sequence and organization of the *lac* operon control region are shown. The colored bars above and below the DNA show regions covered by RNA polymerase and the regulatory proteins. Note that the Lac repressor covers more DNA than that sequence defined as the minimal operator-binding site and RNA polymerase more than that defined by the sequences that make up the promoter.



Where does the activating region of CAP touch RNA polymerase when activating the *lac* genes? This site is revealed by mutant forms of polymerase that can transcribe most genes normally, but cannot be activated by CAP at the *lac* genes. These mutants have amino acid substitutions in the **carboxy-terminal domain (CTD)** of the  **$\alpha$  subunit** of RNA polymerase. As we saw in Chapter 13, this domain is attached to the amino-terminal domain (NTD) of  $\alpha$  by a flexible linker. The  $\alpha$ NTD is embedded in the body of the enzyme, but the  $\alpha$ CTD extends out from it and binds the UP-element of the promoter (when that element is present) (see Fig. 13-7).

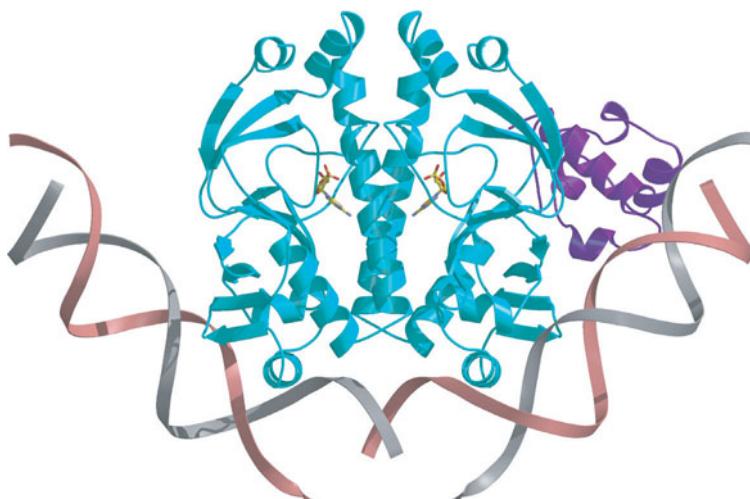
At the *lac* promoter, where there is no UP-element,  $\alpha$ CTD binds to CAP and adjacent DNA instead (Fig. 18-9). This picture is supported by a crystal structure of a complex containing CAP,  $\alpha$ CTD, and a DNA oligonucleotide duplex containing a CAP site and an adjacent UP-element (Fig. 18-10; see also Structural Tutorial 18-1). In Box 18-1, Activator Bypass Experiments, we describe an experiment showing that activation of the *lac* promoter requires no more than polymerase recruitment.

Having seen how CAP activates transcription at the *lac* operon, and how the Lac repressor counters that effect, we now look more closely at how these regulators recognize their DNA-binding sites.

### CAP and Lac Repressor Bind DNA Using a Common Structural Motif

X-ray crystallography has been used to determine the structural basis of DNA binding for a number of bacterial activators and repressors, including CAP and the Lac repressor. Although the details differ, the basic mechanism of DNA recognition is similar for most bacterial regulators.

**FIGURE 18-9 Activation of the *lac* promoter by CAP.** RNA polymerase binding at the *lac* promoter with the help of CAP. CAP is recognized by the CTDs of the  $\alpha$  subunits. The  $\alpha$ CTDs also contact DNA, adjacent to the CAP site, when interacting with CAP. As discussed in Chapter 13, we use this representation of RNA polymerase when indicating specific points of contact between an activator and its target site on polymerase, or between regions of polymerase and the promoter.



**FIGURE 18-10 Structure of CAP- $\alpha$ CTD-DNA complex.** CAP is shown (in turquoise) bound as a dimer to its site on DNA. In addition, the  $\alpha$ CTD of RNA polymerase is shown (in purple) bound to an adjacent stretch of DNA and interacting with CAP. The site of interaction on each protein involves the residues identified genetically. One molecule of cAMP is shown bound to each monomer of CAP. (Benoff B. et al. 2002. *Science* 297: 1562.) Image prepared with MolScript, BobScript, and Raster3D.

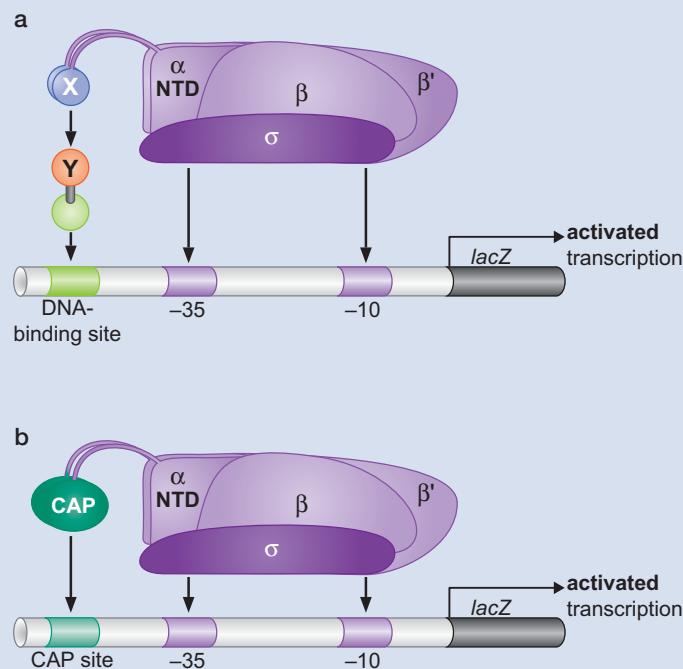
## ► KEY EXPERIMENTS

**Box 18-1 Activator Bypass Experiments**

If an activator has only to recruit polymerase to the gene, then other methods of bringing the polymerase to the gene should work just as well. This turns out to be true of the *lac* genes, as shown by the following experiments (Box 18-1 Fig. 1).

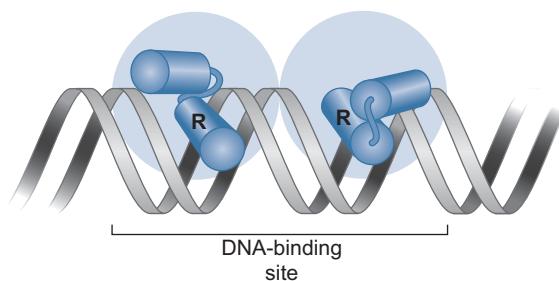
In one experiment, another protein–protein interaction is used in place of that between CAP and polymerase. This is done by taking two proteins known to interact with each other, attaching one to a DNA-binding domain, and, with the other, replacing the carboxy-terminal domain of the polymerase  $\alpha$  subunit ( $\alpha$ CTD). The modified polymerase can be activated by the makeshift “activator” as long as the appropriate DNA-binding site is introduced near the promoter. In another experiment, the  $\alpha$ CTD of polymerase is replaced with a DNA-binding domain (e.g., that of CAP). This modified polymerase efficiently initiates transcription from the *lac* promoter in the

absence of any activator, as long as the appropriate DNA-binding site is placed nearby. A third experiment is even simpler: polymerase can transcribe the *lac* genes at high levels in vitro in the absence of any activator if the enzyme is present at high concentration. So we see that either recruiting polymerase artificially or supplying it at a high concentration is sufficient to produce activated levels of expression of the *lac* genes. These experiments are consistent with the activator having only to help polymerase bind to the promoter. For an explanation of why simply increasing the concentration of a protein (e.g., RNA polymerase) helps it bind to a site on DNA (in this case the promoter), see Box 18-4. The results discussed in this box would not be expected if the activator had to induce a specific allosteric change in either polymerase or DNA to activate transcription.



**BOX 18-1 FIGURE 1** Two activator bypass experiments. (a) The  $\alpha$ CTD is replaced by a protein X, which interacts with protein Y. Protein Y is fused to a DNA-binding domain, and the site recognized by that domain is shown placed near the *lac* genes. (b) The  $\alpha$ CTD is replaced by the DNA-binding portion of CAP.

In the typical case, the protein binds as a homodimer to a site that is an inverted repeat (or near repeat). One monomer binds each half-site, with the axis of symmetry of the dimer lying over that of the binding site (as for the Lac repressor, Fig. 18-7). Recognition of specific DNA sequences is achieved using a conserved region of secondary structure called a **helix-turn-helix** (Fig. 18-11). This motif is composed of two  $\alpha$  helices, one of which—the **recognition helix**—fits into the major groove of the DNA. As discussed in Chapter 6, an  $\alpha$  helix is just the right size to fit into the major groove, allowing amino



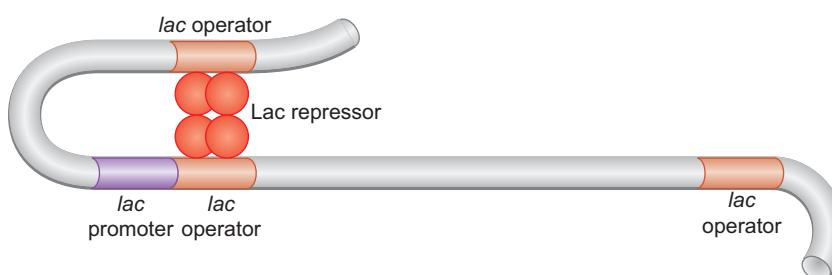
**FIGURE 18-11** Binding of a protein with a helix-turn-helix motif to DNA. The protein, as is typically the case, binds as a dimer, and the two subunits are indicated by the shaded circles. The helix-turn-helix motif on each monomer is indicated; the “recognition helix” is labeled R.

acid residues on its outer face to interact with chemical groups on the edges of base pairs. And we saw how each base pair presents a characteristic pattern of hydrogen bonding acceptors and donors (Fig. 6-14). Thus, a protein can distinguish different DNA sequences in this way without unwinding the DNA duplex. Figure 6-13 illustrates an example of the interactions made by a given recognition helix and its DNA-binding site.

The second helix of the helix-turn-helix motif sits across the major groove and makes contact with the DNA backbone, ensuring proper presentation of the recognition helix and at the same time adding binding energy to the overall protein–DNA interaction.

This description is essentially true not only for CAP (see Fig. 18-10) and the Lac repressor, but for many other bacterial regulators as well. These include the bacteriophage  $\lambda$  repressor (the example shown in Fig. 6-13) and  $\lambda$  Cro proteins we encounter in a later section, as well as the repressors of related lambdoid phages (e.g., that of phage 434 [see Structural Tutorial 18-2]). Despite this, there are differences in detail, as the following examples illustrate.

- Lac repressor binds as a tetramer, not a dimer. Nevertheless, each operator is contacted by only two of these subunits. Thus, the different oligomeric form does not alter the mechanism of DNA recognition. The other two monomers within the tetramer can bind one of two other lac operators, located 400 bp downstream and 90 bp upstream of the primary operator. In such cases, the intervening DNA loops out to accommodate the reaction (Fig. 18-12).
- In some cases, other regions of the protein, outside the helix-turn-helix motif, also interact with the DNA. The  $\lambda$  repressor, for example,



**FIGURE 18-12** Lac repressor binds as a tetramer to two operators. The loop shown is between the Lac repressor bound at the primary operator and the upstream auxiliary one. A similar loop can alternatively form with the downstream operator. The primary operator—the one shown against the promoter—is the operator referred to in discussion of regulation of *lac* gene expression. In this figure, each repressor dimer is shown as two circles, rather than as a single oval (as used in earlier figures) to emphasize its oligomeric structure.

makes additional contacts using amino-terminal arms. These reach around the DNA and interact with the minor groove on the back face of the helix.

- In many cases, binding of the protein does not alter the structure of the DNA. In some cases, however, various distortions are seen in the protein–DNA complex. For example, CAP induces a dramatic bend in the DNA, partially wrapping it around the protein. This is caused by other regions of the protein, outside the helix-turn-helix motif, interacting with sequences outside the DNA segment recognized by the helix-turn-helix motif. In other cases, binding results in twisting of the DNA site.

Not all prokaryotic repressors bind using a helix-turn-helix. A few have been described that employ quite different approaches. A striking example is the Arc repressor from phage P22 (a phage related to  $\lambda$  but that infects *Salmonella*). The Arc repressor binds as a dimer to an inverted repeat operator, but instead of an  $\alpha$  helix, it recognizes its binding site using two antiparallel  $\beta$  strands inserted into the major groove.

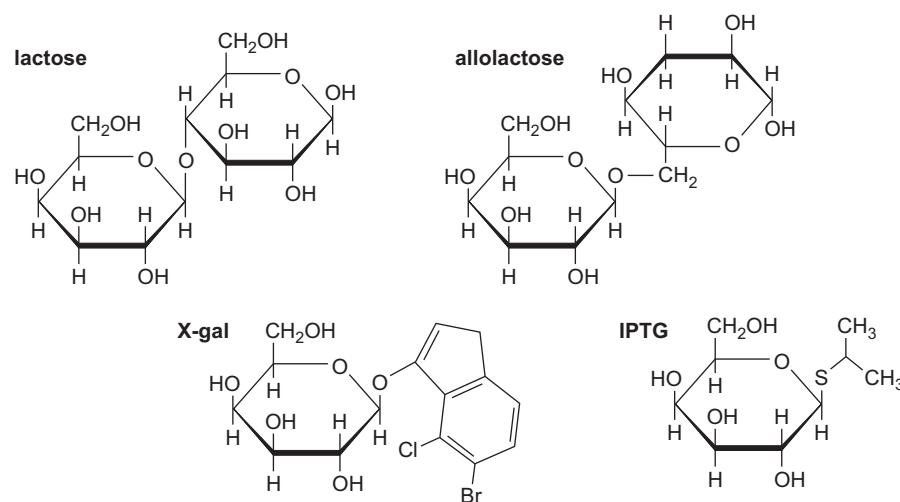
### The Activities of Lac Repressor and CAP Are Controlled Allosterically by Their Signals

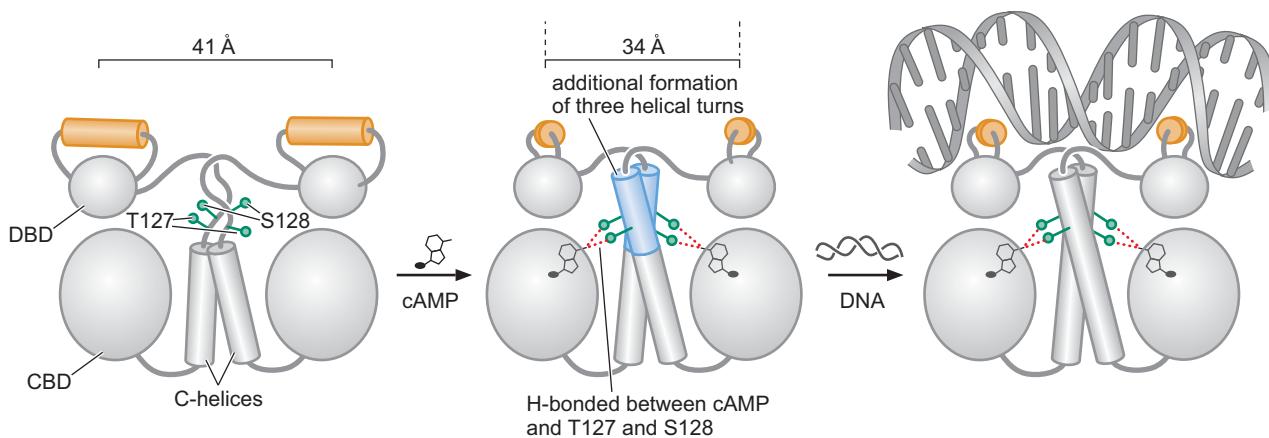
When lactose enters the cell, it is converted to allolactose. It is allolactose (rather than lactose itself) that controls the Lac repressor (Fig. 18-13). Paradoxically, the conversion of lactose to allolactose is catalyzed by  $\beta$ -galactosidase, itself encoded by one of the *lac* genes. How is this possible?

The answer is that expression of the *lac* genes is leaky: even when they are repressed, an occasional transcript gets made. This happens because every so often, RNA polymerase will manage to bind the promoter in place of the Lac repressor. This leakiness ensures that there is a low level of  $\beta$ -galactosidase in the cell even in the absence of lactose, and so there is enzyme poised to catalyze the conversion of lactose to allolactose.

Allolactose binds to the Lac repressor and triggers a change in the shape (conformation) of that protein. In the absence of allolactose, the repressor is present in a form that binds its site on DNA (and so keeps the *lac* genes switched off). Once allolactose has altered the shape of the repressor, the protein can no longer bind DNA, and so the *lac* genes are no longer repressed. In Chapter 6, we described the structural basis of this allosteric change in the

**FIGURE 18-13** The inducer of the *lac* operon is allolactose. Lactose is converted to allolactose, and it is this species that is the direct inducer of the *lac* genes—the molecule that binds Lac repressor and induces the conformational change that stops it from binding DNA (see Fig. 6-20). Other synthetic molecules can also act as inducers, most notably (because often used) is IPTG (isopropyl  $\beta$ -D-thiogalactopyranoside). Other molecules can act as substrates for  $\beta$ -galactosidase, but not as inducers. The most notable of these is X-gal (5-bromo-4-chloro-3-indolyl- $\beta$ -D-galactopyranoside). When acted on by that enzyme, X-gal releases a blue color that makes it useful as substrate in assays.





**FIGURE 18-14** Mechanism of allosteric control of CAP. The crystal structures of CAP in three states have been determined: CAP alone, CAP bound to cAMP, and CAP–cAMP–DNA complex. From these it is clear how binding of cAMP (to the cAMP binding domains [CBDs] of CAP) causes a structural change in the protein that results in its DNA-binding domains being realigned to an optimum configuration for DNA recognition. This happens because cAMP bound to CBPs also makes hydrogen bonds to two residues in a coiled region of the protein, triggering that region to form a helix (shown in blue). The recognition helices of CAP’s DNA-binding domain are shown in orange. (Adapted, with permission, from Popovych N. et al. 2009. Proc. Natl. Acad. Sci. **106**: 6927–6932, Fig. 7, p. 6931.)

Lac repressor (Fig. 6-19). An important point to emphasize is that allolactose binds to a part of the Lac repressor distinct from its DNA-binding domain.

CAP activity is regulated in a similar manner (Fig. 18-14). Glucose lowers the intracellular concentration of a small molecule, cAMP. This molecule is the allosteric effector for CAP: only when CAP is complexed with cAMP does the protein adopt a conformation that binds DNA (thus also explaining CAP’s alternative name, CRP). And so, only when glucose levels are low (and cAMP levels high) does CAP bind DNA and activate the *lac* genes. The part of CAP that binds the effector, cAMP, is separate from the part of the protein that binds DNA.

The *lac* operon of *E. coli* is one of the two systems used by French biologists François Jacob and Jacques Monod in formulating the early ideas about gene regulation. In Box 18-2, Jacob, Monod, and the Ideas behind Gene Regulation, we provide a brief description of those early studies and why the ideas they generated have proved so influential.

### Combinatorial Control: CAP Controls Other Genes As Well

The *lac* genes provide an example of **signal integration**: their expression is controlled by two signals, each of which is communicated to the genes via a separate regulator—the Lac repressor and CAP, respectively.

Consider another set of *E. coli* genes, the *gal* genes. These genes encode enzymes involved in galactose metabolism. As with the *lac* genes, the *gal* genes are only expressed when their substrate sugar, in this case galactose, is present, and the preferred energy source, glucose, is absent. Again, analogous to *lac*, the two signals are communicated to the genes via two regulators—an activator and a repressor. The repressor, encoded by the *galR* gene, mediates the effects of the inducer galactose, but the activator of the *gal* genes is again CAP. Thus, a regulator (CAP) works together with different repressors at different genes. This is an example of **combinatorial control**. In fact, CAP acts at more than 100 genes in *E. coli*, working with an array of partners.

## ► KEY EXPERIMENTS

## Box 18-2 Jacob, Monod, and the Ideas behind Gene Regulation

The idea that the expression of a gene can be controlled by the product of another gene—that there exist regulatory genes the sole function of which is regulating the expression of other genes—was one of the great insights from the early years of molecular biology. It was proposed by a group of scientists working in Paris in the 1950s and early 1960s, in particular François Jacob and Jacques Monod. They sought to explain two apparently unrelated phenomena: the appearance of  $\beta$ -galactosidase in *E. coli* grown in lactose, and the behavior of the bacterial virus (bacteriophage)  $\lambda$  upon infection of *E. coli*. Their work culminated in publication of their operon model in 1961 (and the 1965 Nobel Prize in Physiology or Medicine, which they shared with their colleague, André Lwoff).

It is difficult to appreciate the magnitude of their achievement now that we are so familiar with their ideas and have such direct ways of testing their models. To put it in perspective, consider what was known at the time they began their classic experiments:  $\beta$ -galactosidase activity appeared in *E. coli* cells only when lactose was provided in the growth medium. It was not clear that the appearance of this enzyme involved switching on expression of a gene. Indeed, one early explanation was that the cell contained a general (generic) enzyme and that enzyme took on whatever properties were required by the circumstances. Thus, when lactose was present, the generic enzyme took on the appropriate shape to metabolize lactose, using the sugar itself as a template!

Jacob, Monod, and their coworkers dissected the problem genetically. We will not go through their experiments in any detail, but a brief summary gives a taste of their ingenuity.

First, they isolated mutants of *E. coli* that made  $\beta$ -galactosidase irrespective of whether lactose was present (i.e., mutants in which the enzyme was produced **constitutively**). These mutants came in two classes: in one, the gene encoding the

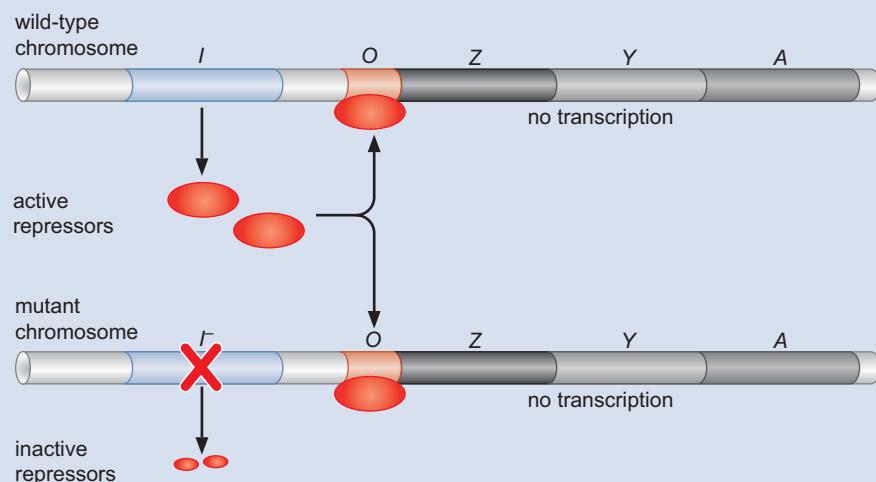
Lac repressor was inactivated; in the other, the operator site was defective. These two classes could be distinguished using a *cis-trans* test, as described later.

Jacob and Monod constructed partially diploid cells in which a section of the chromosome from a wild-type cell carrying the *lac* genes (i.e., the Lac repressor gene, *lacI*, the genes of the *lac* operon, and their regulatory elements) was introduced (on a plasmid called an F') into a cell carrying a mutant version of the *lac* genes on its chromosome. (This genetic trick is described more fully in the Bacteria section in Appendix 1.) This transfer resulted in the presence of two copies of the *lac* genes in the cell, making it possible to test whether the wild-type copy could complement any given mutant copy. When the chromosomal genes were expressed constitutively because of a mutation in the *lacI* gene (encoding repressor), the wild-type copy on the plasmid restored repression (and inducibility); that is,  $\beta$ -galactosidase was once again only made when lactose was present (Box 18-2 Fig. 1). This result is gained because the repressor made from the wild-type *lacI* gene on the plasmid can diffuse to the chromosome (i.e., it can act in *trans*).

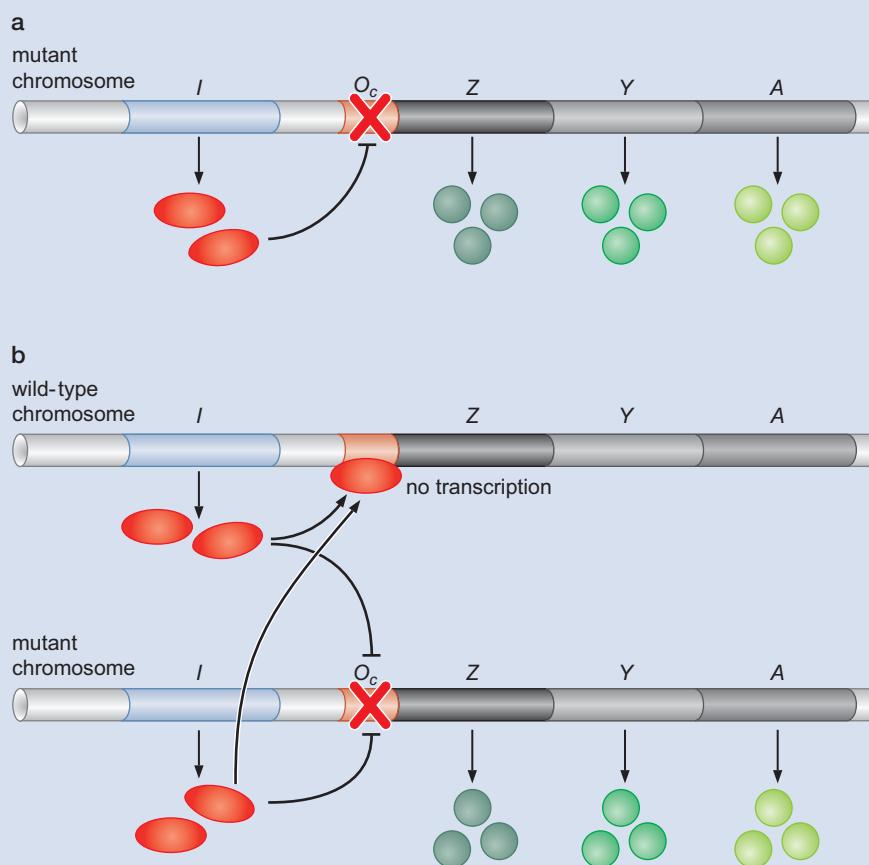
When the mutation causing constitutive expression of the chromosomal genes was in the *lac* operator, it could not be complemented in *trans* by the wild-type genes (Box 18-2 Fig. 2). The operator functions only in *cis* (i.e., it only acts on the genes directly linked to it on the same DNA molecule).

These and other results led Jacob and Monod to propose that genes were expressed from specific sites called promoters found at the start of the gene and that this expression was regulated by repressors that act through operator sites located on the DNA beside the promoter.

But these experiments with the *lac* system were not carried out in isolation; in parallel, Jacob and Monod did similar experiments on bacteriophage  $\lambda$  (a system we consider in detail later



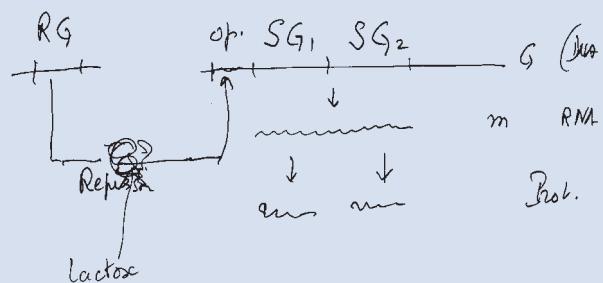
**BOX 18-2 FIGURE 1** Partial diploid cells show that functional repressors work in *trans*. In the absence of lactose, the *lac* genes are not expressed, and thus no significant level of  $\beta$ -galactosidase is made in these cells.

**Box 18-2** (Continued)

**BOX 18-2 FIGURE 2** Partial diploid cells show that operators work only in *cis*. (a) Haploid cell containing mutant operator ( $O_c$ ). (b) Partially diploid cell containing a normal operator ( $O$ ) and a mutant operator ( $O_c$ ). The *lac* genes ( $Z$ ,  $Y$ , and  $A$ ) attached to the mutant operator continue to be expressed constitutively even in the presence of a wild-type operator on another chromosome in the same cell. Thus, the operator only works in *cis*.

in this chapter). Bacteriophage  $\lambda$  can propagate through either of two life cycles. Which one is chosen depends on which of the relevant phage genes are expressed. The French scientists found they could isolate mutants defective in controlling gene expression in this system just as they had in the *lac* case. These mutations again defined a repressor that acted in *trans* through *cis*-acting operator sites. The similarity of these two regulatory systems (despite the very different biology) convinced Jacob and Monod that they had identified a fundamental mechanism of gene regulation and that their model would apply throughout nature. As we will see, although their description

was not complete—most noticeably, they did not include activators (such as CAP) in their scheme—the basic model they proposed of *cis*-regulatory sites recognized by *trans*-regulatory factors has dominated the majority of subsequent thinking about gene regulation.



**BOX 18-2 FIGURE 3** This drawing, showing the *lac* operon and its regulation, was rendered by François Jacob, 2002. (Courtesy of Jan Witkowski.)

François Jacob

Combinatorial control is a characteristic feature of gene regulation. Thus, when the same signal controls multiple genes, it is typically communicated to each of those genes by the same regulatory protein. This regulator will be communicating just one of perhaps several signals involved in regulating each gene; the other signals, different in most cases, will each be mediated by a separate regulator. More complex organisms—higher eukaryotes in particular—tend to have more signal integration, and there we will see greater and more elaborate examples of combinatorial control (Chapter 19).

### Alternative $\sigma$ Factors Direct RNA Polymerase to Alternative Sets of Promoters

Recall from Chapter 13 that it is the  $\sigma$  subunit of RNA polymerase that recognizes the promoter sequences (Fig. 13-6). The *lac* promoter that we have been discussing, along with the bulk of other *E. coli* promoters, is recognized by RNA polymerase bearing the  $\sigma^{70}$  subunit. But *E. coli* encodes several other  $\sigma$  subunits that can replace  $\sigma^{70}$  under certain circumstances and direct the polymerase to alternative promoters.

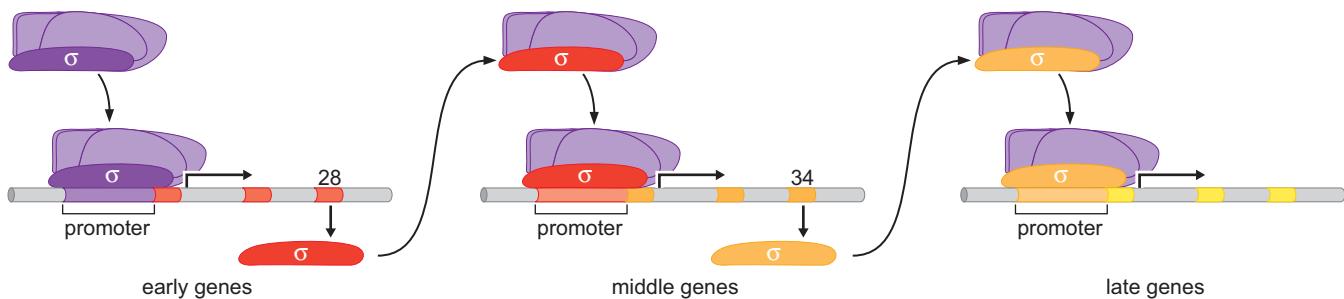
One of these alternatives is the heat shock  $\sigma$  factor,  $\sigma^{32}$ . Thus, when *E. coli* is subject to heat shock, the amount of this new  $\sigma$  factor increases in the cell, it displaces  $\sigma^{70}$  from a proportion of RNA polymerases, and it directs those enzymes to transcribe genes whose products protect the cell from the effects of heat shock. The level of  $\sigma^{32}$  is increased by two mechanisms: first, its translation is stimulated, that is, its mRNA is translated with greater efficiency after heat shock than it was before; and second, the protein is transiently stabilized. Another example of an alternative  $\sigma$  factor,  $\sigma^{54}$ , is considered in the next section.  $\sigma^{54}$  is associated with a small fraction of the polymerase molecules in the cell and directs that enzyme to genes involved in nitrogen metabolism.

Sometimes, a series of alternative sigmas directs a particular program of gene expression. Two examples are found in the bacterium *Bacillus subtilis*. We consider the most elaborate of these, which controls sporulation in that organism, in Chapters 21 and 22. The other we describe briefly here.

Bacteriophage SPO1 infects *B. subtilis*, where it grows lytically to produce progeny phage. This process requires that the phage expresses its genes in a carefully controlled order. That control is imposed on polymerase by a series of alternative  $\sigma$  factors. Thus, upon infection, the bacterial RNA polymerase (bearing the *B. subtilis* version of  $\sigma^{70}$ ) recognizes so called “early” phage promoters, which direct transcription of genes that encode proteins needed early in infection. One of these genes (called gene 28) encodes an alternative  $\sigma$ . This displaces the bacterial  $\sigma$  factor and directs the polymerase to a second set of promoters in the phage genome, those associated with the so-called “middle” genes. One of these genes, in turn, encodes the  $\sigma$  factor for the phage “late” genes (Fig. 18-15).

### NtrC and MerR: Transcriptional Activators That Work by Allostery Rather than by Recruitment

Although the majority of activators work by recruitment, there are exceptions. Two examples of activators that work not by recruitment but by allosteric mechanisms are NtrC and MerR. Recall what we mean by an allosteric mechanism. Activators that work by recruitment simply bring an active form of RNA polymerase to the promoter. In the case of activators that



**FIGURE 18-15** Alternative  $\sigma$  factors control the ordered expression of genes in a bacterial virus. The bacterial phage SPO1 uses three  $\sigma$  factors in succession to regulate expression of its genome. This ensures that viral genes are expressed in the order in which they are needed. (Adapted, with permission, from Alberts B. et al. 2002. *Molecular biology of the cell*, 4th ed., p. 415, Fig. 7-63. © Garland Science/Taylor & Francis LLC.)

work by allosteric mechanisms, polymerase initially binds the promoter in an inactive complex. To activate transcription, the activator triggers an allosteric change in that complex.

NtrC controls expression of genes involved in nitrogen metabolism by inducing a conformational change in a pre-bound RNA polymerase, triggering transition to the open complex. MerR controls expression of a gene involved in mercury resistance. MerR also acts on an inactive RNA polymerase–promoter complex, but in this case, the allosteric effect of the activator is on the DNA, rather than on the polymerase. We now describe these two systems in more detail.

#### NtrC Has ATPase Activity and Works from DNA Sites Far from the Gene

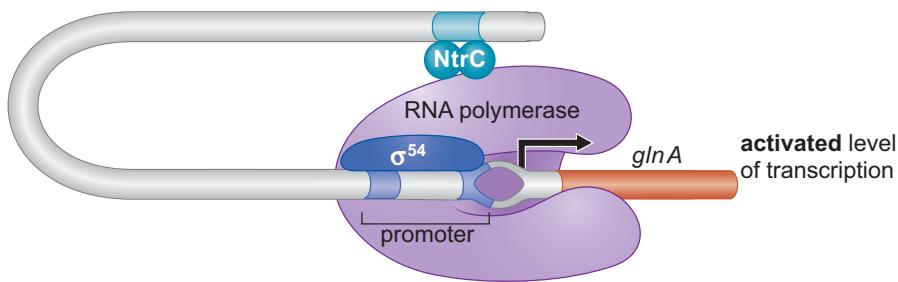
As with CAP, NtrC has separate activating and DNA-binding domains and binds DNA only in the presence of a specific signal. In the case of NtrC, this signal is low nitrogen levels. Under these conditions, NtrC is phosphorylated by a kinase, NtrB, and, as a result, undergoes a conformational change that reveals the activator's DNA-binding domain. Once active, NtrC binds four sites located approximately 150 bp upstream of the promoter (e.g., that of the *glnA* gene). NtrC binds to each of its sites as a dimer and, through protein–protein interactions between the dimers, binds to the four sites in a highly cooperative manner.

The form of RNA polymerase that transcribes the *glnA* gene contains the  $\sigma^{54}$  subunit. This enzyme binds to the *glnA* promoter in a stable closed complex in the absence of NtrC. Once active, NtrC (bound to its sites upstream) interacts directly with  $\sigma^{54}$ . This requires that the DNA between the activator-binding sites and the promoter form a loop to accommodate the interaction (Fig. 18-16). If the NtrC-binding sites are moved further upstream (as much as 1–2 kb), the activator can still work.

NtrC itself has an enzymatic activity—it is an ATPase. This activity provides the energy needed to induce a conformational change in polymerase. This conformational change triggers polymerase to initiate transcription. Specifically, it stimulates conversion of the stable inactive closed complex to an active open complex.

At some genes controlled by NtrC, there is a binding site for another protein, called IHF, located between the NtrC-binding sites and the promoter. Upon binding, IHF bends DNA. When the IHF-binding site, and hence the DNA bend, are in the correct register, this event increases activation by

**FIGURE 18-16 Activation by NtrC.** The promoter sequence recognized by  $\sigma^{54}$ -containing holoenzyme is different from that recognized by  $\sigma^{70}$ -containing holoenzyme. Although not specified in the figure, NtrC contacts the  $\sigma^{54}$  subunit of polymerase. NtrC is shown as a dimer, but in fact forms a higher-order complex on DNA.



NtrC. The explanation is that by bending the DNA, IHF brings the DNA-bound activator closer to the promoter, helping the activator interact with the polymerase bound there (see Fig. 18-4; for a closer look at how IHF bends DNA, see Fig. 12-11).

### MerR Activates Transcription by Twisting Promoter DNA

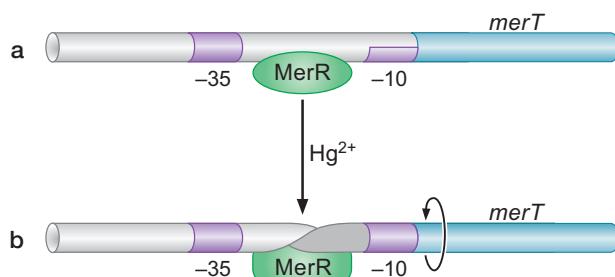
When bound to a single DNA-binding site, in the presence of mercury, MerR activates the *merT* gene. As shown in Figure 18-17, MerR binds to a sequence located between the  $-10$  and  $-35$  regions of the *merT* promoter (this gene is transcribed by  $\sigma^{70}$ -containing polymerase). MerR binds on the opposite face of the DNA helix from that bound by RNA polymerase, and so polymerase can (and does) bind to the promoter at the same time as MerR.

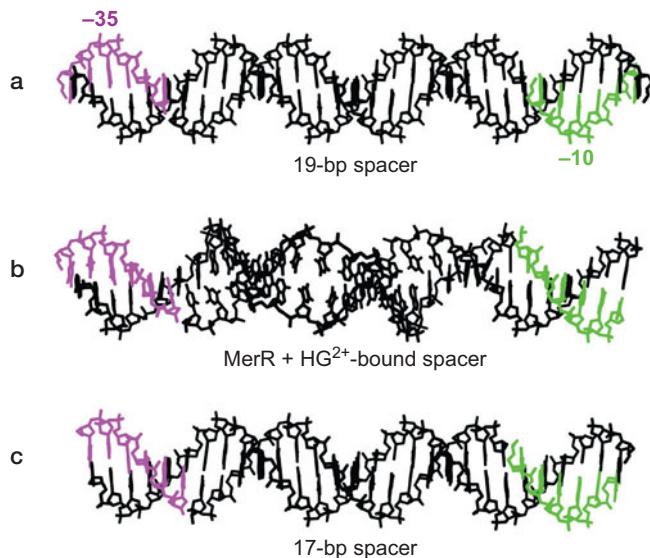
The *merT* promoter is unusual. The distance between the  $-10$  and  $-35$  elements is 19 bp instead of the 15–17 bp typically found in an efficient  $\sigma^{70}$  promoter (see Chapter 13, Box 13-1). As a result, these two sequence elements recognized by  $\sigma$  are neither optimally separated nor aligned; they are somewhat rotated around the face of the helix with respect to each other. Furthermore, the binding of MerR (in the absence of  $Hg^{2+}$ ) locks the promoter in this unpropitious conformation—polymerase can bind, but not in a manner that allows it to initiate transcription. Therefore, there is no basal transcription.

When MerR binds  $Hg^{2+}$ , however, the protein undergoes a conformational change that causes the DNA in the center of the promoter to twist. This structural distortion restores the disposition of the  $-10$  and  $-35$  regions to something close to that found at a strong  $\sigma^{70}$  promoter. In this new configuration, RNA polymerase can efficiently initiate transcription. The structures of promoter DNA in the “active” and “inactive” states have been determined (for another promoter regulated in this manner) and are shown in Figure 18-18.

It is important to note that in this example, the activator does not interact with RNA polymerase to activate transcription, but instead alters the

**FIGURE 18-17 Activation by MerR.** The  $-10$  and  $-35$  elements of the *merT* promoter lie on nearly opposite sides of the helix. (a) In the absence of mercury, MerR binds and stabilizes the inactive form of the promoter. (b) In the presence of mercury, MerR twists the DNA so as to properly align the promoter elements.





**FIGURE 18-18** Structure of a *merT*-like promoter. (a) Promoter with a 19-bp spacer. (b) Promoter with a 19-bp spacer when in complex with active activator. (c) Promoter with a 17-bp spacer. The promoter shown in parts a and b is from the *bmr* gene of *B. subtilis*, which is controlled by the regulator BmrR. BmrR works as an activator when complexed with the drug tetraphenylphosphonium (TPP). The  $-35$  (TTGACT) and  $-10$  (TACAGT) elements of one strand are shown in pink and green, respectively. (Adapted, with permission, from Zheleznova Heldwein E.E. and Brennan R.G. 2001. *Nature* 409: 378; Fig. 3 b–d. © Macmillan.)

conformation of the DNA in the vicinity of the prebound enzyme. Thus, unlike the earlier cases, there is no separation of DNA-binding and activating regions—for MerR, DNA binding is intimately linked to the activation process.

### Some Repressors Hold RNA Polymerase at the Promoter Rather than Excluding It

The Lac repressor works in the simplest possible way: by binding to a site overlapping the promoter, it blocks RNA polymerase binding. Many repressors work in that same way. In the MerR case, we saw a different form of repression: the protein holds the promoter in a conformation incompatible with transcription initiation. There are other ways repressors can work, one of which we now consider.

Some repressors work from binding sites that do not overlap the promoter. These repressors do not block polymerase binding, but instead they bind to sites beside a promoter, interact with polymerase bound at that promoter, and inhibit initiation. One is the *E. coli* Gal repressor. As we mentioned earlier, the Gal repressor controls genes that encode enzymes involved in galactose metabolism, and, in the absence of galactose, the repressor keeps the genes off. In this case, the repressor interacts with the polymerase in a manner that inhibits transition from the closed to open complex.

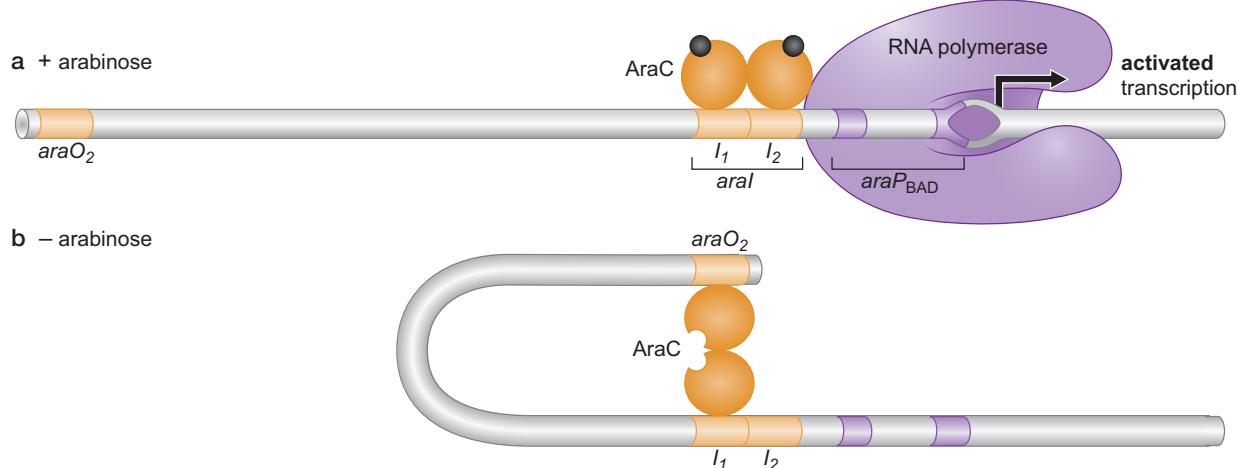
Another example is provided by the P<sub>4</sub> protein from a bacteriophage (φ29) that grows on the bacterium *B. subtilis*. This regulator binds to a site adjacent to one promoter—a weak promoter called P<sub>A3</sub>—and, by interacting with polymerase, serves as an activator. The interaction is with the αCTD, just as we saw with CAP. But this activator also binds at another promoter—a strong promoter called P<sub>A2c</sub>. Here, it makes the same contact with polymerase as at the weak promoter, but the result is repression. It seems that whereas in the former case, the extra binding energy helps recruit polymerase and hence activates the gene, in the latter case, the overall binding energy—provided by the strong interactions between the polymerase and the promoter and the additional interaction provided by the activator—is so strong that the polymerase is unable to escape the promoter.

### AraC and Control of the *araBAD* Operon by Antiactivation

The promoter of the *araBAD* operon from *E. coli* is activated in the presence of arabinose and the absence of glucose and directs expression of genes encoding enzymes required for arabinose metabolism. Unlike the cases of *lac* and *gal* genes, where a repressor and an activator work together, here two activators work together: AraC and CAP. When arabinose is present, AraC binds that sugar and adopts a configuration that allows it to bind DNA as a dimer to the adjacent half-sites, *araI*<sub>1</sub> and *araI*<sub>2</sub> (Fig. 18-19a). Just upstream of these (but not shown in the figure) is a CAP site: in the absence of glucose, CAP binds here and helps activation.

In the absence of arabinose, the *araBAD* genes are not expressed. This is because when not bound to arabinose, AraC adopts a different conformation and binds DNA in a different way: one monomer still binds the *araI*<sub>1</sub> site, but the other monomer binds a distant half-site called *araO*<sub>2</sub>, as shown in Figure 18-19b. As these two half-sites are 194 bp apart, when AraC binds in this fashion, the DNA between the two sites forms a loop. In addition, when bound in this way, there is no monomer of AraC at *araI*<sub>2</sub>, and as this is the position from which activation of *araBAD* promoter is mediated, there is no activation in this configuration.

The magnitude of induction of the *araBAD* promoter by arabinose is very large, and for this reason, the promoter is often used in **expression vectors**. Expression vectors are DNA constructs in which efficient synthesis of any protein can be ensured by fusing its gene to a strong promoter (see Chapter 7). In this case, fusing a gene to the *araBAD* promoter allows expression of the gene to be controlled by arabinose alone: the gene can be kept off when its expression is undesirable, and then both “derepressed” and “induced” when its product is wanted, simply by addition of arabinose. This allows expression even of genes with products that are toxic to the bacterial cells—that is, genes that must be kept very tightly repressed when not induced.



**FIGURE 18-19** Control of the *araBAD* operon. (a) Arabinose binds to AraC, changing the shape of that activator so that it binds as a dimer to *araI*<sub>1</sub> and *araI*<sub>2</sub>. This places one monomer close to the promoter from which it can activate transcription. (b) In the absence of arabinose, the AraC dimer adopts a different conformation and binds to *araO*<sub>2</sub> and *araI*<sub>1</sub>. In this position, there is no monomer at site *araI*<sub>2</sub>, and so the protein cannot activate the *araBAD* promoter (*araP<sub>BAD</sub>*). This promoter is also controlled by CAP (not shown in this figure).

We will now turn to more complicated transcriptional regulatory networks and see how these are created by reiteration of the simple mechanisms we have already encountered. We will focus on the case of bacteriophage  $\lambda$ , where we see how layers of regulators in various combinations can produce positive- and negative-feedback loops that allow alternative patterns of gene expression to be established and maintained, each driving a very different biological response. A fascinating case of simpler feedback loops involved in the biological process called **quorum sensing** is described in Box 18-3. Quorum sensing is used by bacteria in many contexts, but perhaps most importantly as part of their strategy of pathogenesis.

#### MEDICAL CONNECTIONS

##### Box 18-3 Blocking Virulence by Silencing Pathways of Intercellular Communication

In the early days of microbiology, bacteria were considered to be asocial organisms that led individualistic lifestyles involving little interaction with each other. We now know, however, that many bacteria can and do communicate with each other by emitting, detecting, and responding to chemical signals. One prevalent mode of intercellular communication known as **quorum sensing** enables bacteria to turn on genes synchronously in response to increases in the density of cells in the population. Expression of certain genes, such as those for bioluminescence, virulence factors, antibiotics, **biofilm formation**, and competence for uptake of DNA, are advantageous to bacteria only when the cells are in large numbers. For example, the bioluminescent bacterium *Vibrio fischeri* produces the light-emitting enzyme **luciferase** when it reaches a critical cell population density in the light organ of its host, the squid. It is of little benefit to *V. fischeri* to produce luciferase when it is on its own as a single cell. Likewise, the human pathogen *Pseudomonas aeruginosa* produces and secretes virulence factors, such as pyocyanin, cyanide, and lipase, only at a stage of infection when the concerted action of many bacteria can allow the factors to accumulate to high concentration. Elsewhere in this chapter we have discussed how bacteria turn genes on and off in response to signals from the environment. Here we focus on the molecular mechanisms by which bacteria respond to chemical signals known as **autoinducers** that they themselves produce to communicate with each other. As we shall see, an understanding of these mechanisms leads to provocative new strategies for treating infections by pathogenic bacteria based on chemically silencing intercellular communication.

We will focus on one widespread type of signaling molecule known as acyl-homoserine lactones (AHLs) and their recognition by the LuxR family of regulatory proteins. AHLs are simple organic molecules, consisting of a homoserine lactone ring attached to an acyl chain of four to 18 carbon units. Different species of bacteria produce AHLs with different length tails, each species responding just to the type of signaling molecule that it itself produces. This diversity of signals allows quorum-sensing bacteria to have, in effect, private conversations with their kin. AHLs are membrane-permeable molecules. They are able to diffuse into cells where they bind to LuxR. LuxR is an activator protein that is only able to turn on transcription when it is in

a complex with its ligand AHL. Thus, like the activator CAP, which as we have seen depends on the ligand cyclic AMP, LuxR binds to the promoter region of target genes when it is complexed with its appropriate AHL. Because the signaling molecule enters the cell from the outside, LuxR-mediated quorum sensing is an extremely simple signal transduction system in which the ligand directly interacts with the transcriptional regulator. In this sense, LuxR is analogous to certain eukaryotic regulatory proteins, such as the glucocorticoid receptor, that are directly activated by a membrane-permeable ligand (a sterol) that binds to, and thereby activates, its cognate regulatory protein (allowing it to migrate from the cytoplasm into the nucleus in the case of the glucocorticoid receptor). Interestingly, in some cases, AHLs work in a more complicated fashion, binding to a receptor embedded in the cytoplasmic membrane, thereby triggering phosphorylation and activation of a separate transcriptional regulatory protein in the cytoplasm. This is comparable to another eukaryotic system, the STAT pathway, described in the next chapter (Fig. 19-24).

Upon complexing with AHL, LuxR binds to the promoter region of target genes in the chromosome (the luciferase gene, virulence factor genes, and other quorum-sensing-responsive genes depending on the bacterium). One of the targets, *luxI*, specifies the synthetase for AHL. This has an important consequence as it creates a positive feedback loop as we explain. Because *luxI* is under the control of LuxR, the binding of AHL to LuxR simulates the expression of *luxI* and hence enhances the production of AHL. The additional AHL molecules diffuse out of the producing cell and back into other cells in the population. This leads to further activation of LuxR, which, in turn, promotes yet more AHL synthesis, raising the extracellular concentration of AHL still higher. This positive feedback loop only works when the cells are at a population density sufficiently high to enable AHL to reach a threshold concentration. When the cell density is below this threshold, the extracellular concentration of AHL is too low to set the positive feedback loop in motion. Thus, genes under the control of LuxR are only expressed when the cells are present in adequate numbers, a "quorum," to trigger the positive feedback loop.

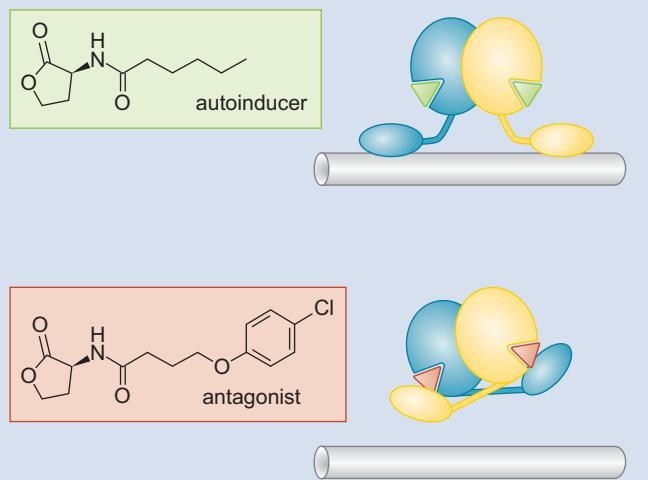
This knowledge has been put to use in an effort to devise quorum-sensing antagonists that block virulence gene

**Box 18-3** (Continued)

expression in pathogenic bacteria. One such pathogen is *Chromobacterium violaceum*, whose LuxR-type transcription factor is activated by an autoinducer called N-hexanoyl homoserine lactone with a six-carbon-long chain. LuxR-type factors are dimers in which each subunit consists of a ligand-binding domain and a DNA-binding domain. X-ray crystallographic studies have shown that N-hexanoyl homoserine lactone forms a complex with its cognate LuxR factor in which the two subunits are side-by-side, allowing the two DNA-binding domains to contact DNA (Box 18-3 Fig. 1). The *C. violaceum* LuxR is, however, potently inhibited by a chlorolactone antagonist whose structure is shown in the figure. The chlorolactone binds to LuxR but traps the transcription factor in a crisscross conformation in which the ligand-binding domain of one subunit is associated with the DNA-binding domain of the other subunit. In this inactive conformation, the dimer is unable to bind DNA and hence unable to activate transcription. Experiments with *Caenorhabditis elegans* as a model host for

**BOX 18-3 FIGURE 1** A quorum-sensing antagonist that functions by stabilizing an inactive conformation of LuxR. (Adapted, with permission, from the graphical abstract in the online Table of Contents of *Mol. Cell*, Vol. 42 [2011], article on pp. 199–209.)

infection by *C. violaceum* show that chlorolactone is indeed able to protect the nematode from quorum-sensing–mediated killing. These findings show how knowledge of the structure and function of a transcription factor that plays a central role in pathogenesis can be exploited for the development of small molecule antagonists that can serve as potential therapeutics.



## THE CASE OF BACTERIOPHAGE λ: LAYERS OF REGULATION

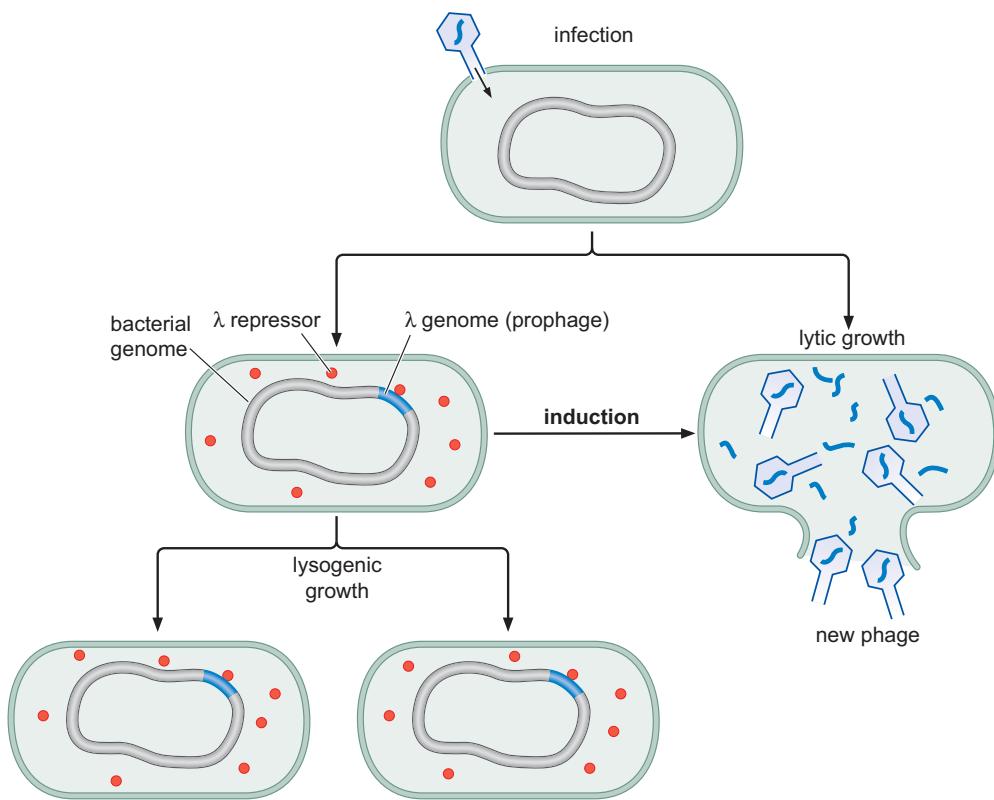
Bacteriophage λ is a virus that infects *E. coli*. Upon infection, the phage can propagate in either of two ways: **lytically** or **lysogenically**, as illustrated in Figure 18-20. Lytic growth requires replication of the phage DNA and synthesis of new coat proteins. These components combine to form new phage particles that are released by lysis of the host cell. Lysogeny—the alternative propagation pathway—Involves integration of the phage DNA into the bacterial chromosome where it is passively replicated at each cell division, as though it were a legitimate part of the bacterial genome.

A lysogen is extremely stable under normal circumstances, but the phage dormant within it—the **prophage**—can efficiently switch to lytic growth if the cell is exposed to agents that damage DNA (and thus threaten the host cell’s continued existence). This switch from lysogenic to lytic growth is called **lysogenic induction**.

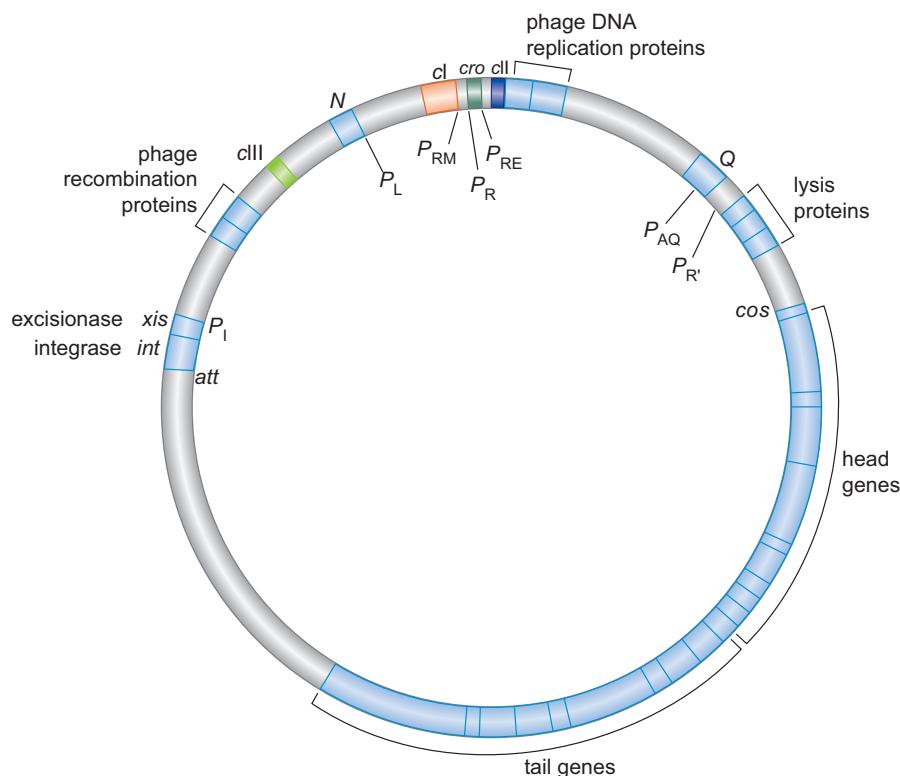
The choice of developmental pathway depends on which of two alternative programs of gene expression is adopted in that cell. The program responsible for the lysogenic state can be maintained stably for many generations but then, upon induction, switch over to the lytic program with great efficiency.

### Alternative Patterns of Gene Expression Control Lytic and Lysogenic Growth

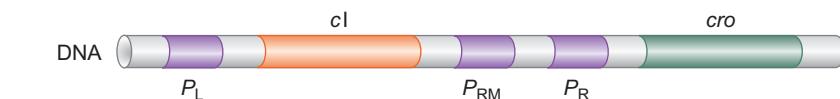
Bacteriophage λ has a 50-kb genome and approximately 50 genes. Most of these genes encode coat proteins, proteins involved in DNA replication, recombination, and lysis (Fig. 18-21). The products of these genes are



**FIGURE 18-20** Growth and induction of  $\lambda$  lysogen. Upon infection,  $\lambda$  can grow either lytically or lysogenically. A lysogen can be propagated stably for many generations or it can be induced. Following induction, the lytic genes are expressed in proper order, leading to the production of new phage particles.



**FIGURE 18-21** Map of bacteriophage  $\lambda$  in the circular form.  $\lambda$  genome is linear in the phage head, but, upon infection, circularizes at the cos site. When integrated into the bacterial chromosome, the phage genome is again linearized, but this time the ends are at the att site (see Chapter 12, Fig. 12-10, for a description of integration).



**FIGURE 18-22** Promoters in the right and left control regions of bacteriophage  $\lambda$ .

important in making new phage particles during the lytic cycle, but our concern here is restricted to the regulatory proteins, and where they act. We can therefore concentrate on just a few of them and start by considering a very small area of the genome shown in Figure 18-22.

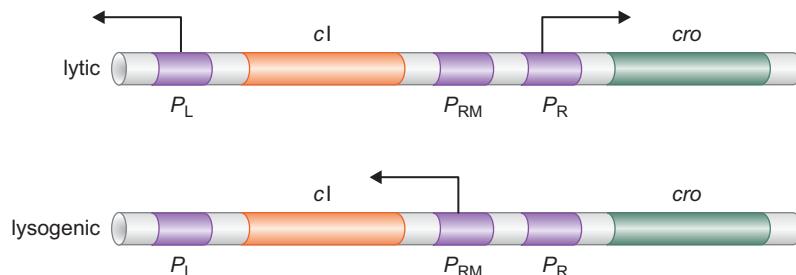
The depicted region contains two genes (*cl* and *cro*) and three promoters ( $P_R$ ,  $P_L$ , and  $P_{RM}$ ). All of the other phage genes (except one minor one) are outside this region and are transcribed directly from  $P_R$  and  $P_L$  (which stand for rightward and leftward promoter, respectively), or from other promoters whose activities are controlled by products of genes transcribed from  $P_R$  and  $P_L$ .  $P_{RM}$  (promoter for repressor maintenance) transcribes only the *cl* gene.  $P_R$  and  $P_L$  are strong, constitutive promoters; that is, they bind RNA polymerase efficiently and direct transcription without help from an activator.  $P_{RM}$ , in contrast, is a weak promoter and only directs efficient transcription when an activator is bound just upstream.  $P_{RM}$  resembles the *lac* promoter in this regard.

Two arrangements of gene expression are depicted in Figure 18-23: one renders growth lytic, the other lysogenic. Lytic growth proceeds when  $P_L$  and  $P_R$  remain switched on while  $P_{RM}$  is kept off. Lysogenic growth, in contrast, is a consequence of  $P_L$  and  $P_R$  being switched off and  $P_{RM}$  switched on. How are these promoters controlled?

### Regulatory Proteins and Their Binding Sites

The *cl* gene encodes  $\lambda$  repressor, a protein of two domains joined by a flexible linker region (Fig. 18-24). The amino-terminal domain contains the DNA-binding region (a helix-turn-helix domain, as we saw earlier). As with the majority of DNA-binding proteins,  $\lambda$  repressor binds DNA as a dimer; the main dimerization contacts are made between the carboxy-terminal domains. A single dimer recognizes a 17-bp DNA sequence, each monomer recognizing one half-site, again just as we saw in the *lac* system. (We have already looked at the details of DNA recognition by  $\lambda$  repressor in Fig. 6-13.)

Despite its name,  $\lambda$  repressor can both activate and repress transcription. When functioning as a repressor, it works in the same way as the Lac repressor: it binds to sites that overlap the promoter and excludes RNA polymerase. As an activator,  $\lambda$  repressor works like CAP—by recruitment.  $\lambda$  repressor's activating region is in the amino-terminal domain of the protein.



**FIGURE 18-23** Transcription in the  $\lambda$  control regions in lytic and lysogenic growth. Arrows indicate which promoters are active at the decisive period during lytic and lysogenic growth, respectively. The arrows also show the direction of transcription from each promoter.

Its target on polymerase is a region of the  $\sigma$  subunit adjacent to the part of  $\sigma$  that recognizes the  $-35$  region of the promoter (region 4, see Chapter 13, Fig. 13-6).

Cro (which stands for control of repressor and other things) only represses transcription, like the Lac repressor. It is a single-domain protein and again binds as a dimer to 17-bp DNA sequences, using a helix-turn-helix motif.

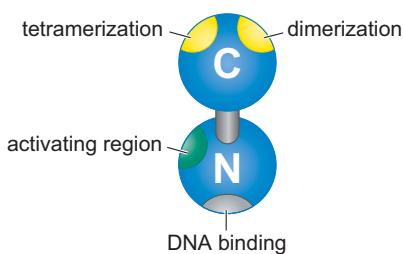
$\lambda$  repressor and Cro can each bind to any one of six operators. These sites are recognized with different affinities by each of the proteins. Three of these sites are found in the left-hand control region and three in the right. We focus on the binding of  $\lambda$  repressor and Cro to the sites in the right-hand region shown in Figure 18-25. Binding to sites in the left-hand control region follows a similar pattern.

The three binding sites in the right operator are called  $O_{R1}$ ,  $O_{R2}$ , and  $O_{R3}$ ; these sites are similar in sequence, but not identical, and each one—if isolated from the others and examined separately—can bind either a dimer of repressor or a dimer of Cro. The affinities of these various interactions, however, are not all the same. Thus, repressor binds  $O_{R1}$  tenfold better than it binds  $O_{R2}$ . In other words, ten times more repressor—a tenfold higher concentration—is needed to bind  $O_{R2}$  than to bind  $O_{R1}$ .  $O_{R3}$  binds repressor with about the same affinity as does  $O_{R2}$ . Cro, on the other hand, binds  $O_{R3}$  with highest affinity, and only binds  $O_{R2}$  and  $O_{R1}$  when present at tenfold higher concentration. The significance of these differences will become apparent presently.

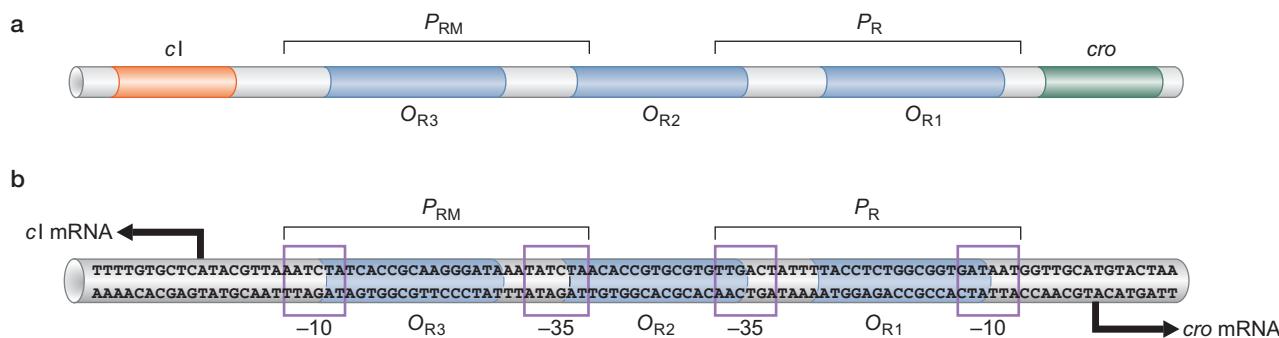
### $\lambda$ Repressor Binds to Operator Sites Cooperatively

$\lambda$  repressor binds DNA cooperatively. This is critical to its function and occurs as follows. Consider repressor binding to sites in  $O_R$ . In addition to providing the dimerization contacts, the carboxy-terminal domain of  $\lambda$  repressor mediates interactions *between* dimers (the point of contact is the patch marked “tetramerization” in Fig. 18-24). In this way, two dimers of repressor can bind cooperatively to adjacent sites on DNA.

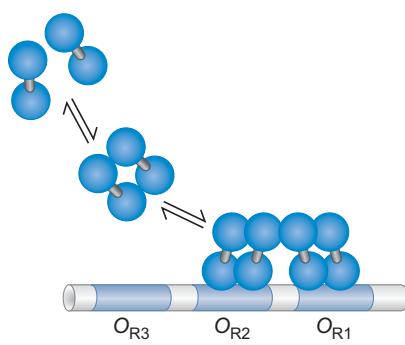
For example, the repressor at  $O_{R1}$  helps the repressor bind to the lower-affinity site  $O_{R2}$  by cooperative binding. Repressor thus binds both sites simultaneously and does so at a concentration that would be sufficient to bind only  $O_{R1}$  were the two sites tested separately (Fig. 18-26). (Recall that, without cooperativity, a tenfold higher concentration of repressor would be needed to bind  $O_{R2}$ .)  $O_{R3}$  is not bound: repressor bound cooperatively at



**FIGURE 18-24**  $\lambda$  repressor. The figure shows a monomer of  $\lambda$  repressor, indicating various surfaces involved in different activities carried out by the protein. N indicates the amino domain, C the carboxyl domain. “Tetramerization” denotes the region where two dimers interact when binding cooperatively to adjacent sites on DNA. (Adapted, with permission, from Ptashne M. and Gann A. 2002. *Genes & signals*, p. 36, Fig. 1.17. © Cold Spring Harbor Laboratory Press.)



**FIGURE 18-25** Relative positions of promoter and operator sites in  $O_R$ . Note that  $O_{R2}$  overlaps the  $-35$  region of  $P_R$  by 3 bp, and that of  $P_{RM}$  by 2 bp. This difference is enough for  $P_R$  to be repressed and  $P_{RM}$  activated by repressor bound at  $O_{R2}$ . (b, Adapted, with permission, from Ptashne M. 1992. *A genetic switch: Phage and higher organisms*, 2nd ed. © Blackwell Science.)



**FIGURE 18-26** Cooperative binding of  $\lambda$  repressor to DNA. The  $\lambda$  repressor monomers interact to form dimers, and those dimers interact to form tetramers. These interactions ensure that binding of repressor to DNA is cooperative. That cooperative binding is helped further by interactions between repressor tetramers at  $O_R$  interacting with others at  $O_L$  (see later in text and Fig. 18-28).

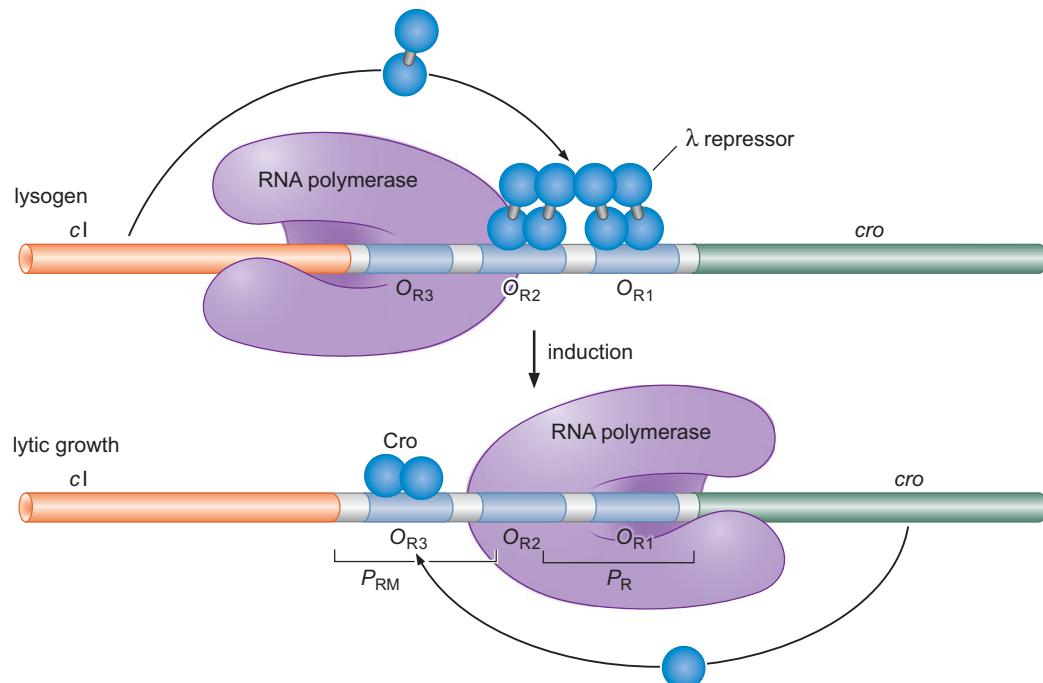
$O_{R1}$  and  $O_{R2}$  cannot simultaneously make contact with a third dimer at that adjacent site.

We have already discussed the idea of cooperative binding and seen an example: activation of the *lac* genes by CAP. As in that case, cooperative binding of repressors is a simple consequence of their touching each other while simultaneously binding to sites on the same DNA molecule.

For a more detailed discussion of the causes and effects of cooperative binding, see Box 18-4, Concentration, Affinity, and Cooperative Binding. Cooperative binding of regulatory proteins is used to ensure that changes in the level of expression of a given gene can be dramatic even in response to small changes in the level of a signal that controls that gene. The lysogenic induction of  $\lambda$ , discussed later, provides an excellent example of this sensitive aspect of control. In some systems, cooperative binding between activators is also the basis of signal integration (see the discussion on  $\beta$ -interferon in Chapter 19).

### Repressor and Cro Bind in Different Patterns to Control Lytic and Lysogenic Growth

How do repressor and Cro control the different patterns of gene expression associated with the different ways  $\lambda$  can replicate? As shown in Figure 18-27, for lytic growth, a single Cro dimer is bound to  $O_{R3}$ ; this site overlaps  $P_{RM}$  and so Cro represses that promoter (which would only work at a low level anyway in the absence of activator because the promoter is weak) (Fig. 18-27). As neither repressor nor Cro is bound to  $O_{R1}$  and  $O_{R2}$ ,  $P_R$  binds RNA polymerase and directs transcription of lytic genes;  $P_L$  does likewise. Recall that both  $P_R$  and  $P_L$  are strong promoters that need no activator.



**FIGURE 18-27** The action of  $\lambda$  repressor and Cro. Repressor bound to  $O_{R1}$  and  $O_{R2}$  turns off transcription from  $P_R$ . Repressor bound at  $O_{R3}$  contacts RNA polymerase at  $P_{RM}$ , activating expression of the *cl* (repressor) gene.  $O_{R3}$  lies within  $P_{RM}$ ; Cro bound there represses transcription of *cl*. (Adapted, with permission, from Ptashne M. and Gann A. 2002. *Genes & signals*, p. 30, Fig. 1.13. © Cold Spring Harbor Laboratory Press.)

## ► ADVANCED CONCEPTS

**Box 18-4** Concentration, Affinity, and Cooperative Binding

What do we mean when we talk about “strong” and “weak” binding sites? When we say two molecules recognize each other, or interact with each other, such as a protein and its site on DNA, we mean that they have some affinity for each other. Whether they are actually found bound together at any given time depends on (1) how high that affinity is (i.e., how tightly they interact), and (2) the concentration of the molecules.

As we emphasized in Chapter 3, the molecular interactions that underpin regulation in biological systems are reversible: when interacting molecules find each other, they stick together for a period of time and then separate. The higher the affinity, the tighter the two molecules stick together and, in general, the longer they remain together before parting. The higher the concentration, the more often they will find each other in the first place. Thus, higher affinity and higher concentration have similar effects: they both result in the two molecules, in general, spending more time bound to each other.

**Cooperativity Visualized**

Cooperativity can be expressed in terms of increased affinity. Repressor has a higher affinity for  $O_{R1}$  than for  $O_{R2}$ . But once repressor is bound to  $O_{R1}$ , repressor can bind  $O_{R2}$  more tightly because it interacts not only with  $O_{R2}$ , but with repressor bound at  $O_{R1}$  as well. Neither of these interactions is very strong alone, but when combined, they substantially increase the affinity of binding of that second repressor. As discussed in Chapter 3, the relationship between binding energy and equilibrium is an exponential one (see Table 3-1). Thus, increasing the binding energy as little as twofold increases affinity by one order of magnitude.

Another way to picture how cooperativity works is to think of it as increasing the local concentration of repressor. Picture repressor bound cooperatively at  $O_{R1}$  and  $O_{R2}$ . Although repressor at  $O_{R2}$  periodically lets go of DNA, it is holding on to repressor at  $O_{R1}$  and so remains in the proximity of  $O_{R2}$ . This effectively increases the local concentration of repressor in the vicinity of that site and ensures that repressor rebinds frequently.

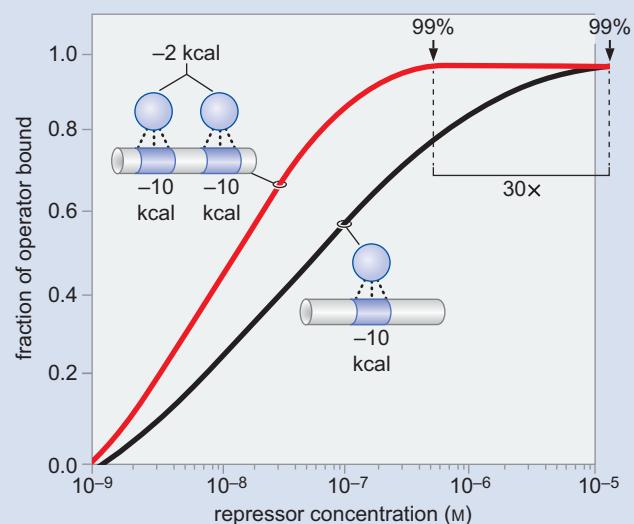
If we dispense with cooperativity and just increase the concentration of repressor in the cell, when repressor falls off  $O_{R2}$ , it will not be held nearby by repressor at  $O_{R1}$  and will usually drift away before it can rebind  $O_{R2}$ . But at the higher concentrations of repressor, another molecule of repressor will likely be close to  $O_{R2}$  and bind there. Thus, even if each repressor dimer only sits on  $O_{R2}$  for a short time, by either holding it nearby or increasing the number of possible replacements, the likelihood of repressor being bound will increase at any given time.

Yet another way of thinking about cooperative binding is as an entropic effect. When a protein goes from being free in solution to being constrained on a DNA-binding site, the entropy of the system decreases. But repressor held close to  $O_{R2}$  by interaction with repressor at  $O_{R1}$  is already constrained compared to its

free state. Rebinding of that constrained repressor has less entropic cost than does binding of free repressor.

We thus see three ways in which cooperativity can be pictured. We should also consider some of the consequences of cooperative binding that make it so useful in biology. For example, cooperativity not only enables a weak site to be filled at a lower concentration of protein than its inherent affinity would predict, it also changes the steepness of the curve describing the filling of that site with changes in concentration. To understand what is meant by that, consider as an example a protein binding cooperatively to two weak sites, A and B. These sites will go from essentially completely empty to almost completely filled over a much narrower range of protein concentration than would a single site (Box 18-4 Fig. 1). In fact, the cooperativity in the  $\lambda$  system is even greater than one might expect because a large fraction of free repressor (i.e., that not bound to DNA) is found as monomer in the cell: it is in essence a cooperative binding of four monomers, rather than two stable dimers, adding to the concerted nature of complex formation on DNA, and so adding to the steepness of the curve. But why does cooperativity make the binding curve steeper?

We have already seen how the site is filled at a lower concentration of repressor than its affinity would suggest, but how is it



**BOX 18-4 FIGURE 1** Cooperative binding reaction. The black line shows the curve that describes the binding of a protein to a single weak site on DNA, and the red line shows the same reaction when the same protein can bind cooperatively to two identical sites side by side (the interaction adding just 2 kcal/mol of extra binding energy). The x-axis shows a log scale of the protein concentration, and the y-axis shows occupancy of the sites. As shown, the site is 99% full at a 30-fold lower protein concentration when binding is cooperative. (Adapted, with permission, from Ptashne, M. 2004. *A genetic switch: Phage lambda revisited*. © Cold Spring Harbor Laboratory Press.)

**Box 18-4** (Continued)

that as repressor concentration decreases, binding falls away so quickly? Consider interactions between components of any system: as the concentration of the components is reduced, any given interaction between two of them will occur less frequently. If the system requires multiple interactions between several different components, this will become very rare at lower concentrations. Thus, binding of four monomers of a protein to two sites requires several (in fact, seven) interactions; the chance of the individual components coming together is drastically reduced as their individual concentrations decrease.

#### **Cooperativity and DNA-Binding Specificity**

A final important aspect of cooperative binding is that it imposes specificity on DNA binding. CAP activation of the *lac* promoter shows this. CAP brings RNA polymerase to promoters that bear CAP sites specifically (as opposed to other promoters of comparable affinity that lack CAP sites). Likewise,  $\lambda$  repressor at  $O_{R1}$  directs another molecule of repressor to bind to the weak site adjacent to it, not some other site of equal affinity elsewhere in the cell. In fact, cooperativity is vital to ensuring that proteins can bind with sufficient specificity for life to work as we know it.

To illustrate this, consider a protein binding to a site on DNA. This protein has a high affinity for its correct site. But the DNA within the cell represents a huge number of potential (but incorrect) binding sites for that protein. What is important, therefore, is not simply the absolute affinity of the protein for its correct site, but its affinity for that site compared to its affinity for all the other incorrect sites. And remember, those incorrect sites are at a much higher concentration than the correct site (representing, as they do, all of the DNA in the cell except the correct site). So even if the affinity for the incorrect sites is lower than that for the correct site, the higher concentration of the incorrect sites ensures that the protein will often sample them while attempting to reach its correct site.

What is needed is a strategy that increases affinity for the correct site without aiding interactions with the incorrect sites. Increasing the number of contacts between the protein

and its DNA site (e.g., by making the protein larger) does not necessarily help because it also tends to increase binding to the incorrect sites. Once affinity for the incorrect sites gets too high, the protein essentially never finds its correct site; it spends too much time sampling incorrect sites. Thus, a kinetic problem replaces the specificity one and it can be just as disruptive.

Cooperativity solves the problem. By binding to two adjacent sites cooperatively, a protein increases dramatically its affinity for those sites, without increasing affinity for other sites. The reason it does not increase affinity for the incorrect sites is simply because the chance of two molecules of protein binding incorrect sites close together at the same time (allowing cooperativity to stabilize that binding) is extremely remote. Only when they find the correct sites do they remain bound long enough to give a second protein a chance to turn up.

#### **Cooperativity and Allostery**

Although in this chapter we use the term *cooperativity* to refer to a particular mechanism of cooperative binding, the term is also used in other contexts where different mechanisms apply. In general, we might say that cooperativity describes any situation in which two ligands bind to a third molecule in such a way that the binding of one of those ligands helps the binding of the other. Thus, for the DNA-binding proteins we considered here, cooperativity is mediated by simple adhesive interactions, which provide increased binding energy, but in other situations, cooperativity can be mediated by allosteric events. Perhaps the best example of that is the binding of oxygen molecules to hemoglobin.

Hemoglobin is a homotetramer, and each subunit binds one molecule of oxygen. This binding is cooperative: when the first oxygen binds, it causes a conformational change that fixes the binding site for the next oxygen in a conformation with a higher affinity for that ligand. Thus, in this case, there is no direct interaction between the ligands, but by triggering an allosteric transition, one ligand increases affinity for a second.

During lysogeny,  $P_{RM}$  is on while  $P_R$  (and  $P_L$ ) are off. Repressor bound cooperatively at  $O_{R1}$  and  $O_{R2}$  blocks RNA polymerase binding at  $P_R$ , repressing transcription from that promoter (Fig. 18-27). But repressor bound at  $O_{R2}$  activates transcription from  $P_{RM}$ .

We will return shortly to the question of how the phage chooses between these alternative pathways. But first we consider induction—how the lysogenic state outlined above switches to the alternative lytic state when the cell is threatened.

#### **Lysogenic Induction Requires Proteolytic Cleavage of $\lambda$ Repressor**

*E. coli* senses and responds to DNA damage. It does this by activating the function of a protein called RecA. This enzyme is involved in recombination (which accounts for its name; see Chapter 11), but it has another function: it stimulates the proteolytic autocleavage of certain proteins. The primary

substrate for this activity is a bacterial repressor protein called LexA that represses genes encoding DNA-repair enzymes. Activated RecA stimulates autocleavage of LexA, releasing repression of those genes. This is called the SOS response (see Chapter 10).

If the cell is a lysogen, it is in the best interests of the prophage to escape under these threatening circumstances. To this end,  $\lambda$  repressor has evolved to resemble LexA, ensuring that  $\lambda$  repressor too undergoes autocleavage in response to activated RecA. The cleavage reaction removes the carboxy-terminal domain of repressor, and so dimerization and cooperativity are immediately lost. As these functions are critical for repressor binding to  $O_{R1}$  and  $O_{R2}$  (at concentrations of repressor found in a lysogen), loss of cooperativity ensures that the repressor dissociates from those sites (as well as from  $O_{L1}$  and  $O_{L2}$ ). Loss of repression triggers transcription from  $P_R$  and  $P_L$ , leading to lytic growth. Transcription from  $P_R$  quickly produces Cro, which binds  $O_{R3}$  and blocks any further synthesis of repressor from  $P_{RM}$ . This action ensures that the decision to induce is irreversible.

For induction to work efficiently, the level of repressor in a lysogen must be tightly regulated. If levels were to drop too low, under normal conditions, the lysogen might spontaneously induce; if levels rose too high, appropriate induction would be inefficient. The reason for the latter is that more repressor would have to be inactivated (by RecA) for the concentration to drop enough to vacate  $O_{R1}$  and  $O_{R2}$ . We have already seen how repressor ensures that its level never drops too low: it activates its own expression, an example of **positive autoregulation**. But how does it ensure levels never get too high? Repressor also regulates itself negatively.

This **negative autoregulation** works as follows. As drawn, Figure 18-27 shows  $P_{RM}$  being activated by repressor (at  $O_{R2}$ ) to make more repressor. But if the concentration gets too high, repressor will bind to  $O_{R3}$  as well and repress  $P_{RM}$  (in a manner analogous to Cro binding  $O_{R3}$  and repressing  $P_{RM}$ ). This prevents synthesis of new repressor until its concentration falls to a level at which it vacates  $O_{R3}$ .

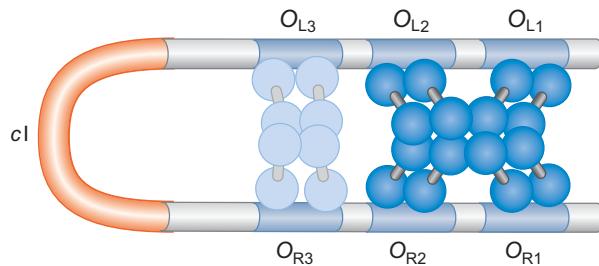
As an aside, it is interesting to note that the term “induction” is used to describe both the switch from lysogenic growth to lytic growth in  $\lambda$  and the switching on of the *lac* genes in response to lactose. This common usage stems from the fact that both phenomena were studied in parallel by Jacob and Monod (see Box 18-2). It is also worth noting that, just as lactose induces a conformational change in Lac repressor to relieve repression of the *lac* genes, so too the inducing signals of  $\lambda$  work by causing a structural change (in this case, proteolytic cleavage) in  $\lambda$  repressor.

### Negative Autoregulation of Repressor Requires Long-Distance Interactions and a Large DNA Loop

We have discussed cooperative binding of repressor dimers to adjacent operators such as  $O_{R1}$  and  $O_{R2}$ . There is yet another level of cooperative binding seen in the prophage of a lysogen, one critical to proper negative autoregulation. Repressor dimers at  $O_{R1}$  and  $O_{R2}$  interact with repressor dimers bound cooperatively at  $O_{L1}$  and  $O_{L2}$ . These interactions produce an octomer of repressor. Each dimer within the octamer is bound to a separate operator.

To accommodate the long-distance interaction between repressors at  $O_R$  and  $O_L$ , the DNA between those operator regions (about 3.5 kb, including the *cI* gene itself) must form a loop (Fig. 18-28). When the loop is formed,  $O_{R3}$  is held close to  $O_{L3}$ . This allows another two dimers of repressor to bind cooperatively to these two sites. This cooperativity means  $O_{R3}$  binds repressor at a lower concentration than it otherwise would—indeed, at a

**FIGURE 18-28** Interaction of repressors at  $O_R$  and  $O_L$ . Repressors at  $O_R$  and  $O_L$  interact as shown. These interactions stabilize binding. In this way, the interactions increase repression of  $P_R$  and  $P_L$  and allow repressor to bind  $O_{R3}$  at a lower concentration than it otherwise could. The repressors bound at  $O_{L3}$  and  $O_{R3}$  are here shown in a lighter shade to indicate that they will be bound only when the concentration of repressor rises above a certain level, as described in the text. (Adapted, with permission, from Ptashne M. and Gann A. 2002. *Genes & signals*, p. 35, Fig. 1.16. © Cold Spring Harbor Laboratory Press.)



concentration only just a little higher than that required to bind  $O_{R1}$  and  $O_{R2}$ . Thus, repressor concentration is very tightly controlled: small decreases are compensated for by increased expression of its gene, and increases by switching the gene off. This explains why lysogeny can be so stable while also ensuring that induction is very efficient.

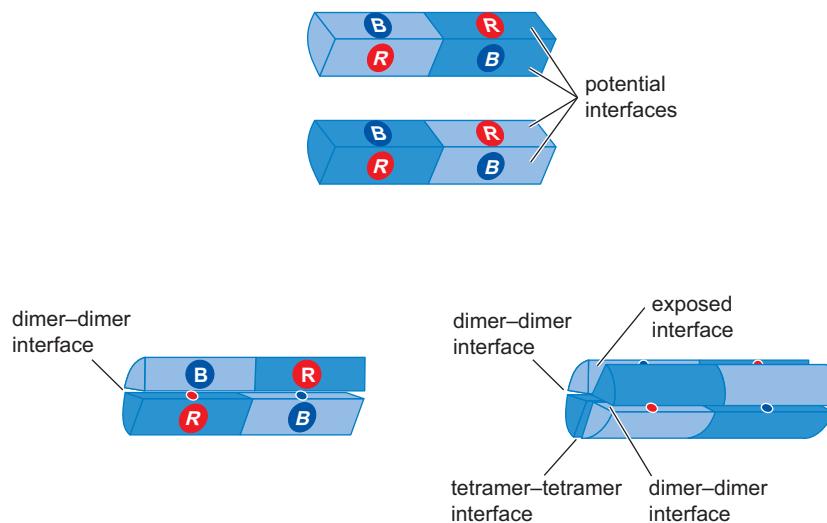
The structure of the carboxy-terminal domain of  $\lambda$  repressor, interpreted in light of earlier genetic studies, reveals the basis of dimer formation, but it also shows how two dimers interact to form the tetramer (as occurs when repressor is bound cooperatively to  $O_{R1}$  and  $O_{R2}$ ). Moreover, the structure reveals the basis for the octomer form and shows that this is the highest-order oligomer repressor can form (Fig. 18-29).

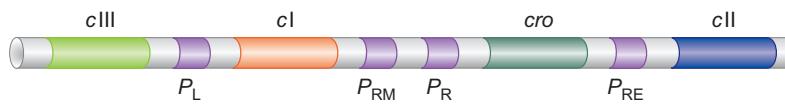
In Box 18-5, Evolution of the  $\lambda$  Switch, we discuss how the control circuits that govern lysogenic and lytic growth, and the process of induction, might have evolved. Specifically, we discuss how the interactions between repressor and Cro, their binding sites, and the promoters they regulate could have evolved to their current elaborate form in small steps from an earlier rudimentary system.

### Another Activator, $\lambda$ CII, Controls the Decision between Lytic and Lysogenic Growth upon Infection of a New Host

We have seen how  $\lambda$  repressor and Cro control lysogenic and lytic growth and the switch from one to the other upon induction. Now we turn to those events early in infection that determine which pathway the phage chooses in the first place. Critical to this choice are the products of two other  $\lambda$  genes,  $cII$  and  $cIII$ . We need only expand slightly our map of the regulatory region of

**FIGURE 18-29** Interactions between the carboxy-terminal domain of  $\lambda$  repressors. The figure shows, at the top, a schematic representation of two dimers of the carboxy-terminal domain of  $\lambda$  repressor. Indicated are the two patches here called B and R on the surface of that domain that mediate interactions between two dimers to give a tetramer, in the first instance, and then between two tetramers to give an octamer (the form found when repressor is bound cooperatively to the four sites,  $O_{R1}$ ,  $O_{R2}$ ,  $O_{L1}$ , and  $O_{L2}$ ). Once the octamer has formed, there is no space left for a further dimer to enter the complex, and so the octamer is the highest-order structure that forms. (Modified, with permission, from Bell et al. 2000. *Cell* 101: 801–811, Figs. 4a,b and 5a–c. © Elsevier.)





**FIGURE 18-30** Genes and promoters involved in the lytic/lysogenic choice. Not shown here is the gene *N*, which lies between  $P_L$  and *cIII* (see Fig. 18-21).

λ to see where *cII* and *cIII* lie: *cII* is on the right of *cl* and is transcribed from  $P_{RE}$ ; *cIII*, on the left of *cl*, is transcribed from  $P_L$  (Fig. 18-30).

Like the λ repressor, the CII protein is a transcriptional activator. It binds to a site upstream of a promoter called  $P_{RE}$  (for repressor establishment) and stimulates transcription of the *cl* (repressor) gene from that promoter.

#### ► KEY EXPERIMENTS

##### Box 18-5 Evolution of the λ Switch

We have emphasized many of the intricacies that underlie the mechanisms of decision-making by bacteriophage λ: how it chooses between lytic and lysogenic development and how it can efficiently switch from a stable prophage to a lytically replicating virus. Many of the subtleties that give the system these characteristics have been discussed: cooperative binding, auto-positive and negative regulation, the use of repressors with opposing effects, and so on. Emphasizing the intricate interplay of these features, and how interdependent they are in the phage we see today, rather begs the question of how such a system could have evolved in simple steps from an earlier primitive version. This is an important question when considering all biological systems, and we address it here for λ.

A proposed step-by-step model of how the λ switch might have evolved from a rudimentary version is shown (Box 18-5 Fig. 1). In each step, one simple addition has been made to a system that already works, to produce one that works a little better.

In the last few years, a series of experiments has explored the issues raised in this scheme. These studies point toward how relatively easily evolution might have molded the λ switch. Thus, each apparently critical feature of the existing switch has been eliminated by mutation, rendering the phage defective in various behaviors; the mutant phage might, for example, lysogenize less efficiently or form lysogens that are unstable, or perhaps too stable, making induction too easy or too difficult.

Other mutations were then found that compensated for the original defect in each case. These experiments revealed that far from the irreducible complexity that might on the surface seem to exist for this system, loss of any individually “essential” feature could be compensated for, at least partially, by alteration of another. For example, positive autoregulation was eliminated by introducing a *pc* mutation into the *cl* gene of the phage. A *pc* mutation, as we discussed earlier for CAP, eliminates the activation function of an activator. Thus, in this case, the mutant phage would make repressor that can still bind DNA and repress transcription but cannot activate expression of more repressor from  $P_{RM}$ . This mutant phage can form lysogens, but they are very unstable because repressor levels are low. Introducing other changes that strengthen the promoter  $P_{RM}$

compensates for this to a great extent, making the lysogens more stable and more like those produced by wild-type λ. The strengthened  $P_{RM}$  can direct expression of more repressor without being activated by the existing repressor. It seems that having autopositive regulation of repressor gives the wild-type phage an advantage (explaining why all known lambdoid phage have this feature), but it is not completely essential for the system to work fairly well. Thus, one can see an intermediate step in the evolution of the modern system.

In another example, cooperative binding by repressor was also shown to be a feature that, although advantageous, is not completely necessary for the phage to function in a rudimentary way. Thus, cooperative binding was substantially weakened by introducing mutations that had previously been shown to disrupt cooperative interactions between repressor dimers. Phage carrying this mutant repressor gene were unable to form lysogens. But addition of other modifications—one again strengthening  $P_{RM}$  and the other strengthening the binding site  $O_{R2}$  for repressor—together generated a phage that now could form lysogens, albeit less efficiently than wild-type λ.

In a further set of remarkable experiments, the λ switch was dismantled and reassembled in ways that test critical ideas about both how it functions and how it arose. In the most recent and ambitious of these, the repressor gene was replaced by a gene for a bacterial repressor protein, the Tet repressor, and in the same phage, the gene for Cro was substituted by *lacI*, the gene encoding Lac repressor. In addition, operator sites within the phage were modified to allow these two bacterial repressors to bind in patterns that mimic some of the critical binding patterns of λ repressor and Cro in wild-type λ.

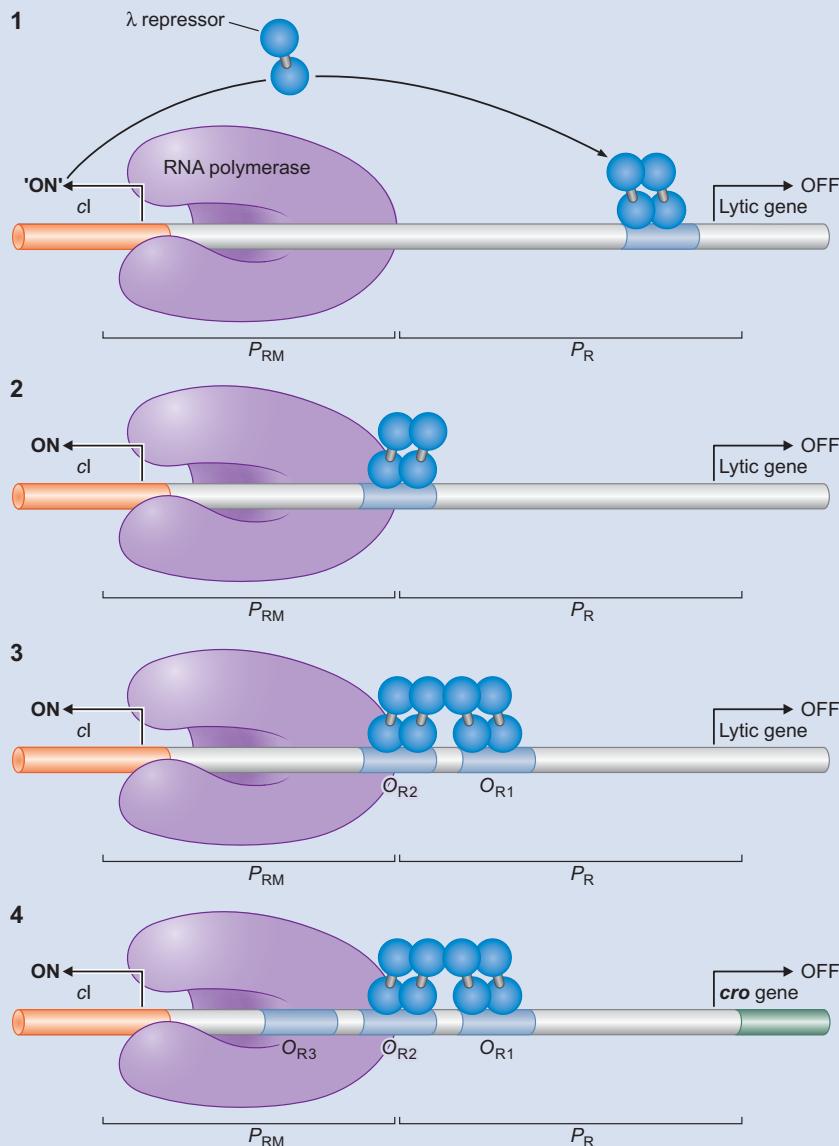
Phage built from these heterologous pieces could recapitulate some of the behaviors of wild-type λ. Because the binding of both the repressors employed in the modified phage can be titrated precisely by small molecules (*lac* and *tet* inducers), further subtle manipulations can now be used to investigate further the workings, and possible origins, of the λ system.

Taken together, these various experimental approaches make clear two points. First, the λ switch could easily have evolved through a series of steps, each adding a new level of

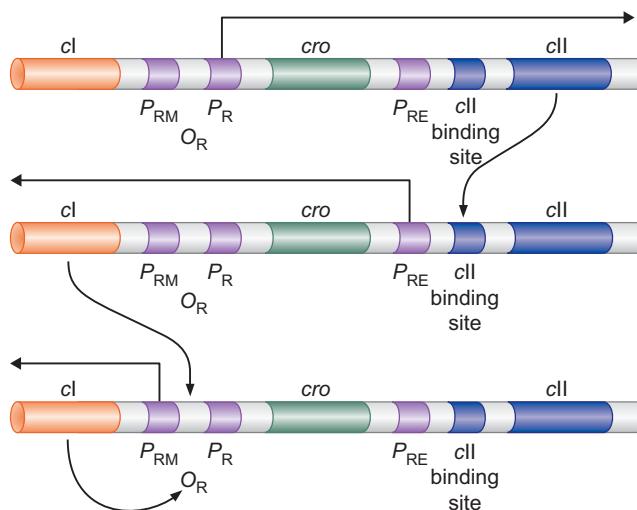
**Box 18-5 (Continued)**

regulation to a system that worked less well, but did work, before. This is what would be expected of any system that evolved through natural selection. Second, there are alternative ways any given behavior can be achieved. Understanding the details

of the solution that finally appeared can make the problem of how it evolved seem much more difficult than it necessarily was; that is, the final solution was only one of a variety that would have worked had they arisen.



**BOX 18-5 FIGURE 1** Hypothetical stages in the evolution of the  $\lambda$  switch. (Stage 1) The primitive  $\lambda$  genome bears two promoters, one directing expression of the lytic genes ( $P_R$ ) and one for the repressor gene ( $P_{RM}$ ). A single  $\lambda$  repressor-binding site overlaps  $P_R$ ; when bound to this site, repressor turns off the lytic genes, but its own synthesis is unregulated. (Stage 2) Here, the single repressor-binding site has moved close to  $P_{RM}$  (now in the position of  $O_{R2}$ ), so that bound repressor contacts polymerase at  $P_{RM}$  and thereby stimulates that promoter while repressing  $P_R$ . (Stage 3) A second repressor-binding site has been introduced (in the position of  $O_{R1}$ ). In addition, a new protein–protein interaction surface has been introduced, allowing cooperative binding of repressor dimers to these adjacent sites. These features contribute additional aspects of cooperativity to the system that increases the efficiency of the switch mechanism. (Stage 4) The third repressor-binding site ( $O_{R3}$ ) is introduced. When bound to this site, repressor negatively regulates its own synthesis such that its concentration remains below a critical level and ensures an efficient switch mechanism. (Adapted, with permission, from Ptashne M. and Gann A. 1998. *Curr. Biol.* 8: R812–R822. © Elsevier.)



**FIGURE 18-31** Establishment of lysogeny. The *cl* gene is transcribed from  $P_{RE}$  when establishing lysogeny and from  $P_{RM}$  when maintaining that state. Repressor bound at  $O_{R1}$  and  $O_{R2}$  not only activates the maintenance mode, but also turns off the establishment mode of expression. Note that  $P_R$  controls not only lytic genes, but also expression of *cII* and is thus important in lysogeny as well as lytic development. Similarly, although not shown in the figure,  $P_L$ , which controls many lytic genes, also controls the *cII* gene which helps establish lysogeny (see text). (Adapted, with permission, from Ptashne M. and Gann A. 2002. *Genes & signals*, p. 31, Fig. 1.14. © Cold Spring Harbor Laboratory Press.)

Thus, the repressor gene can be transcribed from two different promoters ( $P_{RE}$  and  $P_{RM}$ ).

$P_{RE}$  is a weak promoter because it has a very poor  $-35$  sequence. The CII protein binds to a site that overlaps the  $-35$  region but is located on the opposite face of the DNA helix; by directly interacting with polymerase, CII helps polymerase bind to the promoter.

Only when sufficient repressor has been made from  $P_{RE}$  can that repressor bind to  $O_{R1}$  and  $O_{R2}$  and direct its own synthesis from  $P_{RM}$ . Thus, we see that repressor synthesis is **established** by transcription from one promoter (stimulated by one activator) and then **maintained** by transcription from another promoter (under its own control—positive autoregulation).

We can now see in summary how CII orchestrates the choice between lytic and lysogenic development (Fig. 18-31). Upon infection, transcription is immediately initiated from the two constitutive promoters  $P_R$  and  $P_L$ .  $P_R$  directs synthesis of both Cro and CII. Cro expression favors lytic development: once Cro reaches a certain level, it will bind  $O_{R3}$  and block  $P_{RM}$ . CII expression, on the other hand, favors lysogenic growth by directing transcription of the repressor gene (Fig. 18-31). For successful lysogeny, repressor must then bind to  $O_{R1}$  and  $O_{R2}$  and activate  $P_{RM}$ .

The efficiency with which CII directs transcription of the *cI* gene, and hence the rate at which repressor is made, is the critical step in deciding how  $\lambda$  will develop. What determines how efficiently CII works in any given infection?

### The Number of Phage Particles Infecting a Given Cell Affects Whether the Infection Proceeds Lytically or Lysogenically

Multiplicity of infection (moi) is a measure of how many phage particles infect a given bacterial cell within a population. If the average number is one or fewer phage particles per cell, the infection is more likely to result in lysis. If the number of phage particles is two or more, it is more likely to produce lysogeny. And as the numbers of phage per cell become lower and lower, the tendency toward lytic infection increases, and as it becomes higher and higher, the likelihood of lysogeny similarly increases.

Mechanistically, this makes sense. The more phage genomes that enter the cell and start transcribing from  $P_R$  and  $P_L$ , the more CII and CIII gets made, and the greater the chance that at least one of those phage genomes

will establish repressor synthesis and integrate into the bacterial chromosome. As long as one of the infecting phage does this, the others will subsequently be blocked from further lytic development.

One can speculate as to why  $\lambda$  is set up to respond this way—why it would rather develop lysogenically when in a population of many phage and few bacteria, for example. If there are few bacterial cells, then availability of host cells for the next round of infection will be limited, and so the phage might benefit from becoming dormant within a lysogen rather than risk finding no further host cells after a round of lytic infection. The growth conditions of the bacterial cells also influence the outcome of an infection as described later.

### Growth Conditions of *E. coli* Control the Stability of CII Protein and Thus the Lytic/Lysogenic Choice

When the phage infects a population of bacterial cells that are healthy and growing vigorously, it tends to propagate lytically, releasing progeny into an environment rich in fresh host cells. When conditions are poor for bacterial growth, however, the phage is more likely to form lysogens and sit tight: again, there will likely be few host cells in the vicinity for any progeny phage to infect. These different growth conditions impinge on CII as follows.

CII is a very unstable protein in *E. coli*; it is degraded by a specific protease called FtsH (HflB), encoded by the *hfl* gene. The speed with which CII can direct synthesis of repressor is thus determined by how quickly it is being degraded by FtsH. Cells lacking the *hfl* gene (and thus FtsH) almost always form lysogens upon infection by  $\lambda$ : in the absence of the protease, CII is stable and directs synthesis of ample repressor. FtsH activity is itself regulated by the growth conditions of the bacterial cell, and, although it is not understood exactly how this is achieved, we can state the following. If growth is good, FtsH is very active, CII is destroyed efficiently, repressor is not made, and the phage tend to grow lytically. Under poor growth conditions, the opposite happens: low FtsH activity, slow degradation of CII, repressor accumulation, and a tendency toward lysogenic development. Levels of CII are also modulated by the phage protein CIII. CIII stabilizes CII, probably because it acts as an alternative (and thus competing) substrate for FtsH.

The *cI*, *cII*, and *cIII* genes were isolated in elegant genetic screens outlined in Box 18-6, Genetic Approaches That Identified Genes Involved in the Lytic/Lysogenic Choice.

A second CII-dependent promoter,  $P_I$ , has a sequence similar to that of  $P_{RE}$  and is located in front of the phage gene *int* (see Fig. 18-21); this gene encodes the integrase enzyme that catalyzes site-specific recombination of  $\lambda$  DNA into the bacterial chromosome to form the prophage (see Chapter 12). A third CII-dependent promoter,  $P_{AQ}$ , located in the middle of gene *Q*, acts to retard lytic development and thus to promote lysogenic development. This is because the  $P_{AQ}$  RNA acts as an antisense message, binding to the *Q* message and promoting its degradation. *Q* is another regulator, one that promotes the late stages of lytic growth, as discussed the next section.

### Transcriptional Antitermination in $\lambda$ Development

Two examples of transcriptional regulation *after* initiation are found in  $\lambda$  development, as we now describe. We start with a type of positive transcriptional regulation called **antitermination**.

## ► KEY EXPERIMENTS

**Box 18-6** Genetic Approaches That Identified Genes Involved in the Lytic/Lysogenic Choice

Genes involved in lytic/lysogenic choice were identified by screening for  $\lambda$  mutants that efficiently grow only either lytically or lysogenically. To understand how these mutants were found, we need to consider how phage are grown in the laboratory (see Appendix 1). Bacterial cells can be grown as a confluent, opaque lawn across an agar plate. A lytic phage, grown on that lawn, produces clear plaques, or holes (Fig. A-3). Each plaque is typically initiated by a single phage infecting a bacterial cell. The progeny phage from that infection then infect surrounding cells, and so on, killing off (lysing) the bacterial cells in the vicinity of the original infected cell and causing a clear cell-free zone in the otherwise opaque lawn of bacterial cells.

Bacteriophage  $\lambda$  forms plaques too, but they are turbid (or cloudy)—that is, the region within the plaque is clearer than the uninfected lawn, but only marginally so. The reason for this is that  $\lambda$ , unlike a purely lytic phage, kills only a proportion of the cells it infects; the others survive as lysogens. Lysogens are resistant to subsequent infection and so can grow within the plaque unharmed by the mass of phage particles found there. The reason for this “immunity” is quite simple: in a

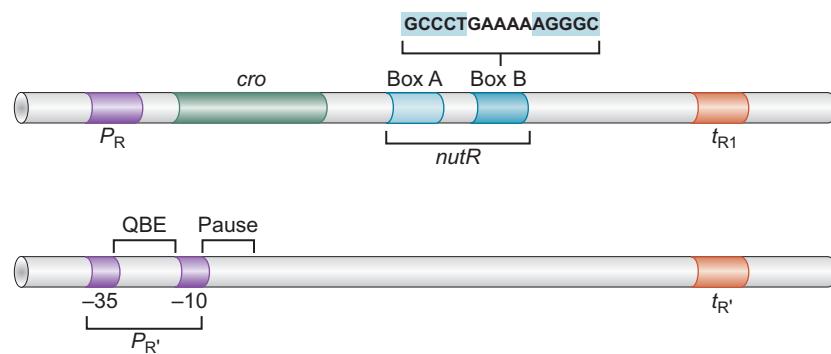
lysogen, the integrated phage DNA (the prophage) continues making repressor from  $P_{RM}$ . Any new  $\lambda$  genome entering that cell will at once be bound by repressor, giving no chance of lytic growth.

In one classic study, mutants of  $\lambda$  that formed clear plaques were isolated. These mutant phage are unable to form lysogens but still grow lytically. The  $\lambda$  clear mutations identified the three phage genes, called  $cl$ ,  $cII$ , and  $cIII$  (for clear I, II, and III). In other studies, so-called virulent ( $vir$ ) mutations were isolated. These mutations define the operator sites where  $\lambda$  repressor binds and were isolated by virtue of the fact that such phage can grow on lysogens. By analogy to the  $lac$  system, the  $cl$  mutants are comparable to the Lac repressor ( $lacI$ ) mutants;  $vir$  mutants are the equivalent of the  $lac$  operator ( $lacO$ ) mutants (see Box 18-2). Another revealing mutation was identified in a different experiment, this one a mutation in a host gene. The mutant is called  $hfl$  for high frequency of lysogeny. When infected with wild-type  $\lambda$ , this strain almost always forms lysogens, very rarely allowing the phage to grow lytically. This bacterial strain lacks the protease that degrades the  $\lambda$  CII protein (see text).

The transcripts controlled by the  $\lambda$  N and Q proteins are initiated perfectly well in the absence of those regulators. But the transcripts terminate a few hundred to a thousand nucleotides downstream from the promoter unless RNA polymerase has been modified by the regulator.  $\lambda$  N and Q proteins are therefore called antiterminators.

N protein regulates early gene expression by acting at three terminators: one to the left of the  $N$  gene itself, one to the right of  $cro$ , and one between genes  $P$  and  $Q$  (Figs. 18-21 and 18-32). Q protein has one target, a terminator 200 nucleotides downstream from the late gene promoter,  $P_{R'}$ , located between the  $Q$  and  $S$  genes (see Figs. 18-21 and 18-32). The late gene operon of  $\lambda$ , transcribed from  $P_{R'}$ , is remarkably large for a prokaryotic transcription unit: about 26 kb, a distance that takes about 10 minutes for RNA polymerase to traverse.

Our understanding of how antiterminators work is incomplete. Like other regulatory proteins, N and Q only work on genes that carry sequences specific for each regulator. Thus, N protein prevents termination in the early operons of  $\lambda$ , but not in other bacterial or phage operons. The specific

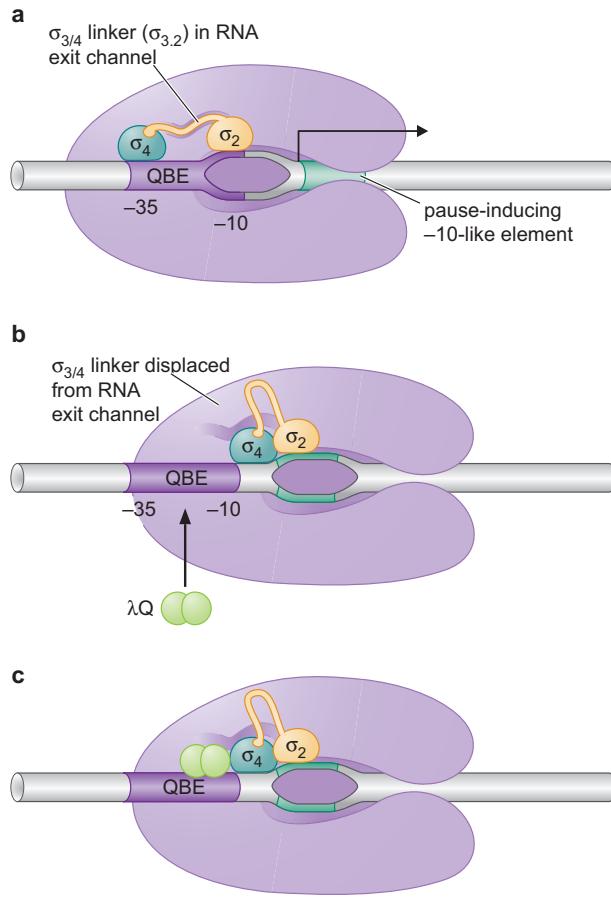


**FIGURE 18-32** Recognition sites and sites of action of the  $\lambda$  N and Q transcription antiterminators. The upper line shows the early rightward promoter  $P_R$  and its initial terminator,  $t_{R1}$ . The  $nutR$  site is divided into two regions, called Box A (7 bp) and Box B, separated by a spacer region of 8 bp. The sequence of Box B has dyad symmetry and forms a stem-loop structure once transcribed into RNA. The sequence of the RNA-like strand of  $nutR$  is shown above. The lower line shows the promoter  $P_{R'}$ , the sequences essential for Q protein function, and the terminator at which Q protein acts.

recognition sequences for antiterminators are not found in the terminators where they act, but instead occur somewhere between the promoter and the terminator. For N, those sites are called *nut* (for *N* utilization) sites which are 60 and 200 nucleotides downstream from  $P_L$  and  $P_R$  (see Fig. 18-32). But N does not bind to these sequences within DNA. Rather, N binds to RNA transcribed from DNA containing a *nut* sequence. Thus, once RNA polymerase has passed a *nut* site, N binds to the RNA and from there is loaded on to the polymerase itself. In this state, the polymerase is resistant to the terminators found just beyond the *N* and *cro* genes. λ N works together with the products of the bacterial genes *nusA*, *nusB*, *nusE*, and *nusG*. The NusA protein is an important cellular transcription factor. NusE is the small ribosomal subunit protein S10, but its role in N protein function is unknown. No cellular function of NusB protein is known. These proteins form a complex with N at the *nut* site, but N can work in their absence if present at high concentration, suggesting that it is N itself that promotes antitermination.

Unlike N protein, the λ Q protein recognizes DNA sequences (QBE) between the -10 and -35 regions of the late gene promoter ( $P_{R'}$ ) (see Fig. 18-32). In the absence of Q, polymerase binds  $P_{R'}$  and initiates transcription, only to pause after a mere 16 or 17 nucleotides; it then continues but terminates when it reaches the terminator ( $t_R'$ ) about 200 bp downstream. If Q is present, it binds to QBE once the polymerase has left the promoter and transfers from there to the nearby paused polymerase. With Q on board, the polymerase is then able to transcribe through  $t_{R'}$ .

The σ factor of polymerase is involved in Q function (see Fig. 18-33). First, the reason polymerase pauses just after initiation at  $P_{R'}$  is because it



**FIGURE 18-33** How λ Q engages RNA polymerase during early elongation. The sequence of events at λ  $P_{R'}$ . (a) The organization of polymerase elements in the initiation complex bound to λ  $P_{R'}$ .  $\sigma_{3/4}$  linker ( $\sigma_{3.2}$ ) is shown within the RNA exit channel. (b) The paused complex. The nascent transcript (not shown) has displaced  $\sigma_{3/4}$  linker from the exit channel. This is just before binding of λ Q. (c) λ Q is shown bound to the paused elongation complex. Further details of the process are provided in the text. (Courtesy of Ann Hochschild.)

encounters a sequence resembling the “–10” element of a promoter. Region 2 of  $\sigma$  typically recognizes that sequence, binding to base pairs in the non-template strand as described in Figure 13-8. It does the same at this pause site, halting polymerase progress temporarily. At the same time, the nascent transcript exiting the RNA channel of the enzyme facilitates rearrangements of the interface between  $\sigma$  and the core enzyme, revealing part of  $\sigma$  region 4 that was previously buried (as described in Chapter 13). This surface of  $\sigma$  is then bound by Q.

Why this new complex of polymerase and Q is impervious to the downstream terminator is still unclear. But we see that  $\sigma$  can be involved in regulation downstream from initiation and that  $\sigma$  region 4 can be a target for regulators working at initiation and afterward as well.

### Retroregulation: An Interplay of Controls on RNA Synthesis and Stability Determines *int* Gene Expression

The CII protein activates the promoter  $P_I$  that directs expression of the *int* gene, as well as the promoter  $P_{RE}$  responsible for repressor synthesis (see Fig. 18-21). The Int protein is the enzyme that integrates the phage genome into that of the host cell during formation of a lysogen (see Chapter 12). Therefore, upon infection, conditions favoring CII protein activity give rise to a burst of both repressor and integrase enzyme.

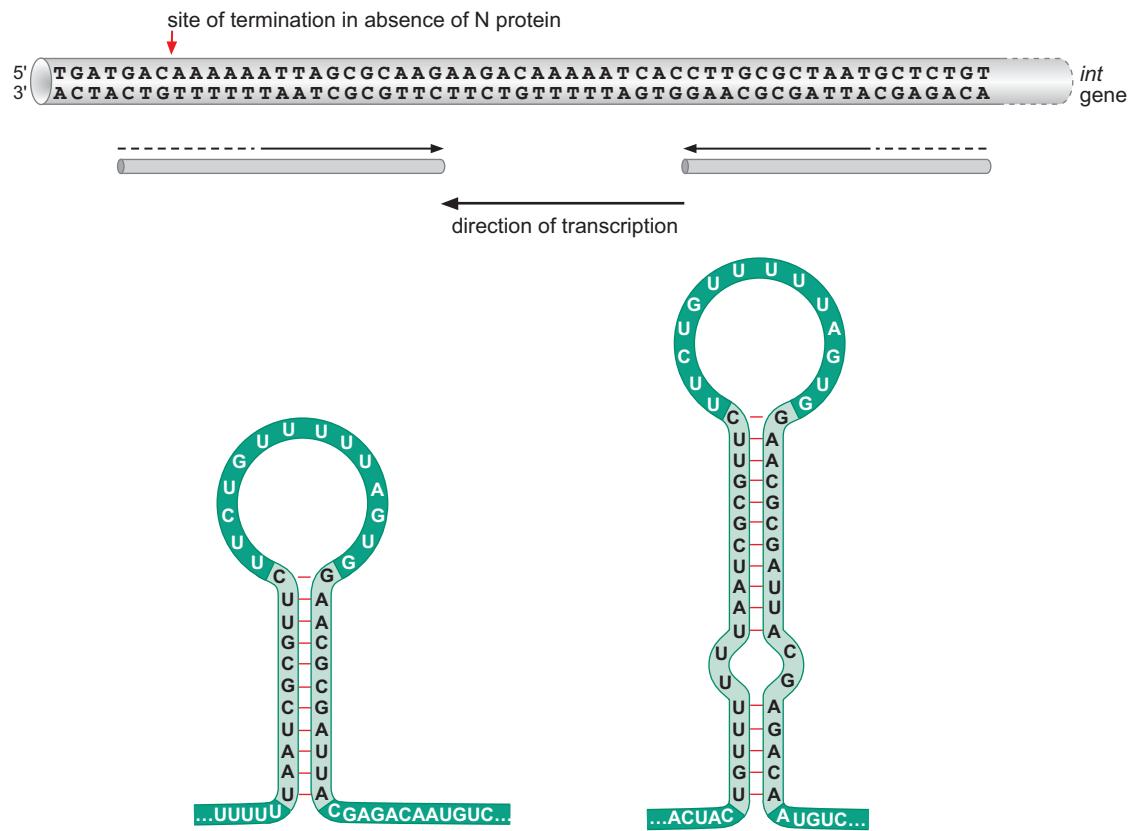
But the *int* gene is transcribed from  $P_L$  as well as from  $P_I$ , so one would have thought that integrase should be made even in the absence of CII protein. This does not happen. The reason is that *int* mRNA initiated at  $P_L$  is degraded by cellular nucleases, whereas mRNA initiated at  $P_I$  is stable and can be translated into integrase protein. This occurs because the two messages have different structures at their 3' ends.

RNA initiated at  $P_I$  stops at a terminator about 300 nucleotides after the end of the *int* gene; it has a typical stem-and-loop structure followed by six uridine nucleotides (Fig. 18-34; see Chapter 13, Fig. 13-13). When RNA synthesis is initiated at  $P_L$ , on the other hand, RNA polymerase is modified by the N protein and thus goes through and beyond the terminator. This longer mRNA can form a stem that is a substrate for nucleases. Because the site responsible for this negative regulation is downstream from the gene it affects, and because degradation proceeds backward through the gene, this process is called **retroregulation**.

The biological function of retroregulation is clear. When CII activity is low and lytic development is favored, there is no need for integrase enzyme; thus, its mRNA is destroyed. But when CII activity is high and lysogeny is favored, the *int* gene is expressed to promote recombination of the repressed phage DNA into the bacterial chromosome.

There is yet a further subtlety in this regulatory device. When a prophage is induced, it needs to make integrase (together with another enzyme, called excisionase; see Chapter 12) to catalyze reformation of free phage DNA by recombination out of the bacterial DNA; it must do this whether or not CII activity is high. Thus, under these circumstances, the phage must make stable integrase mRNA from  $P_L$  despite the antitermination activity of N protein. How is this achieved?

When the phage genome is integrated into the bacterial chromosome during the establishment of lysogeny, the phage attachment site at which recombination occurs is *between* the end of the *int* gene and those sequences encoding the extended stem from which mRNA degradation is begun (see Fig. 18-21). Thus, in the integrated form, the site causing degradation is removed from the end of the *int* gene, and so *int* mRNA made from  $P_L$  is stable.



**FIGURE 18-34** DNA site and transcribed RNA structures active in retroregulation of *int* expression. (Top) The DNA sequence; (below) the small cylinders show the symmetric sequences that form hairpins in RNA. (Bottom) The structure on the left shows the terminator formed in RNA transcribed from  $P_r$  without antitermination by N protein, which is resistant to degradation by nucleases. The structure on the right shows an extended loop formed in RNA transcribed from  $P_L$  under the influence of N protein antiterminator, which is a target for cleavage by RNase III and degradation by nucleases.

## SUMMARY

A typical gene is switched on and off in response to the need for its product. This regulation is predominantly at the level of transcription initiation. Thus, for example, in *E. coli*, a gene encoding the enzyme that metabolizes lactose is transcribed at high levels only when lactose is available in the growth medium. Furthermore, when glucose (a better energy source) is also available, the gene is not expressed even when lactose is present.

Signals, such as the presence of a specific sugar, are communicated to genes by regulatory proteins. These are of two types: *activators*, positive regulators that switch genes on, and *repressors*, negative regulators that switch genes off. Typically, these regulators are DNA-binding proteins that recognize specific sites at or near the genes they control.

Activators, in the simplest (and most common) cases, work on promoters that are inherently weak; that is, RNA polymerase binds to the promoter (and thus initiates transcription) poorly in the absence of any regulator. An activator binds to DNA with one surface and with another surface

binds polymerase and recruits it to the promoter. This process is an example of cooperative binding and is sufficient to stimulate transcription.

Repressors can inhibit transcription by binding to a site that overlaps the promoter, thereby blocking RNA polymerase binding. Repressors can work in other ways as well: for example, by binding to a site beside the promoter and by interacting with polymerase bound at the promoter, inhibiting initiation.

The *lac* genes of *E. coli* are controlled by an activator and a repressor that work in the simplest way just outlined. CAP, in the absence of glucose, binds DNA near the *lac* promoter and, by recruiting polymerase to that promoter, activates expression of those genes. The Lac repressor binds a site that overlaps the promoter and shuts off expression in the absence of lactose.

Another way in which RNA polymerase is recruited to different genes is by the use of alternative  $\sigma$  factors. Thus, different  $\sigma$  factors can replace the most prevalent one ( $\sigma^{70}$  in *E. coli*)

and direct the enzyme to promoters of different sequences. Examples include  $\sigma^{32}$ , which directs transcription of genes in response to heat shock, and  $\sigma^{54}$ , which directs transcription of genes involved in nitrogen metabolism. Phage SPO1 uses a series of alternative  $\sigma$ s to control the ordered expression of its genes during infection.

There are, in bacteria, examples of other kinds of transcriptional activation as well. At some promoters, RNA polymerase binds efficiently unaided and forms a stable, but inactive, closed complex. This closed complex does not spontaneously undergo transition to the open complex and initiate transcription. At such a promoter, an activator must stimulate the transition from a closed to open complex.

Activators that stimulate this kind of promoter work by allosteric: they interact with the stable, closed complex and induce a conformational change that causes transition to the open complex. In this chapter, we saw two examples of transcriptional activators working by allosteric. In one case, the activator (NtrC) interacts with the RNA polymerase (bearing  $\sigma^{54}$ ) bound in a stable closed complex at the *glnA* promoter, stimulating transition to the open complex. In the other example, the activator (MerR) induces a conformational change in the *merT* promoter DNA.

In all the cases that we have considered, the regulators themselves are controlled allosterically by signals: the shape of the regulator changes in the presence of its signal. In one

state, it can bind DNA, and in the other state, it cannot. Thus, for example, the Lac repressor is controlled by the ligand allolactose (a product made from lactose). When allolactose binds repressor, it induces a change in the shape of that protein; in that state, the protein cannot bind DNA.

Gene expression can be regulated at steps after transcription initiation. For example, regulation can be at the level of transcriptional elongation. Examples considered in this chapter were antitermination by the N and Q proteins of bacteriophage  $\lambda$ . The  $\lambda$  proteins N and Q load on to RNA polymerases initiating transcription at certain promoters in the phage genome. Once modified in this way, the enzyme can pass through certain transcriptional terminator sites that would otherwise block expression of downstream genes.

We concluded this chapter with a detailed discussion of how bacteriophage  $\lambda$  chooses between two alternative modes of propagation. Several of the strategies of gene regulation encountered in this system turn out to operate in other systems as well, including, as discussed in later chapters, those that govern the development of animals—for example, the use of cooperative binding to give stringent on/off switches and the use of separate pathways for establishing and maintaining expression of genes. We also considered how complex and intricate gene networks like that found in  $\lambda$  might have evolved from more rudimentary earlier versions.

## BIBLIOGRAPHY

### Books

- Echols H. 2001. *Operators and promoters: The study of molecular biology and its creators*. University of California Press, Berkeley, CA.  
 Müller-Hill B. 1996. *The lac operon*. de Gruyter, Berlin.  
 Ptashne M. 2005. *A genetic switch: Phage lambda revisited*, 3rd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.  
 Ptashne M. and Gann A. 2002. *Genes & signals*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

### Activation and Repression

- Dodd I.B., Shearwin K.E., and Egan J.B. 2005. Revisited gene regulation in bacteriophage  $\lambda$ . *Curr. Opin. Genet. Dev. Biol.* **15**: 145–152.  
 Dove S.L., Darst S.E., and Hochschild A. 2003. Region 4 of  $\sigma$  as a target for transcription regulation. *Mol. Microbiol.* **48**: 863–874.  
 Gann A. 2010. Jacob and Monod: From operons to EvoDevo. *Curr. Biol.* **20**: R718–R723.  
 Gottesmann M. and Wesiberg R. 2004. Little lambda, who made thee? *Microbiol. Mol. Biol. Rev.* **68**: 796–813.  
 Hochschild A. and Dove S.L. 1998. Protein–protein contacts that activate and repress prokaryotic transcription. *Cell* **92**: 597–600.  
 Huffman J.L. and Brennan R.G. 2002. Prokaryotic transcription regulators: More than just the helix-turn-helix motif. *Curr. Opin. Struct. Biol.* **12**: 98–106.  
 Jacob F. and Monod J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**: 318–356.  
 Lawson C.L., Swigon D., Murakami K.S., Darst S.A., Berman H.M., and Ebright R.H. 2004. Catabolite activator protein: DNA binding and transcription activation. *Curr. Opin. Struct. Biol.* **14**: 10–20.  
 Little J. W. 2010. Evolution of complex gene regulatory circuits by addition of refinements. *Curr. Biol.* **20**: R724–R734.

- Magasanik B. 2000. Global regulation of gene expression. *Proc. Natl. Acad. Sci.* **97**: 14044–14045.  
 Murray N.E. and Gann A. 2007. What has phage lambda ever done for us? *Curr. Biol.* **17**: R305–R312.  
 Ng W. L. and Bassler B.L. 2009. Bacterial quorum-sensing network architectures. *Annu. Rev. Genet.* **43**: 197–222.  
 Oppenheim A.B., Oren Kobiler O., Stavans J., Court D.L., and Adhya S. 2005. Switches in bacteriophage lambda development. *Ann. Rev. Genet.* **39**: 409–429.  
 Payankaulam S., Li L.M., and Arnosti D.N. 2010. Transcriptional repression: Conserved and evolved features. *Curr. Biol.* **20**: R764–R771.  
 Ptashne M. 2006. Lambda's switch: Lesson from a module swap. *Curr. Biol.* **16**: R459–R462.  
 Rappas M., Bose D., and Zhang X. 2007. Bacterial enhancer-binding proteins: Unlocking  $\sigma^{54}$ -dependent gene transcription. *Curr. Opin. Struct. Biol.* **17**: 110–116.  
 Rojo F. 2001. Mechanisms of transcriptional repression. *Curr. Opin. Microbiol.* **4**: 145–151.  
 Roy S., Garges S., and Adhya S. 1998. Activation and repression of transcription by differential contact: Two sides of a coin. *J. Biol. Chem.* **273**: 14059–14062.  
 Schleif R. 2003. AraC protein: A love–hate relationship. *Bioessays* **25**: 274–282.  
 Summers A.O. 2009. Damage control: Regulating defenses against toxic metals and metalloids. *Curr. Opin. Microbiol.* **12**: 138–144.

### DNA Binding, Cooperativity, and Allostery

- Hochschild A. 2002. The switch: cI closes the gap in autoregulation. *Curr. Biol.* **12**: R87–R89.

- Lewis M. 2005. The lac repressor. *Crit. Rev. Biol.* **328**: 521–548.
- Luscombe N.M., Austin S.E., Berman H.M., and Thornton J.M. 2000. An overview of the structures of protein–DNA complexes. *Genome Biol.* **1**: REVIEWS001.
- Monod J. 1966. From enzymatic adaptation to allosteric transitions. *Science* **154**: 475–483.
- Vilar J.M.G. and Saiz L. 2005. DNA looping in gene regulation: From the assembly of macromolecular complexes to the control of transcriptional noise. *Curr. Opin. Genet. Dev. Biol.* **15**: 136–144.
- Gottesman M. 1999. Bacteriophage λ: The untold story. *J. Mol. Biol.* **293**: 177–180.
- Roberts J.W., Yarnell W., Bartlett E., Guo J., Marr M., Ko D.C., Sun H., and Roberts C.W. 1998. Antitermination by bacteriophage λ Q protein. *Cold Spring Harb. Symp. Quant. Biol.* **63**: 319–325.
- Roberts J.W., Shankar S., and Filter J.J. 2008. RNA polymerase elongation factors. *Annu. Rev. Microbiol.* **62**: 211–233.
- Santangelo T.J. and Artsimovitch I. 2011. Termination and antitermination: RNA polymerase runs a stop sign. *Nat. Rev. Microbiol.* **9**: 319–329.

## QUESTIONS

## MasteringBiology®

For instructor-assigned tutorials and problems, go to *MasteringBiology*.

For answers to even-numbered questions, see Appendix 2: Answers.

**Question 1.** The most common level of regulation of gene expression occurs at transcription initiation. Explain why.

**Question 2.** Provide three examples of allosteric regulation from this chapter.

**Question 3.** The loss of repression for the *E. coli* lac operon requires β-galactosidase (encoded by *lacZ*).

- A. Explain the role of β-galactosidase in the loss of repression of the *lac* operon.
- B. In the absence of glucose and the *lac* operon in the repressed state, predict how β-galactosidase is present to carry out the necessary function to end repression of the *lac* operon.

**Question 4.** In a genetic screen, researchers isolated mutants of *E. coli* that constitutively expressed the genes from the *araBAD* operon.

- A. Describe what constitutive expression means in terms of the *araBAD* operon.
- B. Give an example of a mutation that could lead to constitutive expression of the *araBAD* genes. (Name the region of DNA or gene encoding a specific protein.)

**Question 5.** In the laboratory, you want to purify a protein that is normally toxic in *E. coli* cells. Your advisor suggests cloning the gene encoding your protein into an expression vector that uses the *araBAD* promoter. Why is it ideal to use the *araBAD* promoter for expression of your gene of interest in *E. coli* cells?

**Question 6.** Given the following mutants and conditions, predict the expression of the *lacZ* gene (no expression, basal level of expression, or activated level of expression).

- A. A mutant of *E. coli* that has a mutation in the operator of the *lac* operon that prevents the repressor from binding
  - a. In the presence of glucose, absence of lactose
  - b. In the presence of glucose, presence of lactose

## Antitermination

- Gottesman M. 1999. Bacteriophage λ: The untold story. *J. Mol. Biol.* **293**: 177–180.
- Roberts J.W., Yarnell W., Bartlett E., Guo J., Marr M., Ko D.C., Sun H., and Roberts C.W. 1998. Antitermination by bacteriophage λ Q protein. *Cold Spring Harb. Symp. Quant. Biol.* **63**: 319–325.
- Roberts J.W., Shankar S., and Filter J.J. 2008. RNA polymerase elongation factors. *Annu. Rev. Microbiol.* **62**: 211–233.
- Santangelo T.J. and Artsimovitch I. 2011. Termination and antitermination: RNA polymerase runs a stop sign. *Nat. Rev. Microbiol.* **9**: 319–329.

- c. In the absence of glucose, absence of lactose
- d. In the absence of glucose, presence of lactose

**B.** A mutant of *E. coli* that has a mutation in the promoter of the *lac* operon (for the *lacZ*, *lacY*, and *lacA* genes) that prevents RNA polymerase from binding

- a. In the presence of glucose, absence of lactose
- b. In the presence of glucose, presence of lactose
- c. In the absence of glucose, absence of lactose
- d. In the absence of glucose, presence of lactose

**Question 7.** List three mechanisms for transcriptional repression in prokaryotes and in each case an example of a protein that uses the mechanism.

**Question 8.** You are studying the *E. coli* ZntR protein, a homolog to MerR, which responds to Zn(II) to regulate transcription of *zntA*, a gene that encodes a protein that aids in Zn(II) detoxification. What are two questions you would design your experiments to answer to support or refute the hypothesis that ZntR uses a mechanism similar to MerR to activate transcription of *zntA*?

**Question 9.** Given a loss-of-function mutation in the tetramerization domain of the λ repressor and using Figure 18-24, predict the specific functional effect of the mutation on the λ repressor.

### Question 10.

- A. Explain why it is to the advantage of bacteriophage λ to tightly regulate the level of λ repressor made in lysogenic *E. coli* cells.
- B. Describe the mechanism of negative autoregulation by λ repressor.

**Question 11.** Review Box 18-4 Figure 1. For the following DNA-binding proteins, do you expect a plot of the % DNA bound versus protein concentration to look like the red line or the black line? Explain why.

- A. NtrC and the regulatory DNA upstream of the *glnA* gene
- B. CAP in the presence of RNA polymerase and the DNA upstream of the *lac* operator

- C. The amino domain of the  $\lambda$  repressor (carboxyl domain removed) and the  $\lambda$  repressor operator sites

**Question 12.** You discover a new operon that is regulated by a repressor in a prokaryotic species. Assuming the repressor binds an operator site, design an in vitro experiment to identify the specific region where the repressor binds DNA under conditions similar to those normally found for repression in the cell.

**Question 13.** Describe how alternative  $\sigma$  factors play a role in the regulation of transcription in prokaryotes.

**Question 14.** Researchers studying  $\lambda$  repressor binding to the three binding sites in the right operator produced the data in the table below. In the experiment, they performed a DNase I protection assay (footprinting assay) using the DNA containing all three binding sites for a range of repressor concentrations. From this, they calculated the relative concentration of repressor dimers required to occupy a specific binding site on half of the DNA molecules present (values given in table).

Relative Repressor Concentration for Each Site			
DNA	$O_{R3}$	$O_{R2}$	$O_{R1}$
Wild-type	25	2	1
Mutant X	5	5	—
Mutant Y	25	—	2

- A. Based on the data for wild-type DNA and the information in Chapter 18, explain why the relative concentrations for  $O_{R1}$  and  $O_{R2}$  are almost the same despite the fact that  $O_{R1}$  has a 10-fold higher affinity for  $\lambda$  repressor than  $O_{R2}$  does (information from this chapter).
- B. Based on the data for wild-type DNA and the information in Chapter 18, explain why the relative concentration of repressor needed to bind  $O_{R3}$  is much higher than for  $O_{R1}$  and  $O_{R2}$ .
- C. Based on the data for mutant DNAs X and Y, identify the binding site ( $O_{R1}$ ,  $O_{R2}$ , or  $O_{R3}$ ) that included the mutation in each respective mutant. Explain your choices.

- D. Given what you know from this chapter, explain why the relative repressor concentration required to bind  $O_{R3}$  goes down to 5 for Mutant X relative to wild-type DNA in terms of the mechanism for  $\lambda$  repressor binding DNA.

Data adapted from Johnson et al. (1979. *Proc. Natl. Acad. Sci.* **76**: 5061–5065).

**Question 15.** The regulation of a novel operon in *E. coli* involves two operators that sandwich a promoter and three structural genes. RNA polymerase transcribes the structural genes from the promoter, and a specific repressor represses transcription from that promoter. In the presence of the relevant signal, the repressor binding to DNA is disrupted and repression is alleviated.

To study the mechanism of repression, researchers created a reporter containing the promoter but replaced the operators with *lac* operator sites. Wild-type *lac* repressor was able to repress expression from this construct in the cell. In vitro, researchers visualized protein–DNA complexes in electron micrographs using the reporter construct after incubation with either wild-type Lac repressor or a mutant of the Lac repressor that binds to the operator sites, dimerizes, but fails to tetramerize. The scored data of their observations are given in the table below.

Protein	Observed DNA-Protein Complexes			
	Free DNA	Single binding	Tandem binding	Loop
Wild-type Lac repressor	53	29	3	15
Mutant Lac repressor	42	44	14	0

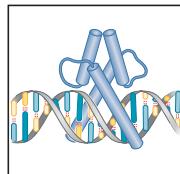
In addition, it was shown that the mutant Lac repressor was *not* able to repress expression from the reporter construct *in vivo*, although it was able to repress the endogenous *lac* operon.

Given the data in the table, propose a model for repression stating how the data supports your model. What other information would you want to know about the novel repression protein to support this model?

Data adapted from Mandal et al. (1990. *Genes Dev.* **4**: 410–418).

*This page intentionally left blank*

CHAPTER 19



# Transcriptional Regulation in Eukaryotes

IN EUKARYOTIC CELLS, EXPRESSION OF A GENE can be regulated at all those steps seen in bacteria and a few additional steps as well. The additional steps include **splicing**, as we saw in Chapter 14. In many cases, a given transcript can be spliced in alternative ways to generate different products, and this too can be regulated.

But just as in bacteria, it is the initiation of transcription that is the most pervasively regulated step. Indeed, many of the principles we encountered when considering how transcription is regulated in bacteria apply to regulation of transcription in eukaryotes as well. These principles are laid out in the first few pages of the chapter on prokaryotic transcriptional regulation (Chapter 18) and in the summary at the end of that chapter. We urge readers who have not previously (or recently) read that chapter to look at those passages before continuing with this chapter.

We have also already seen that the eukaryotic transcriptional machinery is more elaborate than its bacterial counterpart (Chapter 13). This is particularly true of the RNA polymerase II machinery—that which transcribes protein-encoding genes and most regulatory RNA genes. Despite this added complexity, transcription is once again regulated by activators and repressors, DNA-binding proteins that help or hinder transcription initiation at specific genes in response to appropriate signals. There are, however, additional features of eukaryotic cells and genes that complicate the actions of these regulatory proteins. We begin by summarizing the two most significant of these additional complexities.

**Nucleosomes and Their Modifiers** As discussed in Chapter 8, the genome of a eukaryote is wrapped in proteins called **histones** to form **nucleosomes**. Thus, the transcriptional machinery is presented with a partially concealed substrate. This condition reduces the expression of many genes in the absence of regulatory proteins. Eukaryotic cells also contain several enzymes that rearrange, or chemically modify, histones. These modifications alter nucleosomes in ways that affect how easily the transcriptional machinery—and DNA-binding proteins in general—can bind and operate. Thus, nucleosomes present a problem not faced in bacteria, but their modification

## OUTLINE

- Conserved Mechanisms of Transcriptional Regulation from Yeast to Mammals, 659
  - Recruitment of Protein Complexes to Genes by Eukaryotic Activators, 665
  - Signal Integration and Combinatorial Control, 675
  - Transcriptional Repressors, 681
  - Signal Transduction and the Control of Transcriptional Regulators, 682
  - Gene “Silencing” by Modification of Histones and DNA, 687
  - Epigenetic Gene Regulation, 694
- Visit Web Content for Structural Tutorials and Interactive Animations

and remodeling also offer new opportunities for regulation. The ability of transcription factors or RNA polymerase to interact with regulatory DNAs often requires displacement of positioned nucleosomes.

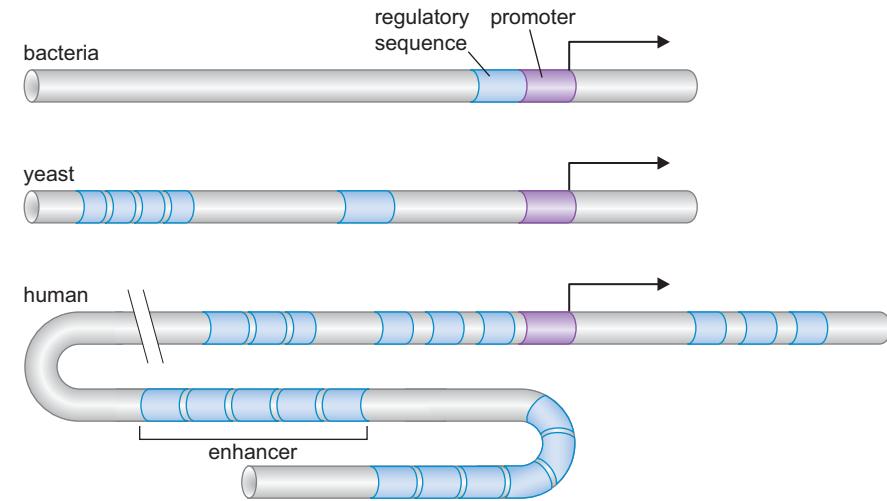
**More Regulators and More Extensive Regulatory Sequences** A further difference between eukaryotes and prokaryotes is the number of regulatory proteins that control a typical gene, as reflected in the number and arrangement of regulator binding sites associated with that gene (Fig. 19-1). As in bacteria, individual regulators bind short sequences, but in eukaryotes, these binding sites are often more numerous and positioned further from the start site of transcription than they are in bacteria. We call the region at the gene where the transcriptional machinery binds the **promoter**, the individual binding sites **regulator binding sites**, and the stretch of DNA encompassing the complete collection of regulator binding sites for a given gene the **regulatory sequences**.

The expansion of regulatory sequences—that is, the increase in the number of binding sites for regulators at a typical gene—is most striking in multicellular organisms such as *Drosophila* and mammals. This situation reflects the more extensive signal integration found in those organisms—the tendency for more signals to regulate a given gene. We saw examples of signal integration in bacteria (Chapter 18), but those examples typically involved just two different regulators integrating two signals to control a gene (e.g., glucose and lactose at the *lac* genes). But not all eukaryotes have extensive signal integration: yeast have less signal integration than multicellular organisms—indeed, they are not so different from bacteria in this regard—and their genes have less extensive regulatory sequences than those of multicellular eukaryotes (Fig. 19-1). Unlike higher eukaryotes, yeast also lack “action at a distance.” Yeast regulatory sequences are usually located within a few hundred base pairs of their promoters.

In multicellular organisms, regulatory sequences can spread thousands of nucleotides from the promoter—both upstream and downstream—and can be made up of tens of regulator binding sites. Often, these binding sites are grouped in units called **enhancers**, and a given enhancer binds regulators responsible for activating the gene at a given time and place. Alternative enhancers bind different groups of regulators and control expression of the same gene at different times and places in response to different signals.

Having more extensive regulatory sequences means that some regulators bind sites far from the genes they control, in some cases hundreds of

**FIGURE 19-1** The regulatory elements of bacterial, yeast, and human genes. Illustrated is the increasing complexity of regulatory sequences from a simple bacterial gene controlled by a repressor to a human gene controlled by multiple activators and repressors. In each case, a promoter is shown at the site where transcription is initiated. Although this is accurate for the bacterial case, in the eukaryotic examples, transcription initiates somewhat downstream from where the transcription machine binds (see Chapter 13). Some groups of regulatory binding sites in the human regulatory sequences represent enhancers, as shown in one case.



kilobases or more. In fact, a critical enhancer of the Sonic hedgehog gene in mammals maps 1 Mb away from the transcription start site. How can regulators act from such a distance? In bacteria, we encountered DNA-binding proteins that communicate over a range of a few kilobases: λ repressors at  $O_R$  interacting with those at  $O_L$ ; and NtrC, which can activate the *glnA* gene from sites placed 1 kb or more upstream. In those examples of mild “action at a distance,” the intervening DNA loops out to accommodate the interaction between the proteins. The same mechanism explains action at a distance in many, if not all, eukaryotic cases as well, although in some cases, the distances over which proteins work are very large, and it is not clear how the looping occurs. In some cases, “tethering” sequences located immediately upstream of the promoter can recruit remote enhancers.

Activation at a distance raises another problem. When bound at an enhancer, there may be several genes within range of an activator, yet a given enhancer typically regulates only one gene. Other regulatory sequences—called **insulators** or **boundary elements**—are found between enhancers and some promoters. Insulators block activation of the promoter by activators bound at the enhancer. These elements, although still poorly understood, ensure that activators do not work indiscriminately.

## CONSERVED MECHANISMS OF TRANSCRIPTIONAL REGULATION FROM YEAST TO MAMMALS

---

In this chapter, we consider transcriptional regulation in organisms ranging from single-celled yeast to mammals. All of these organisms have both the more elaborate transcriptional machinery and the nucleosomes and their modifiers typical of eukaryotes. Therefore, it is not surprising that many of the basic features of gene regulation are the same in all eukaryotes. Because yeast is the most amenable to a combination of genetic and biochemical dissection, much of the information regarding how activators and repressors work comes from that organism. In addition, critical to the generality of the conclusions, when expressed in a mammalian cell, a typical yeast activator can stimulate transcription. This is tested using a **reporter gene**. The reporter gene consists of binding sites for the yeast activator inserted upstream of the promoter of a gene whose expression level is readily measured (as we shall discuss later).

We shall see that the typical eukaryotic activator works in a manner similar to that of the simplest bacterial case: it has separate DNA-binding and activating regions and activates transcription by recruiting protein complexes to specific genes. In contrast, repressors work in a variety of ways, some different from anything we encountered in bacteria. These novel repression mechanisms include examples of what is called **gene silencing**, in which nucleosome and DNA modifiers are recruited to regions of the genome where they act to keep genes switched off, sometimes over large stretches of DNA. There are also examples of “short-range” repression, whereby a sequence-specific repressor inhibits activators bound to neighboring sites within an enhancer.

Despite having so much in common, not all details of gene regulation are the same in all eukaryotes. Most importantly, as we have mentioned, a typical yeast gene has less-extensive regulatory sequences than its multicellular counterpart. Therefore, we must look to higher organisms to see how the basic mechanisms of gene regulation are extended to accommodate more complicated cases of signal integration and combinatorial control. In this

chapter, we restrict discussion to transcriptional regulation mediated by proteins (and their modifications). In the following chapter, we discuss regulation of gene expression mediated by RNA molecules.

### Activators Have Separate DNA-Binding and Activating Functions

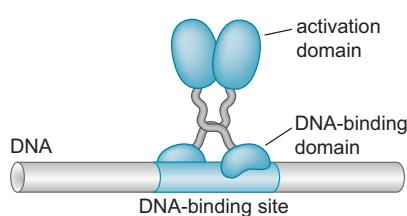
In bacteria, we saw that a typical activator, such as CAP, has separate DNA-binding and activating functions. We described the genetic demonstration of this: positive control (or *pc*) mutants bind DNA normally but are defective in activation. Eukaryotic activators have separate DNA-binding and activating regions as well. Indeed, in this case, the two surfaces are very often on separate domains of the protein.

We take as an example the most studied eukaryotic activator, Gal4 (Fig. 19-2). This protein activates transcription of the galactose genes in the yeast *Saccharomyces cerevisiae*. These genes, like their bacterial counterparts, encode enzymes required for galactose metabolism. One such gene is *GAL1*. Gal4 binds to four sites located 275 bp upstream of *GAL1* (Fig. 19-3). When bound there, in the presence of galactose, Gal4 activates transcription of the *GAL1* gene 1000-fold.

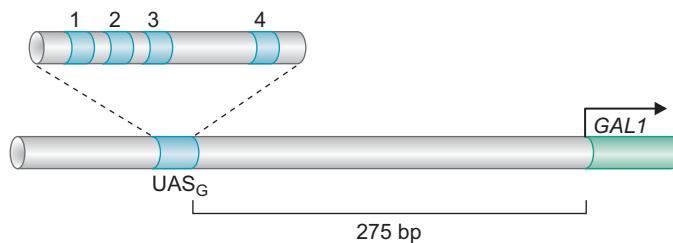
The separate DNA-binding and activating regions of Gal4 were revealed in two complementary experiments. In one experiment, expression of a fragment of the *GAL4* gene—encoding the amino-terminal one-third of the activator—produced a protein that bound DNA normally but did not activate transcription. This protein contained the DNA-binding domain but lacked the activating region and was therefore formally comparable to the *pc* mutants of bacterial activators (Fig. 19-4a).

In a second experiment, a hybrid gene was constructed that encoded the carboxy-terminal two-thirds of Gal4 fused to the DNA-binding domain of a bacterial repressor protein, LexA. The fusion protein was expressed in yeast together with a reporter plasmid bearing LexA-binding sites upstream of the *GAL1* promoter. The fusion protein activated transcription of this reporter (Fig. 19-4b). This experiment shows that activation is not mediated by DNA binding alone, as it was in one of the alternative mechanisms we encountered in bacteria—activation by MerR (Chapter 18, Fig. 18-17). Instead, the DNA-binding domain serves merely to tether the activating region to the promoter just as in the most common mechanism we saw in bacteria.

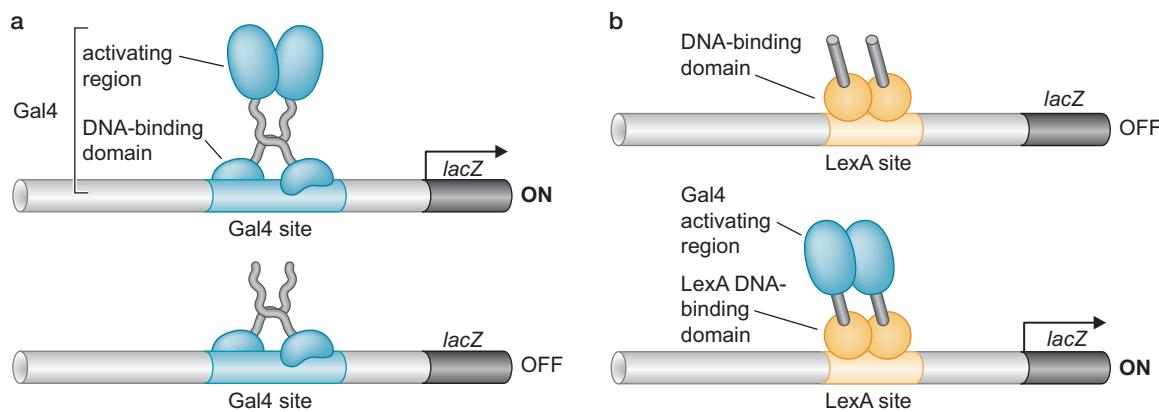
Many other eukaryotic activators have been examined in similar experiments, and whether from yeast, flies, or mammals, the same story typically holds: DNA-binding domains and activating regions are separable. In some cases, they are even carried on separate polypeptides: one has a DNA-binding domain, the other has an activating region, and they form a complex



**FIGURE 19-2** Gal4 bound to its site on DNA. The yeast activator Gal4 binds as a dimer to a 17-bp site on DNA. The DNA-binding domain of the protein is separate from the region of the protein containing the activating region (the activation domain).



**FIGURE 19-3** The regulatory sequences of the yeast *GAL1* gene. The UAS<sub>G</sub> (upstream activating sequence for *GAL1*) contains four binding sites, each of which binds a dimer of Gal4 as shown in Figure 19-2. Although not shown here, there is another site between these and the *GAL1* gene that binds a repressor called Mig1, which is discussed later (see Fig. 19-23).



**FIGURE 19-4** Domain swap experiment. (a) The DNA-binding domain of Gal4, without that protein’s activation domain, can still bind DNA but cannot activate transcription. In another experiment (not shown), the activation domain, without the DNA-binding domain, also does not activate transcription. (b) Attaching the activation domain of Gal4 to the DNA-binding domain of the bacterial protein LexA creates a hybrid protein that activates transcription of a gene in yeast as long as that gene bears a binding site for LexA. Expression is measured using a reporter plasmid in which the *GAL1* promoter is fused to the *E. coli* *lacZ* gene whose product ( $\beta$ -galactosidase) is readily assayed in yeast cells. Levels of expression from the *GAL1* promoter in response to the various activator constructs can therefore easily be measured. Similar reporter plasmids are used in many experiments in this chapter.

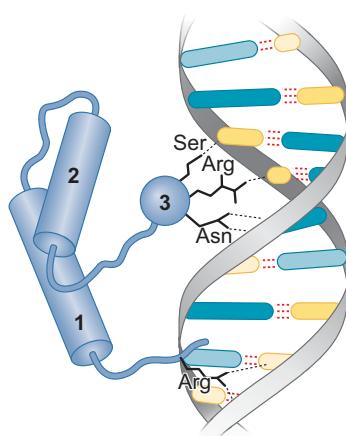
on DNA. An example of this is the herpes virus activator VP16, which interacts with the Oct1 DNA-binding protein found in cells infected by that virus. Another example is the *Drosophila* activator Notch, described in Chapter 21. The separable nature of DNA-binding and activating regions of eukaryotic activators is the basis for a widely used assay to detect protein–protein interactions (see Box 19-1, The Two-Hybrid Assay).

#### Eukaryotic Regulators Use a Range of DNA-Binding Domains, But DNA Recognition Involves the Same Principles as Found in Bacteria

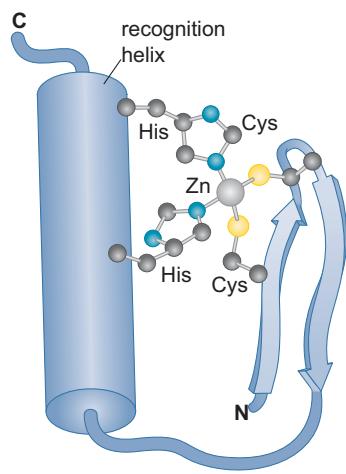
The experiments described above show that a bacterial DNA-binding domain can function in place of the DNA-binding domain of a eukaryotic activator. That result suggests there is no fundamental difference in the ways DNA-binding proteins from these organisms recognize their sites.

Recall from the previous chapter (and from Chapter 6) that most bacterial regulators bind as dimers to DNA target sequences that are twofold rotationally symmetric: each monomer inserts an  $\alpha$  helix into the major groove of the DNA over one-half of the site and detects the edges of base pairs found there. Binding typically requires no significant alteration in the structure of either the protein or the DNA. The vast majority of bacterial regulatory proteins use the so-called helix-turn-helix motif. This motif, as we saw, consists of two  $\alpha$  helices separated by a short turn. One helix (the recognition helix) fits in the major groove of the DNA and recognizes specific base pairs. The other helix makes contacts with the DNA backbone, positioning the recognition helix properly and increasing the strength of binding (see Chapter 6, Figs. 6-13 and 6-14). As discussed in Chapter 21 (and later in this chapter), homeobox genes important for animal development encode helix-turn-helix DNA-binding proteins.

The same basic principles of DNA recognition are used in most eukaryotic cases, despite variations in detail. Thus, proteins often bind as dimers and



**FIGURE 19-5** DNA recognition by a homeodomain. The homeodomain consists of three  $\alpha$  helices, of which two (helices 2 and 3 in this figure) form the structure resembling the helix-turn-helix motif (compare with Fig. 6-13, for example). Thus, helix 3 is the recognition helix, and, as shown, it is inserted into the major groove of DNA. Amino acid residues along its outer edge make specific contacts with base pairs. In the case shown, the yeast  $\alpha 2$  transcriptional repressor, an arm extending from helix 1 makes additional contacts with base pairs in the minor groove. (Adapted, with permission, from Wolberger C. et al. 1991. *Cell* 67: 517–528. © Elsevier.)



**FIGURE 19-6** Zinc finger domain. The  $\alpha$  helix on the left of the structure is the recognition helix, and it is presented to the DNA by the  $\beta$  sheet on the right. The zinc is coordinated by the two histidine residues in the  $\alpha$  helix and two cysteine residues in the  $\beta$  sheet as shown. This arrangement stabilizes the structure and is essential for DNA binding. See Fig. 6-15 for more details. (Adapted from Lee M.S. et al. 1989. *Science* 245: 635–637.)

recognize specific DNA sequences using an  $\alpha$  helix inserted into the major groove. As we just saw, homeobox proteins present the recognition helix as part of a structure very like the helix-turn-helix domain; others present the recognition helix within quite different domain structures. In a variation not seen in prokaryotes, several of the regulatory proteins we encounter in eukaryotes bind DNA as **heterodimers** and, in some cases, even as monomers (although often only when binding cooperatively with other proteins). Heterodimers extend the range of DNA-binding specificities available: when each monomer has a different DNA-binding specificity, the site recognized by the heterodimer is different from that recognized by either homodimer. A brief survey of some eukaryotic DNA-binding domains follows.

**Homeodomain Proteins** The **homeodomain** is a class of helix-turn-helix DNA-binding domains and recognizes DNA in essentially the same way as those bacterial proteins (Fig. 19-5). Homeodomains from different proteins are structurally very similar: Not only is the recognition helix similar, but so is the surrounding protein structure that presents that helix to the DNA. In contrast, as we saw in the previous chapter, the detailed structures of helix-turn-helix domains vary to a greater extent. Homeodomain proteins are found in all eukaryotes. They were first discovered in *Drosophila*, where they control many basic developmental programs, just as they do in higher eukaryotes; we consider their functions in that regard in Chapter 21. Homeodomain proteins are also found in yeast; some of the mating-type control genes discussed later encode homeodomain proteins. Indeed, it is the structure of one of these that is shown in Figure 19-5. Many homeodomain proteins bind DNA as heterodimers.

**Zinc-Containing DNA-Binding Domains** There are various forms of DNA-binding domains that incorporate a zinc atom(s). These include the classically defined **zinc finger** proteins (described and depicted in Chapter 6, Fig. 6-15) and the related **zinc cluster** domain found in the yeast activator Gal4. The zinc atom interacts with cysteine and histidine residues and serves a structural role essential for integrity of the DNA-binding domain (Fig. 19-6). The DNA is again recognized by an  $\alpha$  helix inserted into the major groove (Fig. 6-15b). Some proteins contain two or more zinc finger domains linked end-to-end. Each finger inserts an  $\alpha$  helix into the major groove, extending—with each additional finger—the length of the DNA sequence recognized and thus the affinity of binding.

There are other DNA-binding domains that use zinc. In those cases, the zinc is coordinated by four cysteine residues and stabilizes a rather different DNA-recognition motif, one resembling a helix-turn-helix. An example of this is found in the glucocorticoid receptor, which regulates genes in response to certain hormones in mammals.

**Leucine Zipper Motif** The leucine zipper motif combines dimerization and DNA-binding surfaces within a single structural unit. As shown in Figure 19-7, two long  $\alpha$  helices form a pincer-like structure that grips the DNA, with each  $\alpha$  helix inserting into the major groove half a turn apart. Dimerization is mediated by another region within those same  $\alpha$  helices: in this region, they form a short stretch of coiled-coil, wherein the two helices are held together by hydrophobic interactions between appropriately spaced leucine (or other hydrophobic) residues. We discussed this protein–protein interaction in more detail in Chapter 6 (Fig. 6-9). Leucine zipper-containing proteins often form heterodimers as well as homodimers. This is also true of our next category, the so-called helix-loop-helix proteins (HLH proteins).

**Helix-Loop-Helix Proteins** As in the example of the leucine zipper, an extended  $\alpha$ -helical region from each of two monomers inserts into the major groove of the DNA. As shown in Figure 19-8, the dimerization surface is formed from two helical regions: the first is part of the same helix involved in DNA recognition; the other is a shorter  $\alpha$  helix. These two helices are separated by a flexible loop that allows them to pack together (and gives the motif its name). Leucine zipper and HLH proteins are often called **basic zipper** and **basic HLH proteins**: this is because the region of the  $\alpha$  helix that binds DNA contains basic amino acid residues.

**HMG Proteins** HMG proteins are unusual in that they interact with the minor groove of the DNA helix, using highly conserved peptide motifs called **AT hooks**. Unlike the other DNA binding proteins discussed above, HMG proteins often dramatically alter the conformation of the DNA helix. In this way, HMG proteins often facilitate the formation of higher-order protein–DNA complexes, as seen for the  $\beta$ -interferon enhanceosome (which we describe later in this chapter [Fig. 19-18]). They also play important roles in development. For example, the HMG-containing Sox2 regulator is essential for the pluripotency of embryonic stem cells (see Chapter 21).

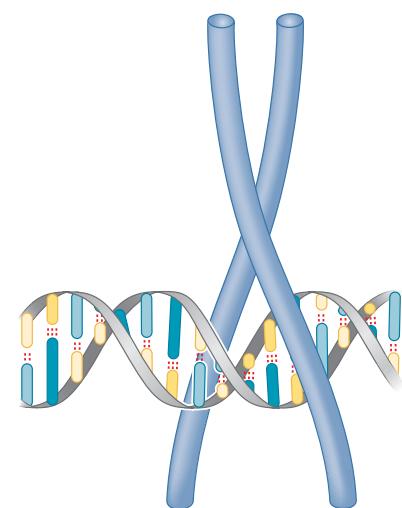
We encountered earlier two other examples of proteins that recognize their DNA target sites via the minor groove and relay of the flexibility of the DNA sequence for specificity. These were the LEF-1 enhancer binding protein (Fig. 6-16) and the general transcription factor TBP (Fig. 13-17).

### Activating Regions Are Not Well-Defined Structures

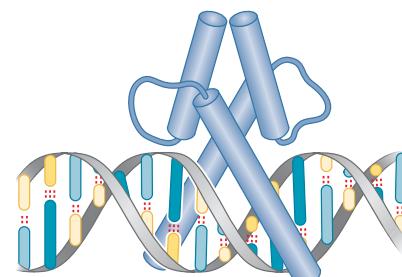
In contrast to DNA-binding domains, activating regions do not always have well-defined structures. They have been shown to form helical structures when interacting with their targets within the transcriptional machinery, but it is believed that these structures are “induced” by that binding. As we shall see, the lack of defined structure is consistent with the idea that activating regions are adhesive surfaces capable of interacting with several other protein surfaces.

Instead of being characterized by structure, therefore, activating regions are grouped on the basis of amino acid content. The activating region of Gal4, for example, is called an “acidic” activating region, reflecting a preponderance of acidic amino acids. The importance of these acidic residues was indicated by mutations that increase the activator’s potency: such mutations increased the overall acidity (negative charge) of the activating region. But despite this, the activating region contains equally critical hydrophobic residues. Many other activators have activating regions like Gal4. Although these show little sequence similarity, they retain the characteristic pattern of acidic and hydrophobic residues.

It is believed that activating regions consist of reiterated small units, each of which has a weak activating capacity on its own. Each unit is a short sequence of amino acids. The greater the number of units, the stronger is the resulting activating region. Again, this is consistent with the idea that activating regions lack an overall structure and act simply as rather indiscriminate “sticky” surfaces. (To understand this reasoning, imagine instead that an activating region folded into a precise, stable three-dimensional structure, comparable to, for example, a DNA-binding domain. Under those circumstances, fragments of that domain would not be expected to retain a fraction of the DNA-binding activity of the intact domain; rather, the entire domain is needed for any significant activity. But if each activating region is simply a general adhesive surface, it is easy to imagine it being made up of



**FIGURE 19-7** Leucine zipper bound to DNA. Two large  $\alpha$  helices, one from each monomer, form both the dimerization and DNA-binding domain at different sections along their length. Thus, as shown, toward the top, the two helices interact to form a coiled-coil that holds the monomers together; further down, the helices separate enough to embrace the DNA, inserting into the major groove on opposite sides of the DNA helix. Once again, specificity is provided by contacts made between amino acid side chains on the  $\alpha$  helices and the edge of base pairs in the major groove. An example of this is found in the yeast transcriptional activator, GCN4. As described in Chapter 6, the helical regions that interact with DNA are disordered until they bind DNA (see Fig. 6-9). (Adapted, with permission, from Ellenberger T.G. et al. 1992. *Cell* 71: 1223. © Elsevier.)

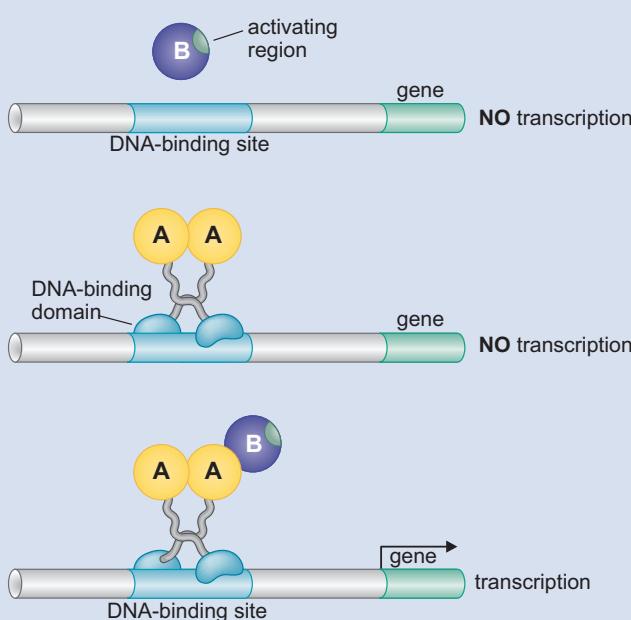


**FIGURE 19-8** Helix-loop-helix motif. In this case, we again see a long  $\alpha$  helix involved in both DNA recognition and, in combination with a second, shorter,  $\alpha$ -helix, dimerization. (Adapted, with permission, from Ma P.C. et al. 1994. *Cell* 77: 451, Fig. 2A. © Elsevier.)

## TECHNIQUES

### Box 19-1 The Two-Hybrid Assay

This assay is used to identify proteins that interact with each other. Thus, in the case shown in Box 19-1 Figure 1, activation of a reporter gene depends on the fact that protein A interacts with protein B (even though these proteins need not themselves normally have a role in transcriptional activation). The assay is predicated on the finding, discussed in the text, that the DNA-binding domain and activating region can be on separate proteins, as long as those proteins interact, and the activating region is thereby tethered to the DNA near the gene to be activated. Practically, the assay is performed as follows. The gene



encoding protein A is fused to a DNA fragment encoding the DNA-binding domain of Gal4. The gene for a second protein (B) is fused to a fragment encoding an activating region. Neither protein alone, when expressed in a yeast cell, activates the reporter gene carrying Gal4-binding sites (as shown in the first two lines of the figure). When both hybrid genes are expressed together in a yeast cell, however, the interaction between proteins A and B generates a complete activator, and the reporter is expressed, as shown in the bottom line of the figure. In a widely used elaboration of this simple assay, the two-hybrid assay is used to screen a library of candidates to find any protein that will interact with a known starting protein. So now, protein A in the figure would be the starting protein (called the **bait**), whereas protein B (the **prey**) represents one of many alternatives encoded by the library (for a description of how libraries are made, see Chapter 7). Yeast cells are transfected with the construct encoding protein A fused to the DNA-binding domain, together with the library encoding many unknown proteins fused to the activating region. Thus, each transfected yeast cell contains protein A tethered to DNA and one or another alternative protein B fused to an activating region. Any cell containing a combination of A and B that interacts will activate the reporter gene. Such a cell will form a colony that can be identified by plating on suitable indicator medium. Typically, the reporter gene would be *lacZ*, and positive colonies (those comprising cells expressing the reporter gene) would be blue on appropriate indicator plates.

**BOX 19-1 FIGURE 1** How the two-hybrid assay works. The reporter gene used in such an assay would typically be *lacZ* or some other gene that makes an easily assayed product.

smaller, weaker units. Activating regions and their targets do not interact in a “lock and key” manner.)

Recently, a series of NMR structural studies has examined the nature of the interaction between an acidic activating region of the yeast activator, Gcn4, and one of its targets in the Gal11 protein (a subunit of Mediator; see Chapter 13). The activating region forms a helical structure upon binding a cleft in the target. But this structure is dynamic and, indeed, can occur in several different conformations and orientations, with only a single hydrophobic interaction between a residue in the activating region and the target being essential and being found in all cases. This kind of “fuzzy complex” may explain why activating regions seem able to interact with several different target proteins when activating transcription, as we discuss in the next section.

There are other kinds of activating regions. These include glutamine-rich activating regions such as that found on the mammalian activator SP1. In addition, proline-rich activating regions have been described—for example, on another mammalian activator CTF1. These, too, lack defined structure. In general, whereas acidic activating regions are typically strong and work in

any eukaryotic organism in which they have been tested, other activating regions are weaker and work less universally than members of the acidic class.

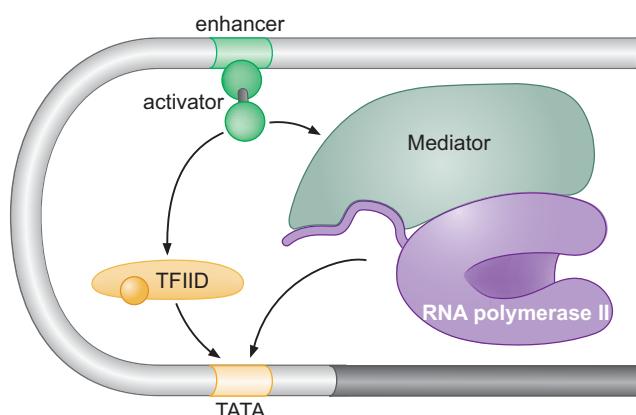
## RECRUITMENT OF PROTEIN COMPLEXES TO GENES BY EUKARYOTIC ACTIVATORS

We saw in bacteria that in the most common case, an activator stimulates transcription of a gene by binding to DNA with one surface, and with another, interacting with RNA polymerase and recruiting the enzyme to that gene (see Chapter 18, Fig. 18-1). Eukaryotic activators also work this way, but rarely, if ever, through a direct interaction between the activator and RNA polymerase. Instead, the activator recruits polymerase indirectly or recruits other factors needed after polymerase has bound. Thus, the activator can interact with parts of the transcriptional machinery other than polymerase and, by recruiting them, recruit polymerase as well. In addition, activators can recruit nucleosome modifiers that alter chromatin in the vicinity of a promoter and thereby help initiation. Finally, activators can recruit factors needed for polymerase to initiate or elongate transcription. In all of these functions, the activator is merely recruiting proteins to the promoter. In bacteria, RNA polymerase is the only protein that needs to be recruited; this is not the case in eukaryotes. Indeed, in eukaryotes, a given activator might work in all three ways: recruitment of nucleosome modifiers and remodelers to “open” the promoter, recruitment of general transcription factor complexes (e.g., Mediator), and recruitment of protein complexes that stimulate Pol II initiation and elongation (e.g., pTEFb/SEC complex). We first consider recruitment of the transcriptional machinery.

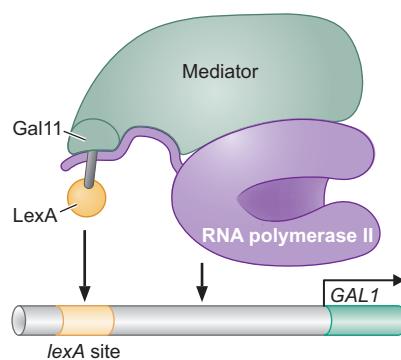
### Activators Recruit the Transcriptional Machinery to the Gene

The eukaryotic transcriptional machinery contains numerous proteins in addition to RNA polymerase, as seen in Chapter 13. Many of these proteins come in preformed complexes such as the **Mediator** and the **TFIID complex** (see Chapter 13, Table 13-2 and Fig. 13-20). Activators interact with one or more of these complexes and recruit them to the gene (Fig. 19-9). Other components that are not directly recruited by the activator bind cooperatively with those that are recruited.

Many proteins in the transcriptional machinery have been shown to bind to activating regions in vitro. For example, a typical acidic activating region can interact with components of the Mediator and with subunits of TFIID.



**FIGURE 19-9** Activation of transcription initiation in eukaryotes by recruitment of the transcription machinery. A single activator is shown recruiting two possible target complexes: the Mediator and, through that, RNA polymerase II, as well as the general transcription factor TFIID. Other general transcription factors are either recruited as part of the Mediator, Pol II, or TFIID complexes or recruited separately by the activator, or they can bind spontaneously in the presence of the recruited components. These are not shown here. In reality, this recruitment would usually be mediated by more than one activator bound upstream of the gene.



**FIGURE 19-10** Activation of transcription through direct tethering of Mediator to DNA. This is an example of an activator bypass experiment, as described in Chapter 18, Box 18-1. In this case, the *GAL1* gene is activated, in the absence of its usual activator Gal4, by the fusion of the DNA-binding domain of LexA to a component of the Mediator complex (Gal11/Med15) (see Chapter 13, Fig. 13-20). Activation depends on LexA DNA-binding sites being inserted upstream of the gene. Other components required for transcription initiation (TFIID, etc.) presumably bind together with Mediator and Pol II.

Recruitment can be visualized using the technique called **chromatin immunoprecipitation (ChIP)**, described in Chapter 7. This technique reveals when a given protein binds to a defined region of DNA within a cell. Elaborations of this method, called ChIP-chip and, now most commonly used, ChIP-Seq, are described in Box 19-2. At most genes (although not all, as we shall see presently), the transcriptional machinery appears at the promoter only upon activation of the gene. That is, the machinery is not prebound, confirming that the role of the activator is to recruit it.

In bacteria, we saw that genes activated by recruitment (such as the *lac* genes) can be activated in so-called activator bypass experiments (Chapter 18, Box 18-1). In such an experiment, activation is observed when RNA polymerase is recruited to the promoter without using the natural activator–polymerase interaction. Similar experiments work in yeast. Thus, the *GAL1* gene (normally activated by Gal4) can be activated equally well by a fusion protein containing the DNA-binding domain of the bacterial protein LexA fused directly to a component of the Mediator complex (Fig. 19-10).

It is important to note that these experiments do not exclude the possibility that at least some activators not only recruit parts of the transcriptional machinery but also induce allosteric changes in them. Such changes might stimulate the efficiency of transcription initiation. Nevertheless, the recruitment of the machinery to one or another gene is the basis of specificity; that is, which gene is activated depends on which gene has the machinery recruited to it. In addition, the success of the activator bypass experiments suggests that any allosteric events that occur during initiation do not require the activator to do anything beyond recruiting proteins to the gene.

## ► TECHNIQUES

### Box 19-2 The ChIP-Chip and ChIP-Seq Assays Are the Best Method for Identifying Enhancers

We described the chromatin immunoprecipitation (ChIP) method in Chapter 7. This method allows an experimenter to identify to what specific DNA sequences a given protein is bound within the genome of a cell—and, indeed, with what other proteins it interacts as well. As we described, in the ChIP procedure, cells, tissues, organs, or even whole embryos are treated with formaldehyde to cross-link DNA-binding proteins to their associated DNA sequences and to other associated proteins. The cross-linked chromatin is sheared to small fragments of ~200 bp. An antibody against the DNA-binding protein of interest is used to isolate the DNA fragments bound by the protein. In conventional ChIP, the cross-linking is reversed and immunoprecipitated DNA is used as a template for amplification by the polymerase chain reaction (PCR) with oligonucleotide primers corresponding to particular genes of interest. Thus, the presence or absence of an amplified sequence reveals whether or not the protein of interest was bound to that DNA sequence in the cells from which the formaldehyde-treated chromatin had been isolated.

The ChIP-chip and ChIP-Seq derivatives of this method make it much more powerful. Rather than just testing to see if the protein is bound to specific previously identified sites, these methods make it possible to detect every site to which the protein of interest is bound in the genome, even if the locations and sequences of those sites were not previously known.

In the ChIP-chip procedure, after the reversal of the cross-linking, all of the immunoprecipitated DNA fragments are amplified by a PCR procedure in which a generic primer is appended nonspecifically to the ends of all of the DNA fragments. After amplification, the DNA molecules are fluorescently labeled and then, in a critical final step, hybridized to a whole-genome array—individual fragments of the genome that together represent coverage of the whole genome. This allows the identification of where the fragments of DNA formerly bound to the protein of interest map in the whole genome of the cell.

ChIP-chip has been used to identify the genome-wide distribution of many interesting regulatory proteins. An example of the power of this technique is its application to the sequence-specific transcription factors Nanog, Sox2, and Oct4. These regulatory proteins are partly responsible for the distinctive properties of human embryonic stem cells, such as their capacity for self-renewal and for generating diverse types of specialized cells. They are also sufficient, when expressed in adult differentiated cells, to induce in those cells stem-cell-like properties that make them pluripotent—so-called induced pluripotent stem cells (iPS cells) (for more details, see Box 21-1).

Antibodies directed against Nanog, Sox2, and Oct4 have been used for the comprehensive identification of the *in vivo* binding sites for these proteins in stem cells (Box 19-2 Fig. 1). More than 100 potential target enhancers have been identified that

**Box 19-2 (Continued)**

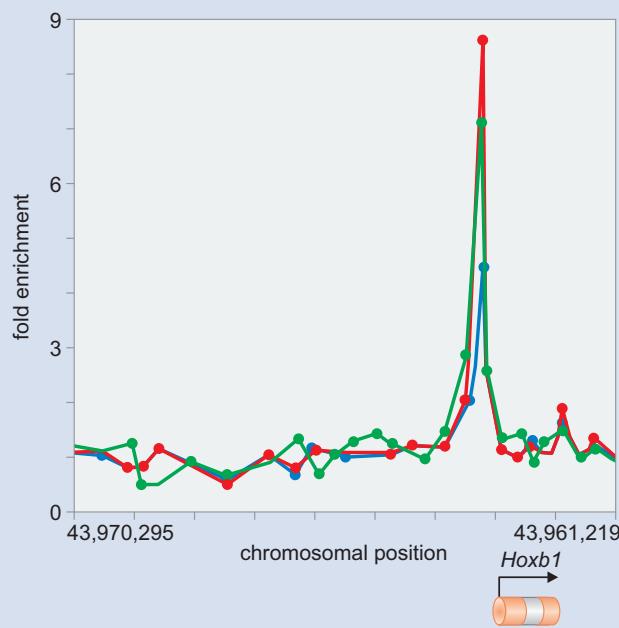
are jointly regulated by all three proteins. Some of these enhancers are associated with genes that are known to be important

regulators of development, such as *Hoxb1*, which is related to the homeotic genes of *Drosophila* (see Chapter 21).

The newest version of this technique—called ChIP-Seq—is even more powerful and simpler to perform. Once the fragments of DNA bound by the protein of choice are liberated by reversing their cross-linking, they are identified by direct sequencing using next-generation sequencing methods (see Chapter 7). In this way, the exact sequence and abundance of each target sequence can readily be detected and measured.

Another extension of these methods is seen in chromosome conformation capture—or 3C (which we described in Chapter 7). This is used to identify when regulatory proteins bound at an enhancer are in close proximity to the transcriptional machinery at a given promoter—a physical proximity interpreted to show looping out of the DNA between the former and the latter.

In 3C, the same basic procedure is followed as we have already described, but in this case, the proteins bound to the enhancer are cross-linked not only to the DNA but also to any other proteins with which they interact. If this includes proteins bound to other DNA sites (e.g., the promoter), then these DNA sites will also be precipitated by the antibody against the original protein of interest. Before reversing the cross-linking, and thus while the various protein-bound DNA fragments are held close together, a ligation reaction is performed, joining the ends of DNA fragments bound by the original protein and those bound by any other proteins in complex with it. The presence of specific hybrid molecules can then be identified in various ways and can reveal where fragments from different locations in the genome (i.e., from a particular enhancer and a known promoter) have been brought together.

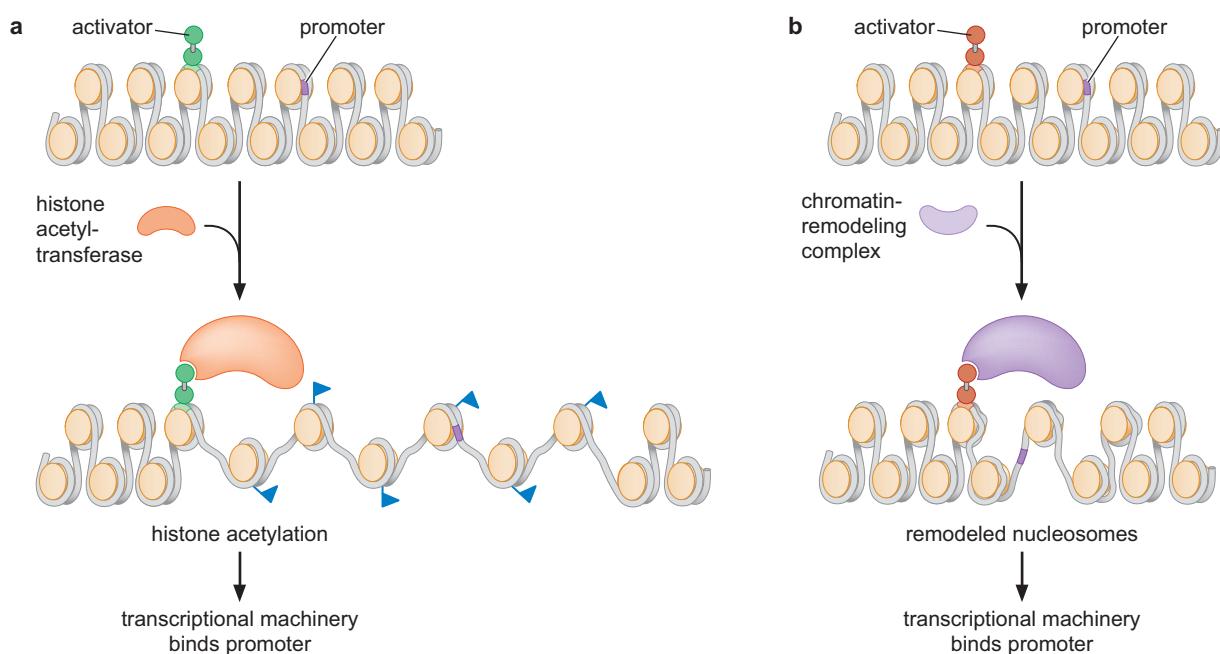


**BOX 19-2 FIGURE 1** ChIP-chip identification of enhancers regulated by stem cell factors. A human whole-genome tiling array was hybridized with DNA fragments associated with Nanog (green), Sox2 (red), and Oct4 (blue). All three proteins bind to a 5'-flanking sequence associated with the *Hoxb1* gene. (Adapted, with permission, from Boyer L.A. et al. 2005. *Cell* **122**: 947–956, Fig. 2b. © Elsevier.)

### Activators Also Recruit Nucleosome Modifiers That Help the Transcriptional Machinery Bind at the Promoter or Initiate Transcription

In addition to direct recruitment of the transcriptional machinery, recruitment of nucleosome modifiers can help activate a gene packaged within **chromatin**. As discussed in Chapter 8, nucleosome modifiers come in two types: those that add chemical groups to the tails of histones, such as **histone acetyltransferases (HATs)**, which add acetyl groups; and those that displace (or “remodel”) the nucleosomes, such as the ATP-dependent activity of **SWI/SNF**. How do these modifications help activate a gene? Two basic models explain how changes in nucleosomes can help the transcriptional machinery bind at the promoter (Fig. 19-11).

First, remodeling, and certain modifications, can uncover DNA-binding sites that would otherwise remain inaccessible within the nucleosome. For example, by removing or increasing the mobility of nucleosomes, remodelers are proposed to free up binding sites for regulators and for the transcriptional machinery. Similarly, the addition of acetyl groups to histone tails alters the interactions between those tails and adjacent nucleosomes. This modification is often said to “loosen” chromatin structure, freeing up sites (for a more complete description, see Chapter 8).



**FIGURE 19-11** Local alterations in chromatin structure directed by activators. Activators capable of binding to their sites on DNA within a nucleosome are shown bound upstream of a promoter that is inaccessible within chromatin. (On the right-hand side) The activator recruits a nucleosome remodeler, which alters the structure of nucleosomes around the promoter, rendering it accessible and capable of binding the transcriptional machinery. (On the left-hand side) The activator is shown recruiting a histone acetylase. That enzyme adds acetyl groups to residues within the histone tails (blue flags). This alters the packing of the nucleosomes somewhat and also creates binding sites for proteins carrying the appropriate recognition domains (bromodomains) (Chapter 8, Figs. 8-41 and 8-42). Together, these effects again allow binding of the transcriptional machinery to the promoter.

But adding acetyl groups also helps binding of the transcriptional machinery (and other proteins) in another way: it creates specific binding sites on nucleosomes for proteins bearing so-called **bromodomains** (Chapter 8, Fig. 8-41). One component of the TFIID complex bears bromodomains and thus binds to acetylated nucleosomes better than to unacetylated nucleosomes. Thus, a gene bearing acetylated nucleosomes at its promoter will likely have a higher affinity for the transcriptional machinery than one with unacetylated nucleosomes. Other proteins contain **chromodomains**: these recognize methylated nucleosomes, examples of which we will encounter later.

Which parts of the transcriptional machinery, and which nucleosome modifiers, are required to transcribe a given gene? Which components are directly recruited by a given activator working at a given gene? The answers to these questions are not known in most cases, but some components of the transcriptional machinery are more stringently required at some genes than at others, and the same applies to nucleosome modifiers as well. These differences are in many cases not absolute. Thus, although all genes absolutely require RNA polymerase itself, a given gene may depend on another particular component of the transcriptional machinery, or a nucleosome modifier, or it may not. In some cases, a component of the transcriptional machinery might help but not be absolutely required (i.e., in the absence of that component, activation is reduced or slowed but not eliminated). In addition, what is needed to activate a given gene can vary depending on circumstances, such as the stage of the cell cycle. For example, Gal4 usually activates the *GAL1* gene efficiently in the absence of a histone acetylase. During mitosis,

however, when chromatin is more condensed (Chapter 8), activation is eliminated unless that acetylase is recruited to the gene.

In yeast, recent experiments have provided good evidence for particular activator–target interactions at specific genes. As we noted above, the acidic activator Gcn4 is known to interact with Gal11 (Med15); it also interacts with the TAF12 subunit of TFIID and other complexes involved in transcription, including the nucleosome remodeler SWI/SNF. And Gal4 appears to contact at least three components: Mediator, TFIID, and a third complex called SAGA (Spt-Ada-Gcn5-acetyltransferase). The last of these complexes harbors acetylation activity (see Chapter 8, Table 8-7) and seems to be capable of interacting with the transcriptional machinery as well. Indeed, SAGA, like TFIID, contains TAFs (TBF-associated factors) (Chapter 13) and is needed in place of TFIID at some promoters. Gal4 also recruits SWI/SNF. The ability of an acidic activator such as Gal4 to work at genes with different requirements (e.g., at those needing TFIID and Mediator, as well as at others needing SAGA and Mediator) can be explained by its ability to interact with multiple targets.

We noted that fusions of a DNA-binding domain to a subunit of Mediator can suffice for activation (Fig. 19-10). But although activation reaches normal levels in such activator bypass experiments, it is slower; that is, it takes longer for expression to reach levels quickly achieved by Gal4. Presumably, activated levels are reached more slowly in the activator bypass experiment because of the need for other components of the machinery to arrive independently in that case, rather than being directly recruited as they are by Gal4.

### Activators Recruit Additional Factors Needed for Efficient Initiation or Elongation at Some Promoters

The elaborate transcriptional machinery of a eukaryotic cell contains numerous proteins required for initiation. It also contains some that aid in elongation (see Chapter 13). At some genes, sequences downstream from the promoter cause pausing or stalling of the polymerase soon after initiation. At those genes, the presence or absence of certain elongation factors greatly influences the level at which the gene is expressed.

One example is the *HSP70* gene from *Drosophila*. This gene, activated by heat shock, is controlled by two activators working together. The GAGA-binding factor is believed to recruit enough of the transcription machinery to the promoter for initiation of transcription. But, in the absence of a second activator, HSF, most of the initiated polymerases stall some 25–50 bp downstream from the promoter. In response to heat shock, HSF binds to specific sites at the promoter and recruits a kinase, P-TEFb (positive transcription elongation factor), to the stalled polymerases. The kinase phosphorylates the carboxy-terminal domain (CTD) of the largest subunit of RNA polymerase (the so-called CTD “tail”; see Chapter 13), freeing the enzyme from the stall and allowing transcription to proceed through the gene. Recent studies suggest that P-TEFb is part of a larger complex, the SEC (super elongation complex), which releases paused Pol II from the proximal promoter. Interestingly, several of the SEC subunits have been implicated in childhood leukemias arising from chromosomal translocations (Box 19-3).

We saw in Chapter 13 that phosphorylation of the CTD tail on Ser5 of the heptad repeat is an important step in the early stages of transcription at all genes, and the kinase TFIIH can perform that phosphorylation. Whether P-TEFb is also needed at most genes is not clear. P-TEFb has been implicated in the phosphorylation of Ser2 of the CTD heptad repeat, and this modification is associated with the release of activated Pol II from promoter se-

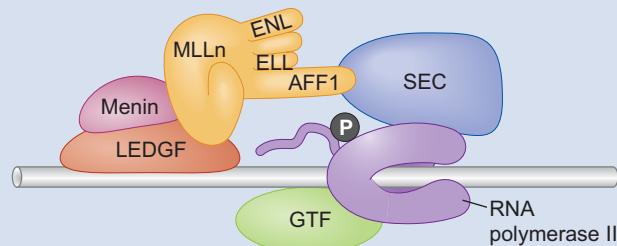
## ► MEDICAL CONNECTIONS

**Box 19-3 Histone Modifications, Transcription Elongation, and Leukemia**

Leukemia is a cancer of the blood arising from an increase in immature white blood cells, or leukocytes. Childhood leukemias are rather common, seen in one in 2000 children under the age of 15. The most common form is called ALL, or acute lymphoblastic leukemia. Approximately three-quarters of infants diagnosed with ALL contain chromosomal rearrangements resulting in the synthesis of MLL fusion proteins. MLL (mixed lineage leukemia) is related to the *Drosophila* Trithorax protein, which is a subunit of a histone-modifying complex called Set1-COMPASS in yeast. These complexes activate promoter regions by methylating lysine 4 of histone H3 (H3K4 trimethylation, or H3K4m3). In *Drosophila*, the Trithorax H3K4 methylation complex counteracts Polycomb repression complexes, which are responsible for trimethylation of H3K27. Thus, MLL, Set1, and Trithorax complexes are positive regulators of gene expression. They impart a “positive” chromatin mark (H3K4me3) in the promoter regions of genes slated for activation.

Remarkably, the majority of MLL fusion proteins resulting in acute lymphoblastic leukemia are fusions between MLL and one or another subunit of the super elongation complex (SEC) described in the text and involved in Pol II elongation. The most common MLL fusion protein is MLL-AFF1, but others include MLL-ENL and other subunits of the SEC. It therefore appears that there is a common underlying molecular mechanism for many different childhood leukemias: the fusion of two major transcription regulatory complexes, MLL (or Set1, Trithorax) and the SEC (see Box 19-3 Fig. 1).

It has been suggested that the functional merging of these two complexes results in “runaway” gene expression in white blood cells. The idea is that MLL complexes normally “mark” certain genes for eventual expression via H3K4me3 in their promoter regions. Such modifications often result in the binding and pausing of Pol II but are not sufficient for the activation of gene expression. However, MLL fusions cause their immediate activation because SEC is now recruited to these genes, where it triggers the release of paused Pol II from the promoter.

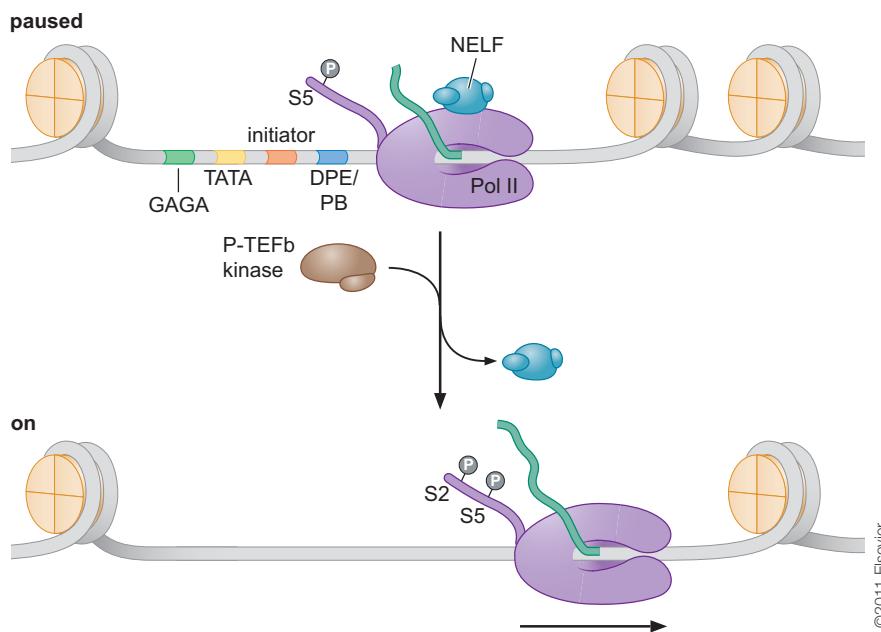


**BOX 19-3 FIGURE 1** Functional merging of two major transcriptional regulatory complexes: MLL chimeras. As illustrated here MLL becomes associated with or fused to any of various proteins in the SEC. Menin, an oncogenic cofactor, tethers MLL and its associated proteins to LEDGF (lens epithelium-derived growth factor), a transcriptional coactivator. GTF, which binds to Pol II, denotes a general transcription factor such as a subunit of the Mediator complex. (Figure kindly provided by Ali Shilatifard.)

quences (Fig. 13-21). A strong acidic activator like Gal4 is able to recruit PTEFb/SEC along with the rest of the machinery. It may be that only at certain genes is the recruitment of the machinery partitioned between regulators in the way we see at this *HSP70* gene, allowing an extra layer of control. For a general picture of paused polymerase and its release, see Fig. 19-12.

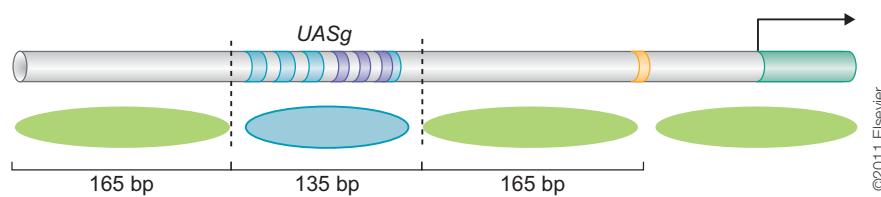
The human immunodeficiency virus (HIV), which causes AIDS, transcribes its genes from a promoter controlled by P-TEFb (and SEC). Again, polymerase initiates transcription at that promoter, under the control of the activator SP1, but stalls soon afterward. In this case, P-TEFb is brought to the stalled polymerase by an RNA-binding protein, not a DNA-bound one. The protein responsible is called TAT. TAT recognizes a specific sequence near the start of the HIV RNA and present in the nascent transcript made by the stalled polymerase. Another domain of TAT interacts with P-TEFb and recruits it to the stalled polymerase. This results in the release of polymerase, the transcription of the viral genome, and infection of the host cell, typically a T-lymphocyte.

It is now believed that paused polymerase is more commonly seen, particularly during development. Thus, recent studies in human embryonic stem cells and the early *Drosophila* embryo suggest that roughly one-third of all protein-coding genes contain paused Pol II before their activation during development. Such genes might be particularly dependent on recruitment of SEC for their expression.



**FIGURE 19-12 Pausing and release of Pol II.** After phosphorylation by TFIIB of Serine 5 in the “tail” (see Fig 13-21), polymerase initiates transcription, but, at some promoters (see text), it then pauses until a second phosphorylation on Serine 2 is achieved through recruitment (by an activator) of the kinase P-TEFb (which is found in the SEC complex, as described in the text). Pausing is mediated by the complex called NELF, which, together with other factors, is shed upon release of the pause. (Redrawn, with permission, from Levine M. 2011. *Cell* 145: 502–511, Fig. 3. © Elsevier.)

It is possible that paused Pol II is also a way of excluding inhibitory nucleosomes from the promoter region, rendering the promoter “poised” for rapid activation by upstream regulatory sequences. Yeast generally lacks paused Pol II but appears to use a different strategy to create poised or “open” promoters. Namely, many yeast promoters contain AT-rich sequences that diminish the formation of nucleosomes. Perhaps paused Pol II and AT-rich promoter sequences accomplish the same goal: to keep promoters relatively free of nucleosomes and therefore poised for rapid induction. As we have already discussed in detail, Gal4 activates the Gal1 gene in yeast, by binding to its sites in the UASg (see Fig. 19-3). There is in fact a second protein (RSC) that binds to sites within this regulatory region (Fig. 19-13). The role of RSC is to nail down and partially unwind a nucleosome over the UASg. This action has two effects: by partially unwinding the DNA from the nucleosome, it ensures the Gal4 binding sites are free for binding by Gal4; and by holding the nucleosome in a set position, it phases the immediately surrounding nucleosomes, ensuring they bind at defined locations. One of these is



**FIGURE 19-13 RSC complex helps Gal4 activate efficiently in the face of nucleosomes.** As we saw in Figure 19-3, Gal4 binds to four sites within the UASg, upstream of the Gal1 gene. Another protein, RSC, also binds sites within the UASg as indicated (in purple). RSC holds a partially unwound nucleosome in such a way that it makes the Gal4 sites available, and it phases the surrounding nucleosomes. These suppress basal expression but are efficiently removed upon activation by SWI/snf recruited by Gal4. The locations of nucleosomes, and the length of DNA they include, is indicated below, in green and blue ovals. (Redrawn, with permission, from Wang X. et al. 2011. *Trends Genet.* 27: 487–492, Fig. 1. © Elsevier.)

positioned over the transcription start site, and upon activation, Gal4 recruits SWI/Snf, as we have already discussed, which efficiently removes this nucleosome (see Fig. 19-11) allowing the rest of the transcription machinery to be recruited to the promoter and transcription to be initiated (Fig. 19-13).

### Action at a Distance: Loops and Insulators

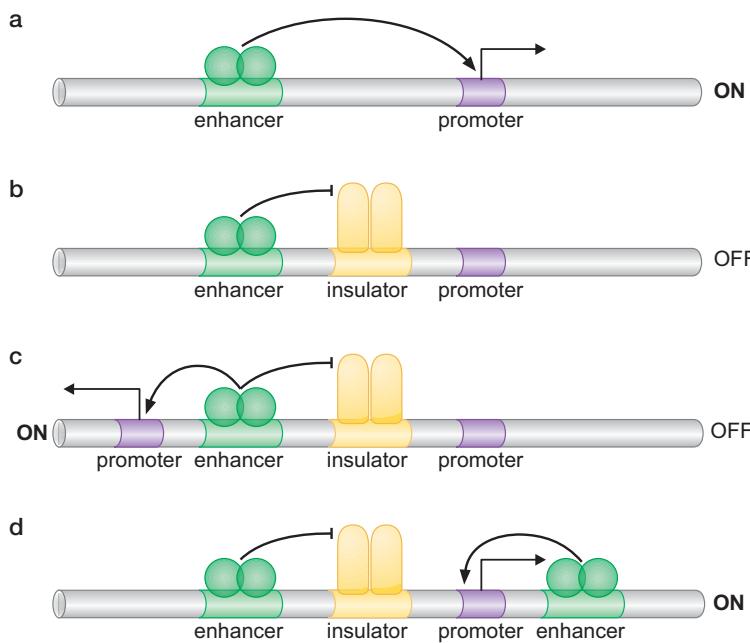
Many eukaryotic activators—particularly in higher eukaryotes—work from a distance. Thus, in a mammalian cell, for example, enhancers can be found several tens or even hundreds of kilobases upstream (or downstream) of the genes they control. We saw in bacteria that proteins bound to separated sites on DNA can interact, a reaction accommodated by DNA looping. But in those cases we were considering proteins binding only a few hundred base pairs apart, sufficiently close to each other that their chance of interacting is much higher *on* DNA than *off* it. Once the sites to which they bind are separated by more than a few kilobases, this advantage is largely lost.

Mechanisms exist to help communication between distantly bound proteins. Recall, from bacteria, one way that this can be done. The “architectural” protein IHF (integration host factor) binds to sites on DNA and bends it. At some genes controlled by NtrC, IHF sites are found between the activator-binding sites and the promoter. By bending the DNA, IHF helps the DNA-bound activator reach RNA polymerase at the promoter (see Chapter 12, Fig. 12-11).

Various models have been proposed to explain how proteins binding between enhancers and promoters might help activation in the cells of higher eukaryotes. In *Drosophila*, the *cut* gene is activated from an enhancer some 100 kb away. A protein called Chip (nothing to do with the technique of that name!) aids communication between enhancer and gene. Thus, mutants in the gene encoding Chip affect the strength of activation. How Chip works is still not clear, but one model is that it binds to multiple DNA sites between the enhancer and the promoter and, by interacting with itself, forms multiple miniloops in the intervening DNA, the cumulative effect of which is to bring the promoter and enhancer into closer proximity. Cohesin, a protein complex engaged in the pairing of homologous chromosomes during cell division (described in Chapter 8; see Fig. 8-14), has also been implicated in stabilizing certain enhancer–promoter loops. Cohesin is found bound to Mediator and, in genome-wide ChIP-seq experiments, is often found associated with proteins bound at enhancers and promoters.

There are other models. In eukaryotes, the DNA is wrapped in nucleosomes as we have seen, and so sites separated by many base pairs may not, in effect, be as far apart in the cell as might have been thought. In addition, chromatin may in some places form special structures that actively bring enhancers and promoters closer together. Enhancers were discovered more than 30 years ago, and yet the basis for long-range enhancer–promoter interactions remains a central mystery of gene regulation.

If an enhancer activates a specific gene 400 kb away, what stops it from activating other genes whose promoters are within that range? On average, because there is a gene every 100–200 kb in a typical vertebrate genome, such an enhancer has to “choose” among two or more genes. Specific elements called **insulators** control the actions of activators. When placed between an enhancer and a promoter, an insulator inhibits activation of the gene by that enhancer. As shown in Figure 19-14, the insulator does not inhibit activation of that same gene by a different enhancer, one placed downstream from the promoter; nor does the insulator inhibit the original activator from working on a different gene. Thus, the proteins that bind insulators do not actively repress the promoter, nor do they inhibit the activities



**FIGURE 19-14** Insulators block activation by enhancers. (a) A promoter activated by activators bound to an enhancer. (b) An insulator is placed between the enhancer and the promoter. When bound by insulator-binding protein CTCF, activation of the promoter by the enhancer is blocked, despite activators binding to the enhancer. (c,d) Neither the activators at the enhancer nor the promoter is inactivated by the action of the insulator. Thus, the activator can activate another promoter nearby (c), and the original promoter can be activated by another enhancer placed downstream (d).

of the activators. Rather, they block communication between the two. Insulators often bind a large zinc-finger protein called CTCF. It is now believed that CTCF also binds cohesin and this complex forms a chromosomal loop with the nearest promoter, thereby precluding enhancers distal to the insulator from forming a similar loop.

In other assays, insulators also inhibit the spread of chromatin modifications. As we have seen, the modification state of local chromatin influences gene expression. We shall see later that propagation of certain repressing histone modifications over stretches of chromatin lies at the heart of a phenomenon called **transcriptional silencing**. Silencing is a specialized form of repression that can spread along chromatin, switching off multiple genes without the need for each to bear binding sites for specific repressors. Insulator elements can block this spreading, and thus insulators protect genes from both indiscriminate activation and repression.

This situation has consequences for some experimental manipulations. A gene inserted at random into the mammalian genome is often “silenced” because it becomes incorporated into a particularly dense form of chromatin called **heterochromatin**. But if insulators are placed upstream and downstream from that gene, they protect it from silencing.

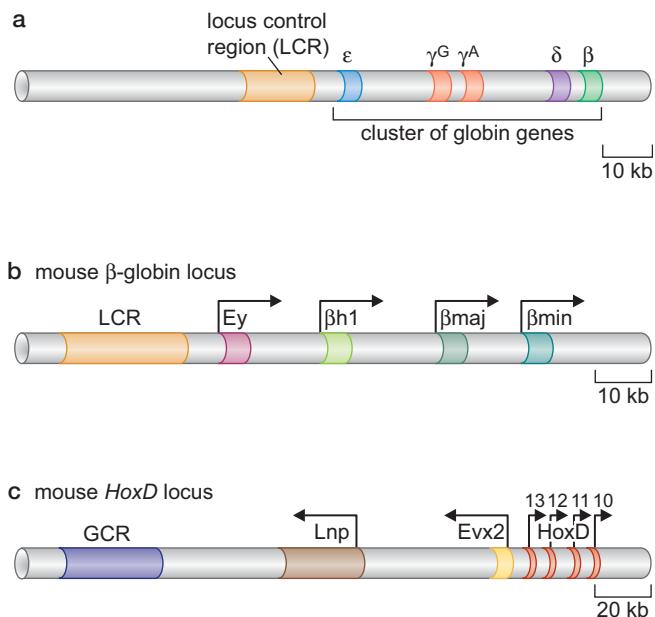
### Appropriate Regulation of Some Groups of Genes Requires Locus Control Regions

The human globin genes are expressed in red blood cells of adults and in various cells in the lineage that forms red blood cells during development. There are five different globin genes in humans (Fig. 19-15a). Although clustered, these genes are not all expressed at the same time; rather, the different genes are expressed at different stages of development starting with  $\epsilon$  (in the fetus) then the  $\gamma$  genes, followed by  $\delta$  and culminating with the expression of  $\beta$ -globin after birth. How is their expression regulated?

Each gene has its own collection of regulatory sites needed to switch that gene on at the right time during development and in the proper tissues (i.e., differentiating blood cells). Thus, the  $\beta$ -globin gene (which is expressed in adult bone marrow) has two enhancers: one upstream of the promoter and

**FIGURE 19-15** Regulation by LCRs.

(a) The human globin genes and the LCR that ensures their ordered expression. Not shown is the  $\alpha$ -globin gene, which is expressed throughout development; its product combines with each of the globins shown here, in turn, to produce different forms of hemoglobins at different stages of development. (b) The globin genes from mice, which are also regulated by an LCR. (c) The *HoxD* gene cluster from the mouse controlled by an element called the GCR, which like the LCRs, appears to impose ordered expression on the gene cluster.



the other downstream. Only in adult bone marrow are the correct regulators all active and present in appropriate concentrations to bind these enhancers. But more than this is required to switch on the various globin genes in the correct order.

A group of regulatory elements collectively called the **locus control region**, or **LCR**, is found 30–50 kb upstream of the whole cluster of globin genes. A similar situation is seen with the *Hoxd* gene cluster in mice. These genes are involved in patterning the developing limbs and are expressed in a precise manner in the embryo (Chapter 21). The *Hoxd* genes are controlled by an element called the **GCR** (global control region) that works like the LCR.

The LCR is made up of multiple sequence elements. Some of these have the properties of enhancers: that is, if these sequences are attached experimentally upstream of a reporter gene, they can activate that gene. Other parts of the LCR act more like insulator elements, and still others seem to have properties of promoters. This diversity of elements has led to numerous models for how LCRs might work. One model proposes that the entire transcriptional machinery is recruited to the LCR and from there transcribes all the way through the locus, opening up the chromatin as it goes and freeing up the local control elements in front of each gene. These individual promoters would then produce high-level expression of each gene as required.

Recent experiments have used techniques that allow the locations of the LCR and promoter to be visualized in cells during activation (e.g., chromosome conformation capture; see Box 19-2). These studies have now been performed with several genes controlled by LCRs or LCR-like elements. In all cases, the results show that regulatory proteins bound to the upstream regulatory sequences are found in close proximity to the promoter as that promoter is activated. This is consistent with the idea that proteins bound at the LCR interact with others at the promoter, with the intervening DNA looping out to accommodate the interaction.

Activation by LCRs is associated with substantial chromatin modification. How this is linked to activation remains unclear. It might help “open up” the chromatin around the LCR itself or around the promoter. It might also alter the chromatin between the two in a manner that helps loop formation.

## SIGNAL INTEGRATION AND COMBINATORIAL CONTROL

### Activators Work Synergistically to Integrate Signals

In bacteria, we saw examples of signal integration in gene regulation. Recall, for example, that the *lac* genes of *Escherichia coli* are efficiently expressed only when both lactose is present and glucose is absent. The two signals are communicated to the gene through separate regulators: one an activator and the other a repressor. In multicellular organisms, signal integration is used extensively. In some cases, numerous signals are required to switch a gene on. But just as in bacteria, each signal is transmitted to the gene by a separate regulator; therefore, at many genes, multiple activators must work together to switch the gene on.

When multiple activators work together, they often do so **synergistically**. That is, the effect of, say, two activators working together is greater (usually much greater) than the sum of each of them working alone. Synergy can result from multiple activators recruiting a single component of the transcriptional machinery, multiple activators each recruiting a different component, or multiple activators helping each other bind to their sites near the gene they control. We briefly consider all three strategies before giving examples. Additional examples of transcriptional synergy are described in Chapter 21, Box 21-4.

Two activators can recruit a single complex (e.g., the Mediator) by touching different parts of it. The combined binding energy will have an exponential effect on recruitment (see Chapter 3, Table 3-1). In cases in which the activators recruit different complexes (neither of which would bind efficiently without help), synergy is even easier to picture.

Synergy can also result from activators helping each other bind under conditions in which the binding of one depends on binding of the other. This **cooperativity** can be of the type we encountered in bacteria, whereby the two regulators touch each other when they bind their sites on DNA (e.g., in the case of  $\lambda$  repressor shown in Chapter 18, Fig. 18-26). But it can work in other ways as well: one activator can recruit something that helps the second activator bind. Figure 19-16 illustrates the different ways activators help each other bind DNA, including “classical” cooperative binding, recruitment of a modifier by one activator to help a second bind, and binding of one activator to nucleosomal DNA uncovering the binding site for another.

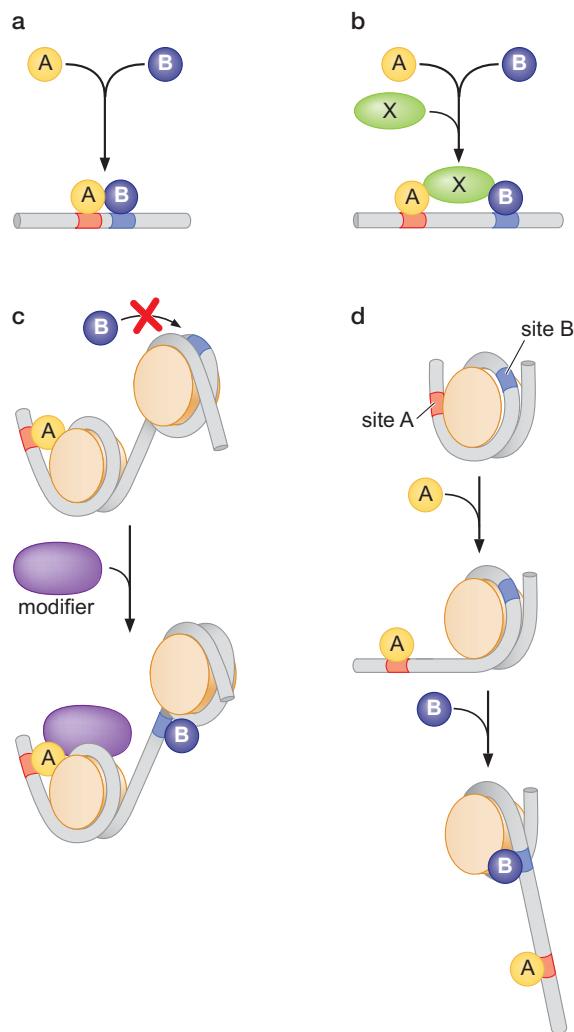
Synergy is critical for signal integration by activators. Consider a gene whose product is only needed when two signals are received, and each signal is communicated to the gene by a separate activator. The gene must be efficiently expressed when both activators are present but be relatively impervious to the action of either activator alone. Thus, the template serves as a matrix for the selective action of multiple signals.

### Signal Integration: The *HO* Gene Is Controlled by Two Regulators—One Recruits Nucleosome Modifiers, and the Other Recruits Mediator

The yeast *Saccharomyces cerevisiae* divides by budding: instead of dividing to produce two identical daughter cells, the so-called mother cell buds to produce a daughter cell. We focus here on the expression of a gene called *HO*. (We need not concern ourselves here with the function of this gene, which is described in Chapter 12.) The *HO* gene is expressed only in mother cells and only at a certain point in the cell cycle ( $G_1$ - $S$  transition) (see Interactive Animation 19-1). These two conditions are communicated to the gene through two activators: SWI5 and SBF.



**FIGURE 19-16 Cooperative binding of activators.** Four ways that the binding of one protein to a site on DNA can help the binding of another to a nearby site. (a) Cooperative binding through direct interaction between the two proteins is shown, as we saw for the  $\lambda$  repressor in Chapter 18 and shall see between many regulators in eukaryotes as well. (b) A similar effect is achieved by both proteins interacting with a common third protein. (c,d) Indirect effects in which binding of one protein to its site on DNA within nucleosomes helps binding of a second protein. (c) The first protein recruits a nucleosome remodeler whose action reveals a binding site for a second protein. (d) The binding of the first protein to its site on the DNA just where it exits the nucleosome. By binding there, it unwinds the DNA from the nucleosome a little, revealing the binding site for the second protein. Each of these mechanisms can explain how one regulator can help others bind or, indeed, how an activator can help the transcriptional machinery bind to a promoter.

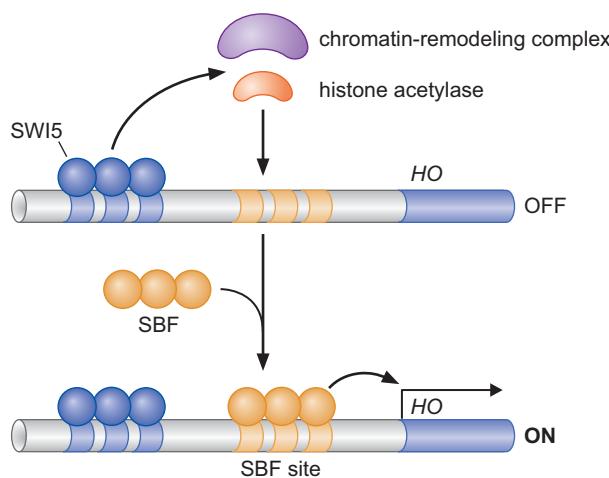


SWI5 binds to multiple sites some distance from the gene, the nearest being  $>1$  kb from the promoter (Fig. 19-17). SBF also binds multiple sites, but these are located closer to the promoter. Why does expression of the gene depend on both activators?

SBF (which is active only during the G<sub>1</sub>-S transition of the cell cycle) cannot bind its sites unaided; their disposition within chromatin prohibits it. SWI5 (which acts only in the mother cell) can bind to its sites unaided but cannot, from that distance, activate the *HO* gene. (Remember that, in yeast, activators typically do not work over long distances.) SWI5 can, however, recruit nucleosome modifiers (a histone acetyltransferase and the remodeling enzyme SWI/SNF). These act on nucleosomes over the SBF sites. Thus, if both activators are present and active, the action of SWI5 enables SBF to bind, and that activator, in turn, recruits the transcriptional machinery (by directly binding Mediator) and activates expression of the gene.

### Signal Integration: Cooperative Binding of Activators at the Human $\beta$ -Interferon Gene

The human  $\beta$ -interferon gene is activated in cells upon viral infection. Infection triggers three activators: NF- $\kappa$ B, IRF, and Jun/ATF. These proteins



**FIGURE 19-17 Control of the *HO* gene.** SWI5 can bind its sites within chromatin unaided, but SBF cannot. Remodelers and histone acetylases recruited by SWI5 alter nucleosomes over the SBF sites, allowing that activator to bind near the promoter and activate the gene. SWI5 is only active in mother cells because a repressor (Ash1), which is only made in daughter cells, binds the *HO* promoter and inhibits action of SWI5. In the figure, for simplicity, the nucleosomes are not drawn. (Adapted, with permission, from Ptashne M. and Gann A. 2002. *Genes and signals*, p. 95, Fig. 2.18. © Cold Spring Harbor Laboratory Press.)

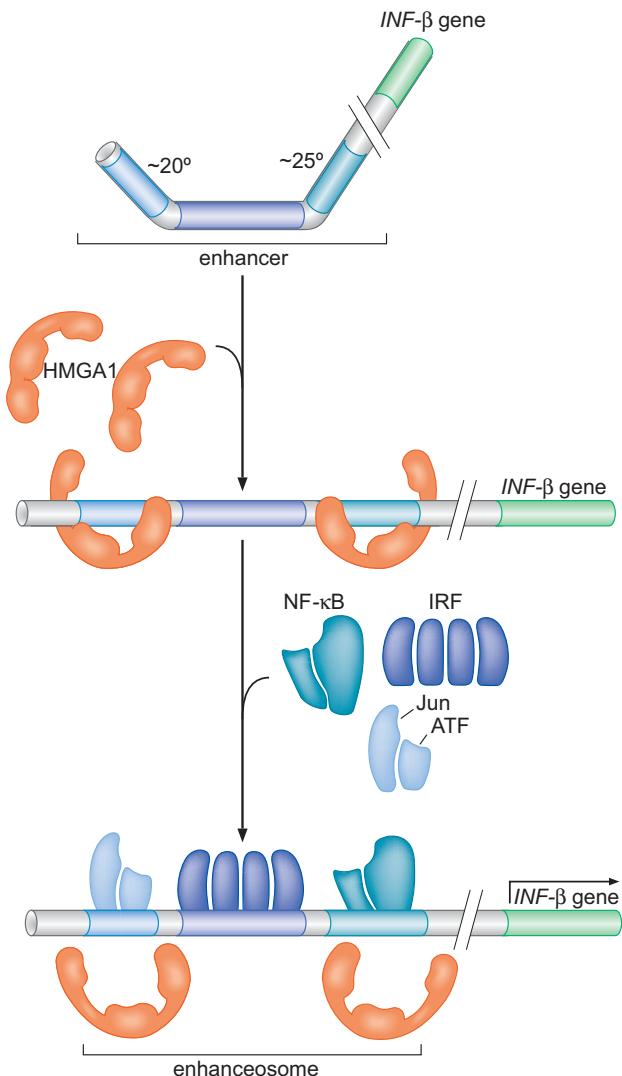
bind cooperatively to sites tightly packed within an enhancer located  $\sim 1$  kb upstream of the promoter. The activators bind the enhancer in a highly cooperative manner to form a structure called an **enhanceosome** (Fig. 19-18). The activators then recruit a so-called coactivator, a protein called CBP (CREB-binding protein) or its close relative p300. This protein has histone-modifying activities and can recruit nucleosome-remodeling activities (e.g., SWI/SNF), as well as the transcriptional machinery itself.

In addition to the activators listed above, another protein binds the enhancer—HMGA1. This protein binds in the minor groove on the opposite face of the DNA and helps in the assembly of the enhanceosome, although it is probably not part of the final structure. Indeed, it seems that it would be impossible for it to remain bound once all of the activators are present; the DNA is simply too crowded. Figure 19-18 shows how the enhanceosome is believed to assemble, and Figure 19-19a shows the crystal structure of the DNA-binding domains of all of the activators bound to the enhancer DNA. As shown in Figure 19-18, the enhancer DNA is bent, but once the activators are bound it is relatively straight; HMGA1 straightens the DNA and thus helps the final structure form.

A striking feature of the structure is that essentially every base pair of DNA within the enhancer is involved in activator binding, which is why there is thought not to be room for HMGA1 in the final structure. This exhaustive use of the sequence information in the enhancer also likely explains why this enhancer is so highly conserved across organisms as diverse as human, mouse, and horse; indeed, it is even more highly conserved than the coding sequence of the gene itself (see Fig. 19-19b).

As we have noted, the activators bind—and the enhanceosome forms—in a highly cooperative manner, ensuring that all three activators must be present. The following are three ways the regulators might be binding cooperatively: (1) through direct protein–protein interactions between them; (2) by changes in the DNA caused by binding of one protein helping binding of another; and (3) by the fact that the activators all interact simultaneously with the coactivator, CBP. All three might operate in this case, although it is hard to know the extent of protein–protein interactions between the activators: not much evidence of direct interactions is apparent in the structure, but then again, only the DNA-binding domain of the activators is in most cases present in the structure.

**FIGURE 19-18** The human  $\beta$ -interferon enhanceosome. Cooperative binding of the three activators, together with the activity of architectural protein HMGA1, activates the  $\beta$ -interferon gene.

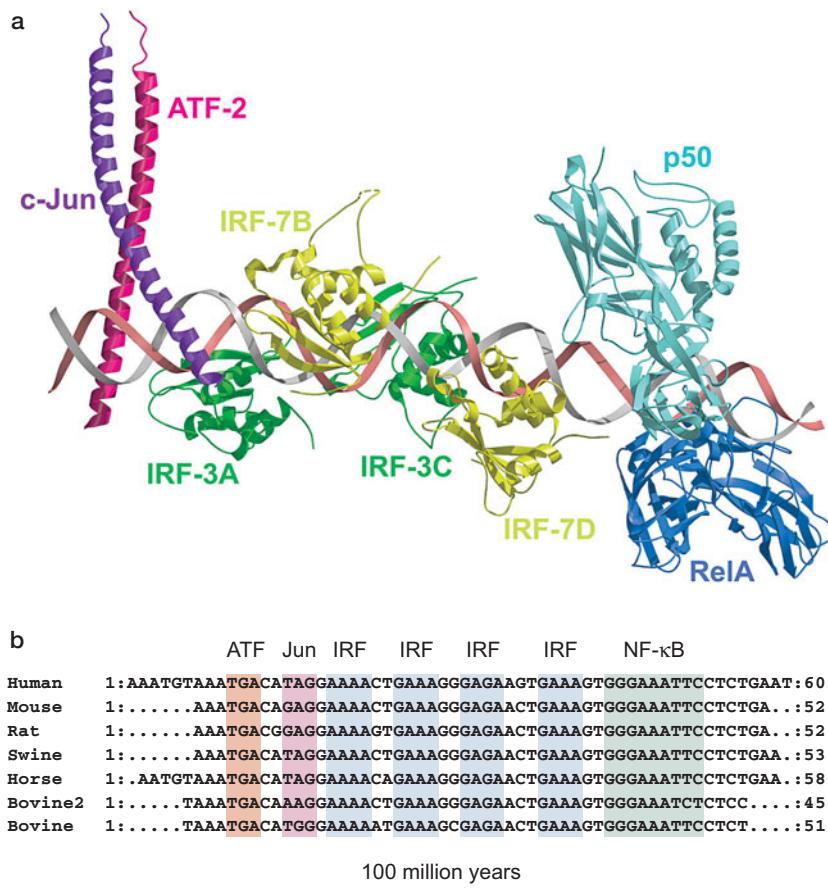


### Combinatorial Control Lies at the Heart of the Complexity and Diversity of Eukaryotes

We encountered simple cases of **combinatorial control** in bacteria. For example, CAP is involved in regulating many genes, in collaboration with other regulators. At the *lac* genes, it works with the Lac repressor; at the *gal* genes, it works with the Gal repressor (see Chapter 18).

There is extensive combinatorial control in eukaryotes. We first consider a generic case (Fig. 19-20). Gene *A* is controlled by four signals (1, 2, 3, and 4), each working through a separate activator (activators 1, 2, 3, and 4). Gene *B* is controlled by three signals (3, 5, and 6), working through activators 3, 5, and 6. Note that there is one signal in common between these two cases, and the activator through which that signal works is the same at both genes. In complex multicellular organisms, such as *Drosophila* and humans, combinatorial control involves many more regulators and genes than shown in this kind of example, and, of course, repressors as well as activators can be involved. How is it that the regulators can intermix so promiscuously?

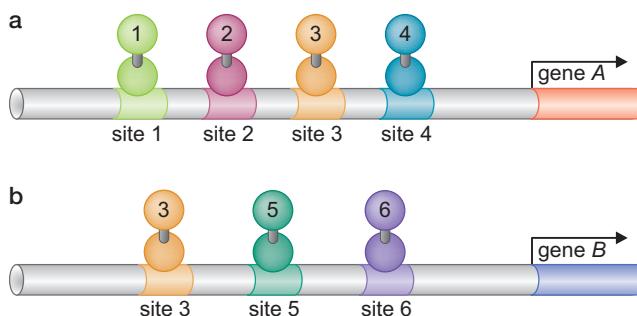
As we discussed above, multiple activators work synergistically. In fact, even multiple copies of a single activator work synergistically, suggesting



**FIGURE 19-19** The enhanceosome structure and sequence. (a) The crystal structure of the enhanceosome, revealing the DNA-binding domains of the activators bound to the enhancer DNA. (Panne D. et al. 2007. *Cell* 129: 1111. PDB Codes: 2O61, 2O6G.) (This image is a combination of structures assembled as a model by Leemor Joshua-Tor. Image prepared with MolScript, BobScript, and Raster3D.) (b) The conservation of the interferon- $\beta$  enhancer DNA sequences across species separated by 100 million years. Also indicated are the sequences within the enhancer recognized by each activator.

that a given activator can interact with multiple targets (just as we saw earlier). This provides an explanation for why different regulators can work together in so many combinations: because each can use any of an array of targets, the combinations that work together are unrestricted.

Both of the examples of signal integration we considered above—the *HO* gene in yeast and the human  $\beta$ -interferon gene—involve activators that also regulate other genes in examples of combinatorial control. Thus, from the yeast example, SWI5 is involved in regulating several other genes. In the mammalian case, NF- $\kappa$ B regulates not only the  $\beta$ -interferon gene, but numerous other genes including the immunoglobulin  $\kappa$  light-chain gene in B cells. Jun/ATF likewise works with other regulators to control other genes. We described earlier that some DNA-binding proteins bind as heterodimers with alternative partners. This offers another level of combinatorial control.



**FIGURE 19-20** Combinatorial control. Two genes are shown, each controlled by multiple signals—four in the case of gene A (a); three in the case of gene B (b). Each signal is communicated to a gene by one regulatory protein. Regulatory protein 3 acts at both genes, in combination with different additional regulators in the two cases.

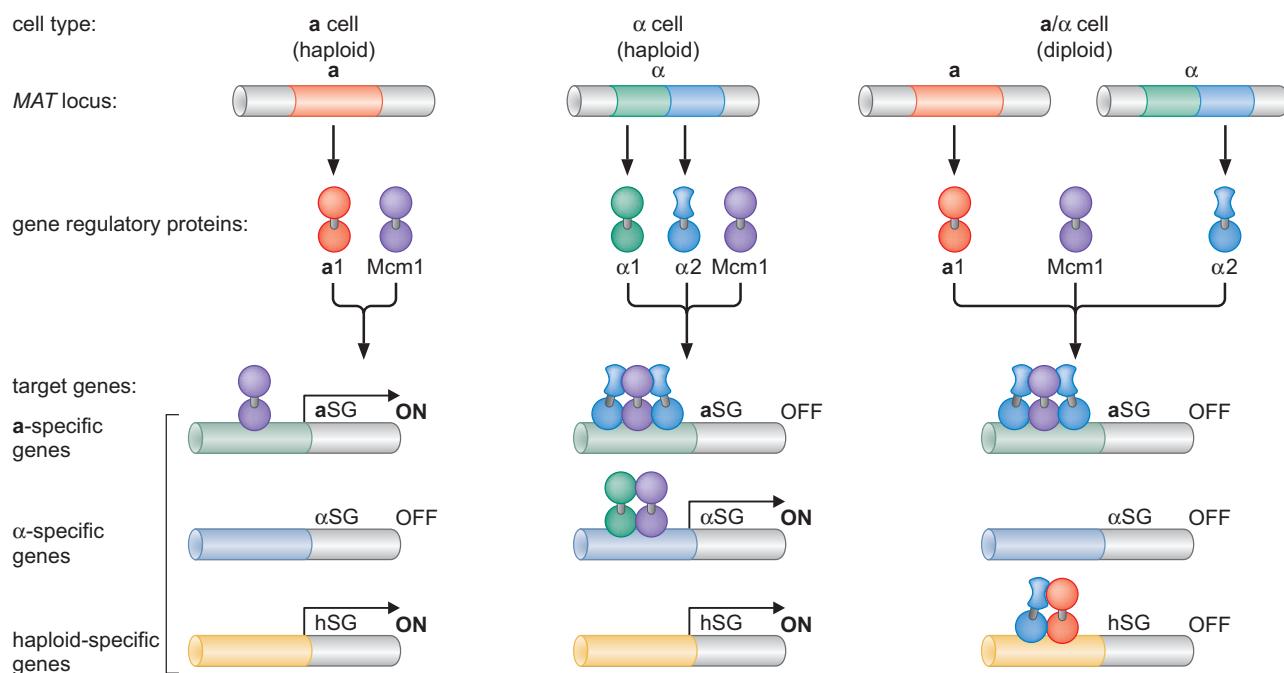
### Combinatorial Control of the Mating-Type Genes from *S. cerevisiae*

The yeast *S. cerevisiae* exists in three forms: two haploid cells of different mating types—**a** and **α**—and the diploid formed when an **a** and an **α** cell mate and fuse. Cells of the two mating types differ because they express different sets of genes: **a**-specific genes and **α**-specific genes. These genes are controlled by activators and repressors in various combinations, as we now briefly describe.

The **a** cell and the **α** cell each encodes cell-type-specific regulators: **a** cells make the regulatory protein **a1**, and **α** cells make the proteins **α1** and **α2**. A fourth regulatory protein, called **Mcm1**, is also involved in regulating the mating-type-specific genes (and many other genes) and is present in both cell types. How do these various regulators work together to ensure that in **a** cells, **a**-specific genes are switched on and **α**-specific genes are off; vice versa in **α** cells; and in diploid cells, both sets are kept off?

The arrangement of regulators at the promoters of **a**-specific genes and **α**-specific genes is shown in Figure 19-21.

- In **a** cells, the **α**-specific genes are off because no activators are bound there, whereas the **a**-specific genes are on because **Mcm1** is bound and activates those genes.
- In **α** cells, the **α**-specific genes are on because **Mcm1** is bound upstream and activates them. At these genes, **Mcm1** binds to a weak site and does so only when it binds cooperatively with a monomer of the protein **α1**. This ensures that **Mcm1** activates these genes only in **α** cells. The **a**-specific genes are kept off in **α** cells by the repressor **α2**. This repressor binds, as a dimer, cooperatively with **Mcm1** at these genes. Two pro-



**FIGURE 19-21** Control of cell-type-specific genes in yeast. As described in detail in the text, the three cell types of the yeast *S. cerevisiae* (the haploid **a** and **α** cells, and the **a/α** diploid) are defined by the sets of genes they express. One ubiquitous regulator (**Mcm1**) and three cell-type-specific regulators (**a1**, **α1**, and **α2**) together regulate three classes of target genes. The **MAT** locus is the region of the genome that encodes the mating-type regulators (Chapter 12).

perties of  $\alpha 2$  ensure that **a**-specific genes are not expressed here: it covers the activating region of Mcm1, preventing that protein from activating; it also actively represses the genes. The mechanism by which  $\alpha 2$  acts as a repressor is described in the next section.

- In diploid cells, both **a**-specific and  $\alpha$ -specific genes are off. This is done as follows: the **a**-specific genes bind Mcm1 and  $\alpha 2$ , just as they do in  $\alpha$  cells. This keeps those genes off. The  $\alpha$ -specific genes are off because, as in **a** cells, no activators bind there.
- Both the haploid cell types (**a** and  $\alpha$ ) express another class of genes called **haploid-specific genes**. These are switched off in the diploid cell by  $\alpha 2$ , which binds upstream of them as a heterodimer with the **a1** protein. Only in diploid cells are both of these regulators present.

The molecular details of mating-type gene regulation are now known for other yeast species. In Box 19-4, Evolution of a Regulatory Circuit, we compare how **a**-specific and  $\alpha$ -specific genes are regulated in *S. cerevisiae* and *Candida albicans*. The comparison reveals how a gene regulatory circuit can evolve, a topic we return to in subsequent chapters.

## TRANSCRIPTIONAL REPRESSORS

---

In bacteria, we saw that many repressors work by binding to sites that overlap the promoter and thus block binding of RNA polymerase. But we also saw other ways they can work: they can bind to sites adjacent to promoters and, by interacting with polymerase bound there, inhibit the enzyme from initiating transcription. They can also interfere with the action of activators.

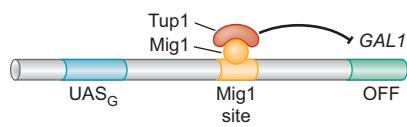
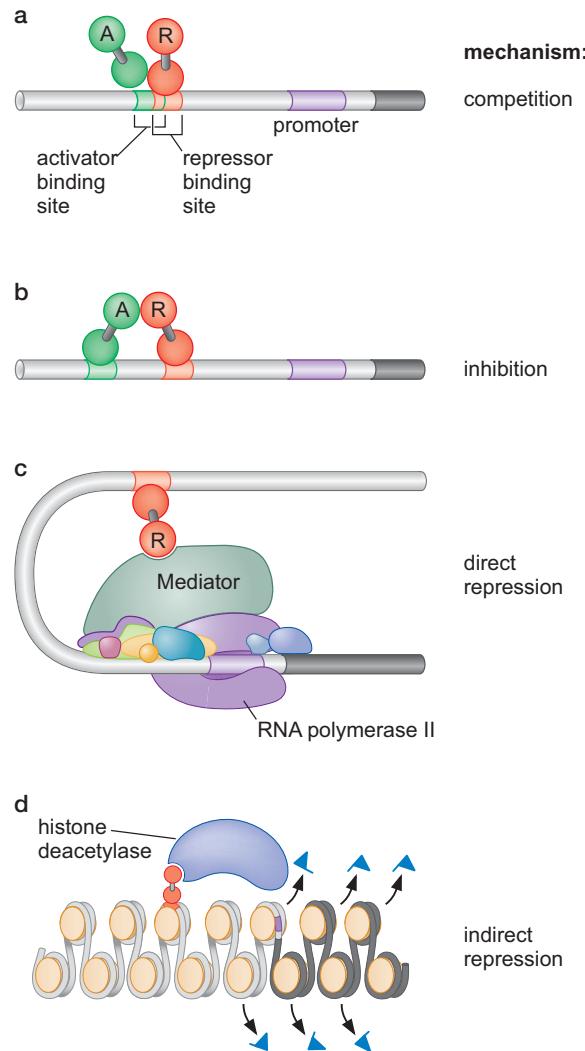
In eukaryotes, we see all of these except the first (ironically, the most common in bacteria). We also see another form of repression, perhaps the most common in eukaryotes, that works as follows. As with activators, repressors can recruit nucleosome modifiers, but in this case, the enzymes have effects opposite to those recruited by activators—they compact the chromatin or remove groups recognized by the transcriptional machinery. Therefore, for example, **histone deacetylases** repress transcription by removing acetyl groups from the tails of histones in *S. cerevisiae*; as we have already seen, the presence of acetyl groups helps transcription. Paradoxically, the histone deacetylase Rpd3 is also recruited to active genes to ensure transcription fidelity. Nucleosomes are deacetylated behind elongating Pol II to prevent the use of “cryptic” promoters within the transcription unit.

Other enzymes add methyl groups to histone tails, and this frequently represses transcription, although in some cases it is associated with an actively transcribed gene (see Chapter 8). Histone (and DNA) modifications also form the basis of a type of repression called **silencing**, which we consider in some detail later in this chapter.

These various examples of repression are shown schematically in Figure 19-22. Here, we consider just one specific example, the repressor called Mig1, which, like Gal4, is involved in controlling the *GAL* genes of the yeast *S. cerevisiae*.

Figure 19-23 shows the *GAL* genes as we saw them above (see Fig. 19-3), but with the addition of a site between the Gal4-binding sites and the promoter: this is where, in the presence of glucose, Mig1 binds and switches off the *GAL* genes. Thus, just as in *E. coli*, the cell only makes the enzymes needed to metabolize galactose if the preferred energy source, glucose, is not present. How does Mig1 repress the *GAL* genes?

**FIGURE 19-22** Ways in which eukaryotic repressors work. Transcription of eukaryotic genes can be repressed in various ways. These include the four mechanisms shown in the figure. (a) By binding to a site on DNA that overlaps the binding site of an activator, a repressor can inhibit binding of the activator to a gene and thus block activation of that gene. In a variation on this theme, a repressor can be a derivative of the same protein as the activator but lack the activating region. In another variation, an activator that binds to DNA as a dimer can be inhibited from doing so by a derivative that retains the region of the protein required for dimerization but lacks the DNA-binding domain. Such a derivative forms inactive heterodimers with the activator. (b) A repressor binds to a site on DNA beside an activator and interacts with that activator, occluding its activating region. (c) A repressor binds to a site upstream of a gene and, by interacting with the transcriptional machinery at the promoter in some specific way, inhibits transcription initiation. (d) Repression is caused by recruiting histone modifiers that alter nucleosomes in ways that inhibit transcription (e.g., deacetylation, as shown here, but also methylation in some cases, or even remodeling at some promoters).



**FIGURE 19-23** Repression of the *GAL1* gene in yeast. In the presence of glucose, Mig1 binds a site between the  $UAS_G$  and the *GAL1* promoter. By recruiting the Tup1 repressing complex, Mig1 represses expression of *GAL1*. Repression is likely the result of deacetylation of local nucleosomes (Tup1 recruits a deacetylase) and also perhaps of directly contacting and inhibiting the transcriptional machinery. In an experiment not shown, if Tup1 is fused to a DNA-binding domain and a site for that domain is placed upstream of a gene, then expression of the gene is repressed.

Mig1 recruits a “repressing complex” containing the Tup1 protein. This complex is recruited by many yeast DNA-binding proteins that repress transcription, including the  $\alpha 2$  protein involved in controlling the mating-type-specific genes described above. Tup1 also has counterparts in mammalian cells. Two mechanisms have been proposed to explain the repressing effect of Tup1. First, Tup1 acts on nucleosomes either through recruiting histone deacetylases and/or by positioning a nucleosome at or near the transcription start site. Second, Tup1 interacts directly with the transcriptional machinery at the promoter and inhibits initiation.

## SIGNAL TRANSDUCTION AND THE CONTROL OF TRANSCRIPTIONAL REGULATORS

### Signals Are Often Communicated to Transcriptional Regulators through Signal Transduction Pathways

As we have seen, whether or not a given gene is expressed very often depends on environmental signals. Signals come in many forms: they can,

## ► KEY EXPERIMENTS

**Box 19-4 Evolution of a Regulatory Circuit**

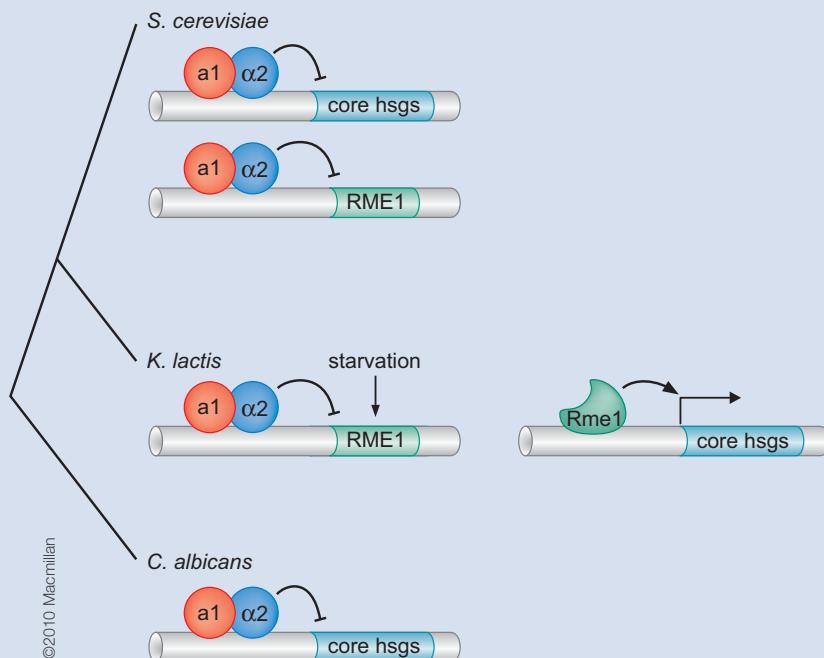
As described in the text (and shown in Fig. 19-21), the different mating types of the yeast *S. cerevisiae* express some of their genes in a cell-type-specific way. Thus, in  $\alpha$  cells, the  $\alpha$ -specific genes are expressed and the  $a$ -specific genes are not, whereas in  $a$  cells, the  $a$ -specific genes are expressed and the  $\alpha$ -specific genes are not. Another class of genes—haploid-specific genes (hsgs)—is expressed in both of these cell types ( $a$  and  $\alpha$ ) but not in the diploid  $a/\alpha$  cell. The products of the hsgs are required for mating—something  $a$  and  $\alpha$  cells do, but that  $a/\alpha$  cells (the product of the mating) cannot. We know in detail how these programs are controlled by regulators encoded by the *MAT* loci working together with the ubiquitous regulator *Mcm1*.

Recent studies provide an illuminating example of how gene regulatory networks evolve. Consider three yeast species—*S. cerevisiae*, *K. lactis*, and *C. albicans*. *C. albicans* and *S. cerevisiae* last shared a common ancestor somewhere between 300 and 900 million years ago. If expressed in terms of the divergence of conserved proteins, these two yeast are more divergent than fish and mammals. *S. cerevisiae* is used in beer and bread making, as well as in laboratory experiments, whereas *C. albicans* is a human pathogen. Nevertheless, just as with *S. cerevisiae*, *C. albicans* comes in two mating types— $a$  and  $\alpha$ —each characterized by the expression of distinct sets of genes ( $a$ -specific genes in  $a$  cells,  $\alpha$ -specific genes in  $\alpha$  cells, and hsgs in both  $a$  and  $\alpha$  cells).

Both species of yeast share a common mechanism for the repression of hsgs in the  $a/\alpha$  diploid cells (Box 19-4 Fig. 1). Diploid cells contain both the  $a1$  and  $\alpha2$  homeodomain regulatory proteins. These proteins form a weak heterodimer by binding to regulatory regions of hsgs, and repress their expression. In *S. cerevisiae*,  $a1/\alpha2$  heterodimers also repress *RME1*, a gene that inhibits entry of diploid cells into meiosis during starvation conditions.

The analysis of mating in a third species (*K. lactis*) reveals an extra layer in the regulation of hsgs by the “intercalation” of *RME1* into that regulatory network (Box 19-4 Fig. 1). In this species, *RME1* is a critical activator of hsgs in  $a$  cells and  $\alpha$  cells. In diploid cells, the  $a1/a2$  heterodimer represses hsgs, just as it does in *S. cerevisiae* and *C. albicans*. But although in those cases the repression is direct, in *K. lactis* the repression is indirect:  $a1/a2$  represses expression of *RME1* and thereby eliminates activation of hsgs.

Network evolution through intercalation of regulators allows new signals to be added to existing pathways. Intercalation of *RME1* into the hsgs repression network adds another tier of regulation by permitting nutritional conditions to impinge on the decision of whether  $a$  and  $\alpha$  cells will mate. These studies also show the importance of examining multiple species within a phylogeny to gain an understanding of how gene regulatory networks evolve.



**BOX 19-4 FIGURE 1** Simplified model for the evolution of regulation of core hsgs in three yeasts. In all three species the core hsgs are repressed by  $a1 - \alpha2$ ; thus, they are ON in  $a$  and  $\alpha$  cells and OFF in  $a/\alpha$  cells. In *S. cerevisiae* and *C. albicans*, the repression is direct ( $a1 - \alpha2$  binds to the promoters of these genes), but in *Kluyveromyces lactis*, repression is indirect, through *Rme1*. The circuit rewiring in the *K. lactis* lineage has resulted in a new mating behavior; this species is able to mate only when starved. (Adapted, with permission, from Booth L. et al. 2010. *Nature* **468**: 959–965. Fig. 4. © Macmillan.)

as was typically the case in bacteria, be small molecules such as sugars, but they can also be proteins released by one cell and received by another. This is particularly common during the development of multicellular organisms (Chapter 21).

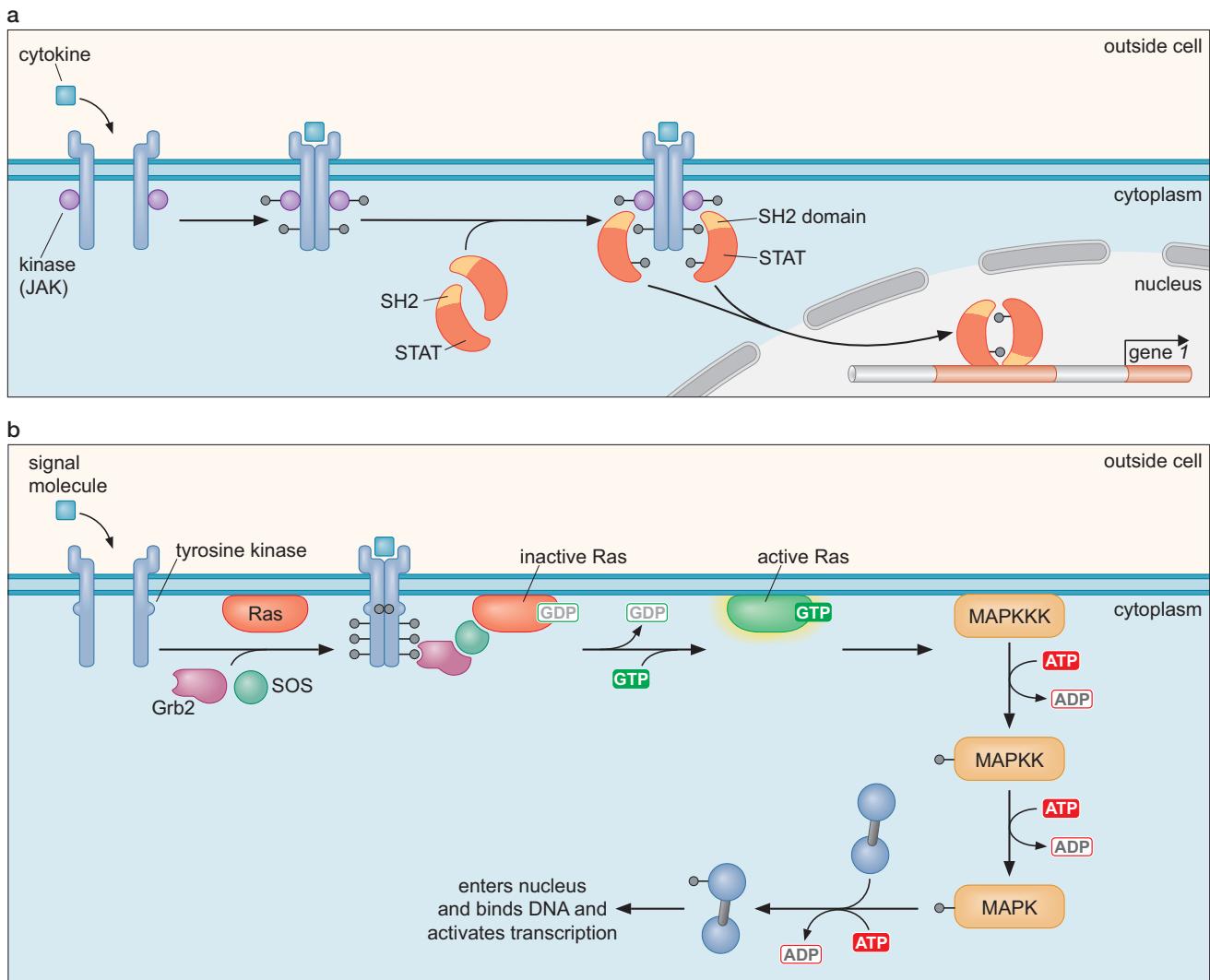
There are various ways that signals are detected by a cell and communicated to a gene. In bacteria, we saw that signals control the activities of regulators by inducing allosteric changes in those regulators. Often, this effect is direct: a small molecular signal, such as a sugar, enters the cell and binds the transcriptional regulator directly. But we saw one example where the effect of the signal is indirect (control of the activator NtrC). In that case, the signal (low ammonia levels) induces a kinase that phosphorylates NtrC. This type of indirect signaling is an example of a **signal transduction pathway**.

The term **signal** refers to the initiating ligand itself—the sugar or protein, for example. This is how we have defined it previously. It can also refer to the “information” as it passes from detection of that ligand to the regulators that directly control the genes—that is, as it passes along a signal transduction pathway. In the simplest of bacterial cases, there was no distinction, of course, but once a signal transduction pathway is involved, there is. In addition, in eukaryotes, we see—particularly in Chapter 21—that most signals are communicated to genes through signal transduction pathways, sometimes very elaborate ones. In this section, we first look at a few signal transduction pathways in eukaryotes. We then consider more generally how signals, emerging from such pathways, control the transcriptional regulators themselves.

In a signal transduction pathway, the initiating ligand is typically detected by a specific **cell surface receptor**: the ligand binds to an extracellular domain of the receptor, and this binding is communicated to the intracellular domain. From there, the signal is relayed to the relevant transcriptional regulator, often through a cascade of kinases. How is the binding of ligand to the extracellular domain communicated to the intracellular domain? This can be through an allosteric change in the receptor, whereby binding of ligand alters the shape (and thus activity) of the intracellular domain. Alternatively, the ligand can act simply to bring together two or more receptor chains, allowing interactions between the intracellular domains of those receptors to activate each other.

Figure 19-24 shows two examples of signal transduction pathways. The first is a relatively simple case, the **STAT** (signal transducer and activator of transcription) pathway (Fig. 19-24a). In this example, a kinase is bound to the intracellular domain of a receptor. When the receptor is activated by its ligand (a cytokine), it brings together two receptor chains and triggers the kinase in each chain to phosphorylate a particular sequence in the intracellular domain of the opposing receptor. This phosphorylated site is then recognized by a particular STAT protein that, once bound, gets phosphorylated itself. Once phosphorylated, the STAT dimerizes, moves to the nucleus, and binds DNA.

The other example is more elaborate (Fig. 19-24b): the mitogen-activated protein kinase (**MAPK**) pathway that controls activators such as Jun, one of the activators that works at the interferon- $\beta$  enhancer we described above (Fig. 19-18). In this case, the activated receptor induces a cascade of signaling events, ending in activation of an MAPK that phosphorylates Jun (and other transcriptional regulators). The most common way in which information is passed through signal transduction pathways is via phosphorylation, but proteolysis, dephosphorylation, and other modifications are also used.



**FIGURE 19-24** Two signal transduction pathways from mammalian cells. Shown are the STAT and Ras pathways. (a) A cytokine is shown binding its receptor, bringing together two receptor chains. Each chain has a kinase called a JAK attached to its intracellular domain. Bringing the chains together (probably accompanied by a conformational change triggered by cytokine binding) leads to phosphorylation of the receptor chains by the JAK kinases (which also phosphorylate each other, stimulating their kinase activity). The sites phosphorylated in the receptor chain are then recognized by cytosolic proteins called STATs. Each STAT has a so-called SH2 domain. These domains are found in many proteins involved in signal transduction. They recognize phosphorylated Tyr residues in certain sequence contexts, and this is the basis of specificity in this pathway. That is, the particular STAT recruited to a given receptor determines which genes will subsequently be activated. Once recruited to the receptor, that STAT itself gets phosphorylated by the JAK kinase. This allows two STAT proteins to form a dimer (the SH2 domain on each STAT recognizing the phosphorylated site on the other). The dimer moves to the nucleus, where it binds specific sites on DNA (different for different STATs) and activates transcription of nearby genes. (b) The Ras pathway leading into the downstream MAPK pathway. A growth factor (such as epidermal growth factor) binds its receptor, bringing together the chains, which, as in the STAT case, then phosphorylate each other. This phosphorylation recruits an adaptor protein called Grb2, which has an SH2 domain that recognizes a phosphorylated tyrosine residue in the activated receptor. The other end of Grb2 binds SOS, a guanine nucleotide exchange factor (Ras GEF). SOS, in turn, binds the Ras protein, which is attached to the inside face of the cell membrane. Ras is a small GTPase, a protein that adopts one conformation when bound to GTP and another when bound to GDP; interaction with SOS triggers Ras to exchange its bound GDP for a GTP and hence undergo a conformational change. In this new conformation, Ras activates a kinase at the top of the so-called MAPK cascade. The first kinase in this pathway is called a MAPK kinase kinase (MAPKKK) (Raf); once activated by Ras, this phosphorylates serine and threonine residues in the next kinase (a MAPK kinase [MAPKK], called Mek). This activates Mek, which, in turn, phosphorylates and activates the MAPK (Erk). This MAPK then phosphorylates several substrates, including transcriptional activators (e.g., Jun) that regulate a number of specific genes, including interferon- $\beta$  (Fig. 19-18).

## Signals Control the Activities of Eukaryotic Transcriptional Regulators in a Variety of Ways

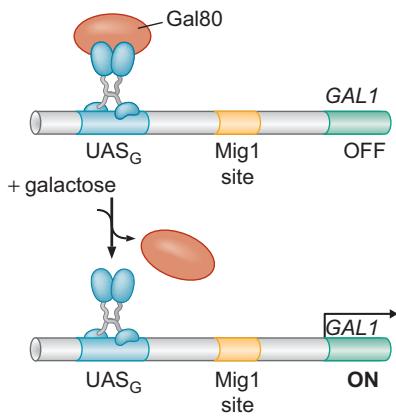
Once a signal has been communicated, directly or indirectly, to a transcriptional regulator, how does it control the activity of that regulator? In bacteria, we saw that the allosteric changes that control transcriptional regulators very often affect the ability of the regulator to bind DNA. This is true in cases in which the signaling ligand itself acts directly on the transcriptional regulator and in cases in which the presence of the signaling ligand is communicated to the regulator through a signal transduction pathway. Thus, Lac repressor binds DNA only when free of allolactose, and phosphorylation of NtrC triggers an allosteric change controlling DNA binding by that activator.

In eukaryotes, transcriptional regulators are not typically controlled at the level of DNA binding (although there are exceptions). Regulators are instead usually controlled in one of the following two basic ways.

**Unmasking an Activating Region** Unmasking an activating region is done either by a conformational change in the DNA-bound activator, revealing a previously buried activating region, or by release of a masking protein that previously interacted with, and eclipsed, an activating region. The conformational changes required in each case can be triggered either by binding ligand directly or through a ligand-dependent phosphorylation.

Gal4 is controlled by a masking protein. In the absence of galactose, Gal4 is bound to its sites upstream of the *GAL1* gene, but it does not activate that gene because another protein, Gal80, binds to Gal4 and occludes its activating region. Galactose triggers the release of Gal80 and activation of the gene (Fig. 19-25).

In many cases, the masking protein not only blocks the activating region but also is itself (or recruits) a deacetylase, and thus actively represses the gene. An example is the mammalian activator E2F, which binds sites upstream of its target genes, whether or not it is activating them. A second protein—the repressor called Rb (retinoblastoma protein)—controls the activity of E2F by binding to it and both blocking activation and recruiting a deacetylase enzyme that represses the target genes. Phosphorylation of Rb causes release of that protein from E2F and thus activation of the genes. E2F controls genes required to take a mammalian cell through the S phase of the cell cycle (Chapter 9). Phosphorylation of Rb thus controls proliferation in these cells. Mutations affecting this pathway are often associated with uncontrolled cell proliferation and cancer.



**FIGURE 19-25** The yeast activator Gal4 is regulated by the Gal80 protein. Gal4 is active only in the presence of galactose. Even in the absence of galactose, however, Gal4 is found bound to its sites upstream of the *GAL1* gene. But it does not under these circumstances activate that gene because the activating region is bound by a protein called Gal80. In the presence of galactose, a protein called Gal3 binds to Gal80, triggers a conformational change, and reveals the activating regions of Gal4. In the figure, Gal80 is shown dissociating from Gal4 in the presence of galactose. It may in reality change its position and weaken its binding but not completely fall off. As shown, Mig1 is not bound at its site because there is no glucose present (see Fig. 19-23).

**Transport into and out of the Nucleus** When not active, many activators and repressors are held in the cytoplasm. The signaling ligand causes them to move to the nucleus, where they act. There are many variations on this theme. Thus, the regulator can be held in the cytoplasm through interaction with an inhibitory protein or with the cell membrane, or it can be in a conformation in which a signal sequence required for its nuclear import is concealed.

Release and transport into the nucleus in response to a signal can be mediated through proteolysis of an inhibitor or tethering region or by allosteric changes. We see an example of this in Chapter 21, when we consider the formation of the dorsoventral axis of the *Drosophila* embryo. There, Cactus is an inhibitory protein that binds the transcriptional regulator Dorsal in the cytoplasm. In response to a specific signal, Cactus is phosphorylated and destroyed, allowing Dorsal to enter the nucleus, bind specific sites within appropriate enhancers, and regulate the transcription of associated genes (Chapter 21, Fig. 21-12). This same mechanism is used to control the activity of NF- $\kappa$ B, one of the regulators of  $\beta$ -interferon, as discussed above. NF- $\kappa$ B is held in the cytoplasm by I $\kappa$ B; NF- $\kappa$ B is related to Dorsal and I $\kappa$ B to Cactus.

## GENE “SILENCING” BY MODIFICATION OF HISTONES AND DNA

We have thus far considered regulation by activators and repressors that bind near a gene and switch it on or off. The effects are local, and the actions of the regulators are often controlled by specific extracellular signals. We now turn to the mechanisms of **transcriptional silencing**. Silencing, in this context (we see the term applied to a rather different situation in Chapter 20), is a position effect: a gene is silenced because of where it is located, not in response to a specific environmental signal. In addition, silencing can “spread” over large stretches of DNA, switching off multiple genes, even those quite distant from the initiating event. Despite these differences, understanding silencing does not require entirely new principles, just extensions of those we have already encountered in this chapter.

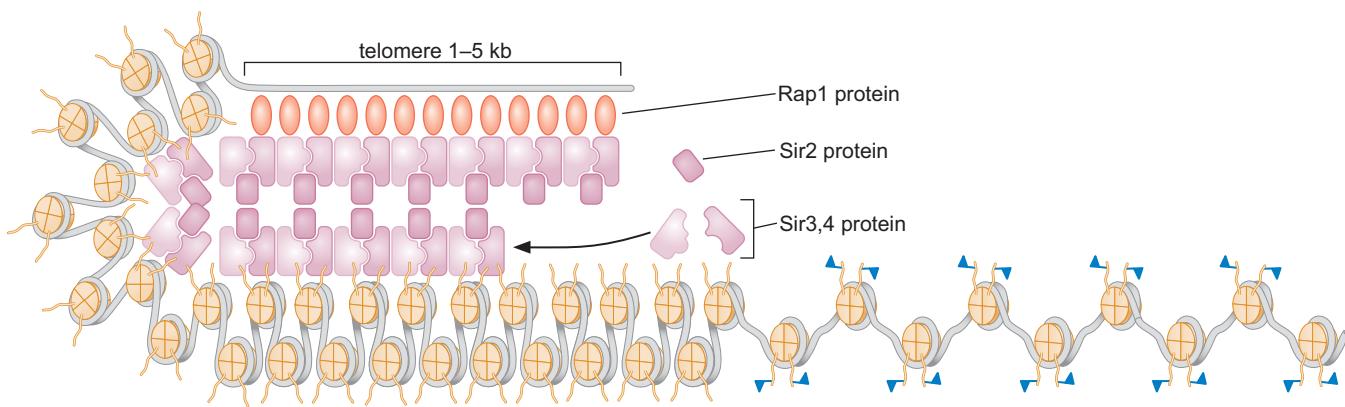
The most common form of silencing is associated with a dense form of chromatin called **heterochromatin**. Heterochromatin was named for its appearance under the light microscope (Fig. 19-26) and represents regions of the chromosome that remain densely packed, even during interphase (Chapter 8). Heterochromatin is frequently associated with particular regions of the chromosome, notably the **telomeres**—the structures found at the ends of chromosomes—and the **centromeres**. As discussed in Chapter 8, telomeres and centromeres are typically composed of repetitive sequences and contain few, if any, protein-coding genes. If a gene is experimentally moved into these regions, that gene is typically switched off. In fact, there are other regions of the chromosome that are also in a heterochromatic state, and in which genes are found, such as in the silent mating-type locus in yeast. And in mammalian cells, ~50% of the genome is estimated to be in some form of heterochromatin. However, there are essential genes located within the heterochromatin of the *Drosophila* genome, including the Rolled MAP kinase. Such genes often require heterochromatin for their normal activities and show erratic behavior when translocated into euchromatin. Nonetheless, for the most part, genes are located in euchromatin and show diminished or erratic activities (variegation) when placed in heterochromatin (as discussed in more detail later).

We have already seen that chromatin can be altered by enzymes that chemically modify the tails of histones or alter the positioning of nucleosomes. Such modifications affect accessibility of the DNA and therefore affect processes such as replication, recombination, and transcription. As we have described, both activation and repression of transcription are often associated with modification of nucleosomes to alter the accessibility of a gene to the transcriptional machinery and other regulatory proteins. We have also encountered proteins that recognize modified nucleosomes and bind specifically to them. Heterochromatic silencing can be understood as an extension of these same principles and mechanism.

Transcription can also be silenced by methylation of DNA by enzymes called **DNA methylases**. This kind of silencing is not found in yeast but is common in mammalian cells. Methylation of DNA sequences can inhibit binding of proteins, including the transcriptional machinery, and thereby block gene expression. But methylation can also inhibit expression in another way: some DNA sequences are recognized only when methylated by specific repressors that then switch off nearby genes, often by recruiting histone modifying enzymes.



**FIGURE 19-26** DAPI staining shows regions of heterochromatin in the *Arabidopsis* genome. DAPI stains DNA, and the brighter staining reflects where the chromatin is more densely packed into heterochromatin. The fainter staining is the euchromatin. (Figure kindly provided by Paul Fransz.)



**FIGURE 19-27 Silencing at the yeast telomere.** Rap1 recruits Sir complex to the telomere. Sir2, a component of that complex, deacetylates nearby nucleosomes. The unacetylated tails themselves then bind Sir3 and Sir4, recruiting more Sir complex, allowing the Sir2 within it to act on nucleosomes further away, and so on. This explains the spreading of the silencing effect produced by deacetylation. (Adapted, with permission, from Grunstein M. et al. 1998. *Cell* **93**: 325–328. © Elsevier.)

### Silencing in Yeast Is Mediated by Deacetylation and Methylation of Histones

The telomeres, the silent mating-type locus, and the rDNA genes are all “silent” regions in *S. cerevisiae*. We consider the telomere as an example.

The final 1–5 kb of each chromosome is found in a folded, dense structure, as shown in Figure 19-27. Genes taken from other chromosomal locations and moved to this region are often silenced, particularly if they are only weakly expressed in their usual location. The chromatin at the telomere is less acetylated than that found in most of the rest of the genome—the so-called **euchromatin**—where genes are more readily expressed.

Mutations have been isolated in which silencing is relieved—that is, in which a gene placed at the telomere is expressed at higher levels. These studies implicate three genes encoding regulators of silencing—*SIR2*, *SIR3*, and *SIR4* (*SIR* stands for silent information regulator). The three proteins encoded by these genes form a complex that associates with silent chromatin, and one of them—Sir2—is a histone deacetylase.

The silencing complex is recruited to the telomere by a DNA-binding protein that recognizes the telomere’s repeated sequences. At the silent mating-type locus, recruitment is also initiated by a specific DNA-binding protein. In both cases, recruitment of Sirs triggers local deacetylation of histone tails. The deacetylated histones are, in turn, recognized directly by the silencing complex, and thus the local deacetylation readily spreads along the chromatin in a self-perpetuating manner, producing an extended region of heterochromatin.

Unlike repression by Tup1, in which the mechanism is still uncertain, here silencing is clearly caused by the deacetylation of histone tails: loss of Sir2 completely alleviates silencing, and acetylation of the histone tail has a similar effect. The entire heterochromatic structure depends on the continued presence of the DNA-binding protein (Rap1) to remain intact. Thus, despite the reinforcing and spreading of deacetylation by Sir’s recognition of deacetylated histones, the DNA-binding protein continues to play a critical part. In addition, of course, it is the DNA-binding protein that gives specificity to the whole process, that is, defines where the silencing complex forms. In some cases of silencing, RNA molecules, rather than proteins, provide this critical specificity. In Chapter 20, we discuss such a case, in which the RNAi

machinery of another yeast (*Schizosaccharomyces pombe*) is required for silencing at the mating-type loci and centromeres of that organism.

How is the spreading of silenced regions contained—that is, how is it limited to appropriate regions and prevented from spreading too far into the genome? We mentioned above that insulator elements can block the spread of histone modifications (Fig. 19-12). In addition, other kinds of histone modifications block binding of the Sir2 proteins, and thereby stop spreading. Methylation of the tail of histone H3 is believed to do this.

Histone methyltransferases attach methyl groups to histone tails. As we saw in Chapter 8, these enzymes add methyl groups to specific lysine residues in the tails of histones H3 and H4. Histone methyltransferases have recently been described in *S. cerevisiae*, where they are believed to help repression of some genes and, as just noted, block spreading of Sir2-mediated silencing in others. But histone methylases have been better characterized in higher eukaryotes and in the yeast *S. pombe*. In these organisms, silencing is typically associated with chromatin containing histones that are not only deacetylated, but methylated as well. Thus, methylation of Lys-9 in the H3 tail (H3K9) is a modification associated with silenced heterochromatin in these organisms (Chapter 8, Table 8-7). In contrast, other sites of methylation (e.g., Lys-4 on that same tail—H3K4) are associated with increased transcription.

### In *Drosophila*, HP1 Recognizes Methylated Histones and Condenses Chromatin

Just as acetylated residues within histones are recognized by proteins bearing bromodomains, methylated residues bind proteins with chromodomains (see Chapter 8, Fig. 8-41). One such protein is the *Drosophila* protein HP1, a component of silent heterochromatin in that organism (and with homologs serving a similar function in other organisms).

The HP1 protein interacts with modified chromatin containing methylated histone H3. This particular modification is produced by an enzyme encoded by *Su(Var)3-9*, a suppressor of so-called **variegation**. Variegation is seen in some cases when a gene is moved into a region of heterochromatin. Instead of being silenced in all cells all the time, the gene switches between the silenced and expressed state apparently at random, being “on” in some cells and “off” in others. Variegation is particularly evident for the so-called *white* gene, which is responsible for the normal red pigmentation of the eyes of adult flies. The gene is called *white* because the mutant phenotype is white eyes (see Appendix 1). When inserted into heterochromatin, expression of the *white* gene becomes “variegated,” producing eyes with salt and pepper red pigmentation (Fig. 19-28). Mutations



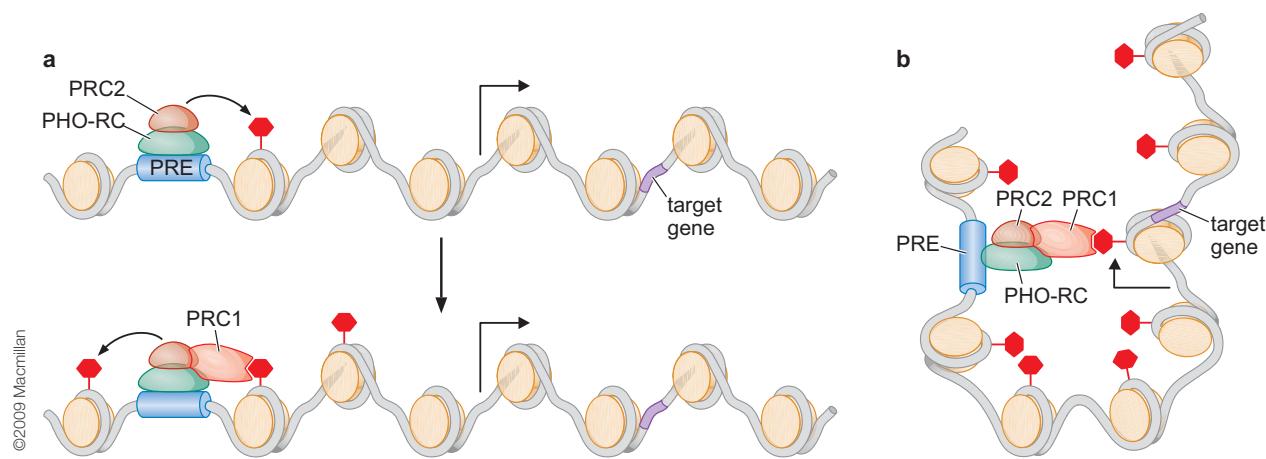
©2004 Macmillan

**FIGURE 19-28 Position effect variegation.** As described in the text, the *white* gene of *Drosophila* produces the red coloring of the wild-type eye (shown here on the right). When the gene is mutated, eyes are white (hence the name of the gene). When the wild-type gene is placed adjacent to heterochromatin, the expression is variegated, with some cells expressing the gene and some not. This results in the mottled coloring shown on the left. (Reprinted, with permission, from Lippman Z. and Martienssen R. 2004. *Nature* **431**: 364–370, Fig. 1a. © Macmillan.)

in the *Su(Var)3-9* gene suppress this variegation, producing eyes with a more uniform red pigmentation; expression of the *white* gene is no longer silenced in so many cells. The *Su(Var)3-9* protein is a histone H3 lysine-9 (H3K9) methyltransferase. Through mechanisms that are not presently understood, the *Su(Var)3-9* protein is recruited to heterochromatin, where it attaches methyl groups to histone H3 tails. This modification is essential for the binding of the HP1 protein, which, in turn, participates in the compaction of the heterochromatin. It is thought that *Su(Var)3-9* can also be recruited to specific euchromatic genes by sequence-specific DNA-binding proteins, thereby leading to gene-specific histone methylation and transcriptional repression by HP1 (see Box 19-5, Is There a Histone Code?, for further details and a figure).

### Repression by Polycomb Also Uses Histone Methylation

Histone methylation-triggered chromosome condensation is also used by **Polycomb** (Pc), an important group of repressors in animal cells. Pc repressors are found in two major protein complexes, Polycomb repressive complexes 1 and 2 (PRC1 and PRC2). The PRC2 complex is recruited by sequence-specific DNA-binding proteins (the repressive complex Pho-RC) that interact with so-called Polycomb Response Elements (PREs) (Fig. 19-29). PRC2 contains a histone methyltransferase (Enhancer of Zeste) that trimethylates lysine-27 (K27) on the tails of histone H3. This methylation leads to the recruitment of the PRC1 complex, which is thought to either condense chromatin or lead to the positioning of a nucleosome at or near the transcription start site. It was previously thought that Pc repression worked like HP1, with long-range spreading of histone methylation and chromosome condensation. But there is now evidence that PREs are often located near promoters. Nonetheless, H3K9 trimethylation is linked to gene silencing by HP1, whereas H3K27 trimethylation underlies gene silencing by Polycomb.



**FIGURE 19-29** Potential roles for trimethylated histone H3 on Lys27 (H3K27me3) in Polycomb group function. (a) The possible role of trimethylated histone H3 on Lys27 (H3K27me3) in Polycomb repressive complex 1 (PRC1) targeting (see text for details). (b) The potential role of H3K27me3 in chromatin looping. PRC1, anchored at the PRE, could interact with H3K27me3 to stabilize contact with the gene. The resulting DNA loop would juxtapose the Polycomb group complexes and the transcribed region, where PRC2 could further spread H3K27 methylation, compact local nucleosomes, and/or impede RNA polymerase. (Adapted, with permission, from Simon J.A. and Kingston R. 2009. *Nat. Rev. Mol. Cell Biol.* **10**: 697–709, Fig. 3. © Macmillan)

## ► ADVANCED CONCEPTS

**Box 19-5** Is There a Histone Code?

It has been proposed that a **histone code** exists. According to this idea, different patterns of modifications on histone tails at a given gene could be “read” to mean different things (Chapter 8, Fig. 8-39). The “meaning” would result from the particular pattern of modifications in each case recruiting a distinct set of proteins; the particular set depends on the number, type, and disposition of recognition domains carried by those proteins.

We have already encountered proteins that recognize specific acetylation or methylation “marks” on histones (e.g., TFIID and HP1). There are also proteins that phosphorylate serine residues in H3 and H4 tails and proteins that bind those modifications. Thus, multiple modifications at several positions in the histone tails are possible (Chapter 8, Fig. 8-39). Add to this the observation that many of the proteins that carry modification-recognition domains are themselves enzymes that modify histones further, and we start to see how a process of recognizing and maintaining patterns of modification could in principle be achieved.

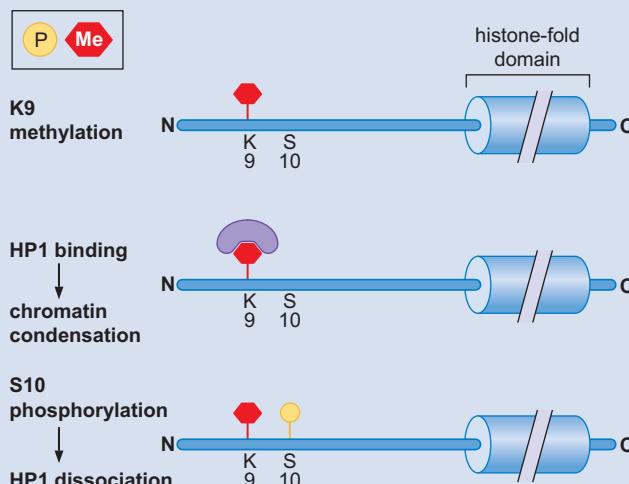
Consider one simple case—Lys-9 on the tail of histone H3 (see Chapter 8, Fig. 8-39). Different modification states of this residue could be interpreted to have different meanings. Thus, acetylation of this residue is associated with actively transcribed genes. This residue is recognized by various histone acetylases

bearing bromodomains, and these stimulate acetylation of nearby nucleosomes. When Lys-9 is unmodified, it is associated with silenced regions (as we saw in *S. cerevisiae* above). Unacetylated histones often recruit deacetylating enzymes, reinforcing and maintaining the deacetylated state (as we saw in the spreading of silenced regions in *S. cerevisiae*). Finally, that same lysine can in some organisms be methylated: in that case, the modified residue binds proteins (e.g., HP1) that establish and maintain a heterochromatic state.

But can *combinations* of modifications have distinct meanings? One example of how a histone modification can apparently influence a second modification present nearby is again illustrated by the *Drosophila* HP1 protein (Box 19-5 Fig. 1). During metaphase, the HP1 protein is temporarily lost from mitotic chromosomes, even though they retain the essential “mark”—namely, histone H3 methylation of lysine 9 (H3K9). Loss of HP1 binding is associated with phosphorylation of the neighboring serine residue at position 10 of H3. This phosphorylation is mediated by a cell cycle kinase called Aurora B. That kinase becomes active only during the M phase of the cell cycle, thereby causing the release of HP1 from the heterochromatin of metaphase chromosomes.

The dissociation of HP1 by the Aurora B kinase seems to be required for the attachment of the mitotic spindles to the centrosomes and the subsequent separation of sister chromatids during cytokinesis. When this process is complete, the phosphorylation of serine 10 is lost because of diminished Aurora B kinase activity, and HP1 reassociates with the chromosome to maintain the heterochromatin. Consistent with this model, mutations that eliminate Aurora B kinase lead to aberrant segregation.

Despite these observations, it remains highly controversial that a specific code exists, with complex patterns of histone modifications at a given locus generating a highly specific readout. Many of the modifications seen at a gene are likely just to be part of the process by which a gene is activated or repressed, rather than being the initiating signal; that is, they are a *consequence* of the gene being “on” or “off,” not the cause. Indeed, a recent study suggests that apparently normal strains of *Drosophila* can be propagated without any H3K4 methylation. Site-specific DNA-binding proteins (or, in some cases, as we shall see in Chapter 20, RNA molecules) remain the strongest provider of specificity determining when a given gene is expressed. Nonetheless, certain histone modifications are often associated with a particular state of gene activity. For example, H3K27 methylation is seen at many repressed genes, whereas H3K4 methylation is associated with active (or “poised”) genes.



**BOX 19-5 FIGURE 1** Influence of one chromatin modification on another. Modifications are shown on the tail of histone H3. Methylation of Lys-9 (K9) recruits HP1, which then effects chromatin condensation. Phosphorylation of the adjacent Ser residue (S10) displaces HP1 from methylated Lys-9, without removing the methyl group.

We have seen how individual types of modification can be involved in gene regulation. But what happens when multiple forms of modification occur at the same gene? How do their influences interact? It has been proposed that complicated patterns of modification operate as a “histone code.” The interactions between histone modifications and the idea of a histone code are described in Box 19-5, Is There a Histone Code?

### DNA Methylation Is Associated with Silenced Genes in Mammalian Cells

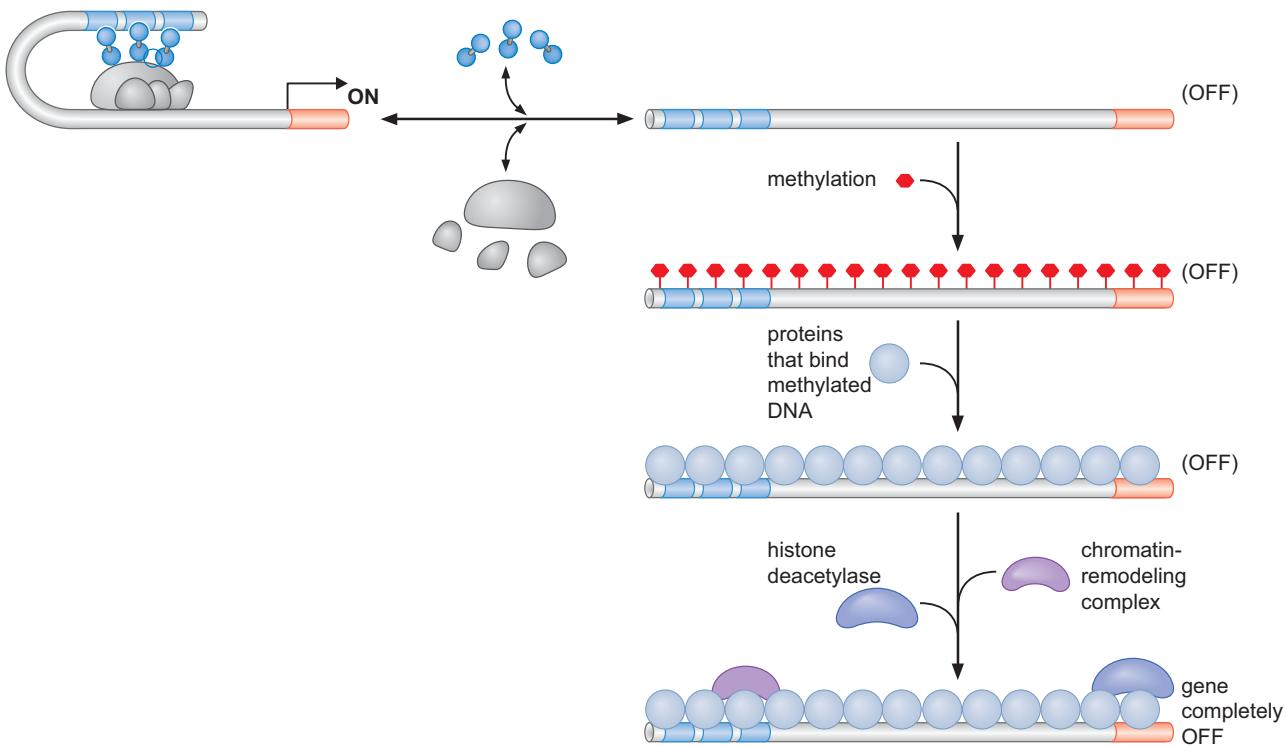
Some mammalian genes are kept silent by methylation of nearby DNA sequences (we are now talking about *DNA* methylation, not *histone* methylation). In fact, large regions of the mammalian genome are marked by methylation of DNA sequences, and often DNA methylation is seen in regions that are also heterochromatic. This is because methylated sequences are often recognized by DNA-binding proteins (such as MeCP2) that recruit histone deacetylases and histone methylases, which then modify nearby chromatin. Thus, methylation of DNA can mark sites where heterochromatin subsequently forms (Fig. 19-30).

DNA methylation lies at the heart of a phenomenon called **imprinting**, as we now describe. In a diploid cell, there are two copies of most genes: one copy on a chromosome inherited from the father and the other copy on the equivalent chromosome from the mother. In the majority of cases, the two alleles are expressed at comparable levels. This is hardly surprising—they carry the same regulatory sequences and are in the presence of the same regulators; they are also located at the same position along homologous chromosomes. But there are a few cases in which one copy of a gene is expressed while the other is silent.

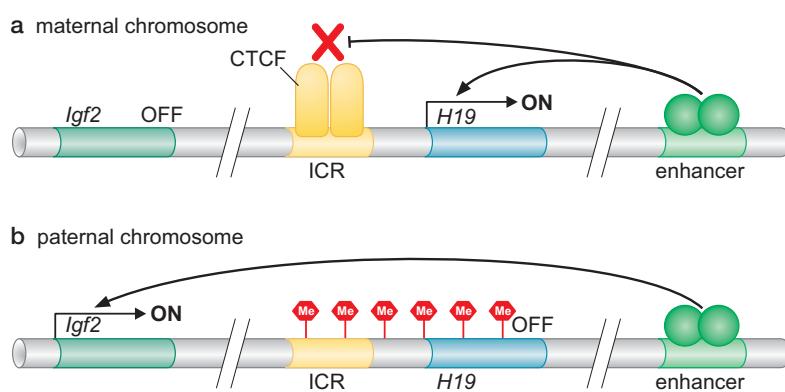
Two well-studied examples are the human *H19* and insulin-like growth factor 2 (*Igf2*) genes (Fig. 19-31). These are located close to each other on human chromosome 11. In a given cell, one copy of *H19* (that on the maternal chromosome) is expressed, whereas the other copy (on the paternal chromosome) is switched off; for *Igf2* the reverse is true—the paternal copy is on and the maternal copy is off.

Two regulatory sequences are critical for the differential expression of these genes: an enhancer (downstream from the *H19* gene) and an insulator (called the imprinting control region [ICR], located between the *H19* and *Igf2* genes). The enhancer (when bound by activators) can, in principle, activate either of the two genes. Why, therefore, does it activate only *H19* on the maternal chromosome and *Igf2* on the paternal chromosome? The answer lies in the role of the ICR and its methylation state. Thus, the enhancer cannot activate the *Igf2* gene on the maternal chromosome because on that chromosome, the ICR binds a protein, CTCF, that blocks activators at the enhancer from activating the *Igf2* gene (we discussed the function of insulators earlier, see Fig. 19-14). On the paternal chromosome, in contrast, the ICR element and the *H19* promoter are methylated. In that state, the transcriptional machinery cannot bind the *H19* promoter, and CTCF cannot bind the ICR. As a result, the enhancer now activates the *Igf2* gene. The *H19* gene is further repressed on the paternal chromosome by the binding of MeCP2 to the methylated ICR. This, as we have seen, recruits histone modifiers that repress the *H19* promoter.

Box 19-6, Transcriptional Repression and Human Disease, describes two cases in which loss of repression causes human diseases: one involves MeCP2 and the other involves defects in imprinting.



**FIGURE 19-30** Switching a gene off through DNA methylation and histone modification. In its unmodified state, the mammalian gene shown can readily switch between being expressed or not expressed in the presence of activators and the transcription machinery, as shown in the top line. In this situation, expression is never firmly shut off—it is leaky. Often that is not good enough; sometimes, a gene must be completely shut off, on occasion permanently. This is achieved through methylation of the DNA and modification of the local nucleosomes. Thus, when the gene is not being expressed, a DNA methyltransferase (a methylase) can gain access and methylate cytosines within the promoter sequence, the gene itself, and the upstream activator binding sites. The methyl group is added to the 5' position in the cytosine ring, generating 5-methylcytosine (see Chapter 4). This modification alone can disrupt binding of the transcription machinery and activators in some cases. But it can also increase binding of other proteins (e.g., MeCP2) that recognize DNA sequences containing methylcytosine. These proteins, in turn, recruit complexes that remodel and modify local nucleosomes, switching off expression of the gene completely.



**FIGURE 19-31** Imprinting. Shown are two examples of genes controlled by imprinting—the mammalian *Igf2* and *H19* genes. As described in the text, in a given cell, the *H19* gene is expressed only from the maternal chromosome, whereas *Igf2* is expressed from the paternal chromosome. The methylation state of the insulator element determines whether or not the ICR binding protein (CTCF) can bind and block activation of the *H19* gene from the downstream enhancer.

## EPIGENETIC GENE REGULATION

---

Patterns of gene expression must sometimes be inherited. A signal released by one cell during development causes neighboring cells to switch on specific genes. These genes may have to remain switched on in those cells for many cell generations, even if the signal that induced them is present only fleetingly. The inheritance of gene expression patterns, in the absence of the initiating signal, is called **epigenetic regulation**.

Contrast this with some of the examples of gene regulation we have discussed. If a gene is controlled by an activator and that activator is only active in the presence of a given signal, then the gene will remain on only as long as the signal is present. Indeed, under normal conditions, the *lac* genes of *E. coli* will only be expressed while lactose is present and glucose is absent. Likewise, the *GAL* genes of yeast are expressed only as long as glucose is absent and galactose is present, and human  $\beta$ -interferon is made only while cells are stimulated by viral infection.

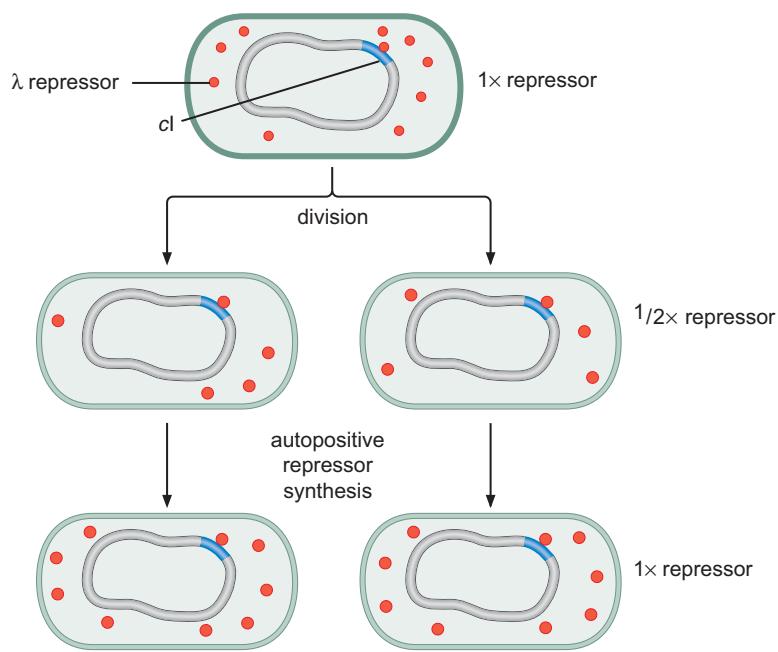
### Some States of Gene Expression Are Inherited through Cell Division Even When the Initiating Signal Is No Longer Present

We have already encountered examples of gene regulation that can be inherited epigenetically. Consider the maintenance of a bacteriophage  $\lambda$  lysogen (Chapter 18). In a lysogen, the phage is in a dormant state within the bacterial host cell. This state is associated with a specific pattern of gene expression and in particular with sustained expression of the  $\lambda$  repressor protein (see Chapter 18, Fig. 18-27).

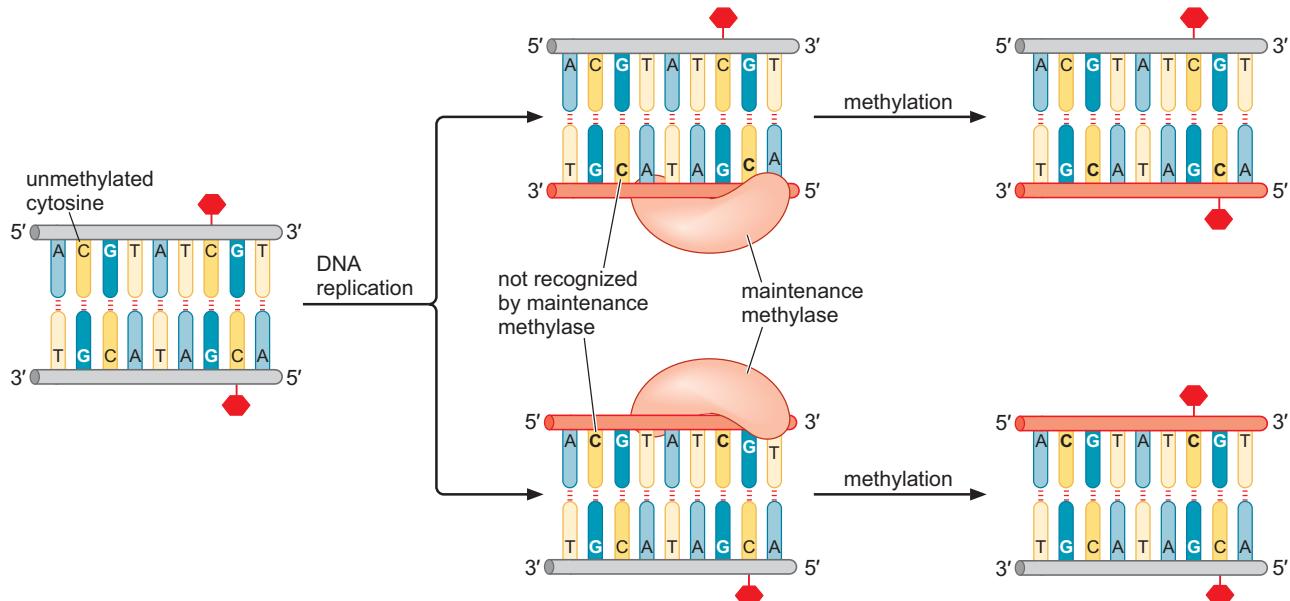
Lysogenic gene expression is established in an infected cell in response to poor growth conditions. Once established, however, the lysogenic state is maintained stably despite improvements in growth conditions: moving a lysogen into rich growth medium does not lead to induction. Indeed, induction essentially never occurs until a suitable inducing signal (such as UV light) is received.

Maintenance of the lysogenic state through cell division is thus an example of epigenetic regulation. This epigenetic control results from a two-step strategy for repressor synthesis. In the first step, synthesis is initially established through activation of the repressor (*cI*) gene by the activator CII (which is sensitive to growth conditions). In the second step, repressor synthesis is maintained by autoregulation: repressor activates expression of its own gene (see Chapter 18, Fig. 18-31). In this way, when the lysogenic cell divides, each daughter cell inherits a copy of the dormant phage genome and some repressor protein. This repressor is sufficient to stimulate further repressor synthesis from the phage genome in each cell (Fig. 19-32). Much of gene regulation during the development of multicellular organisms works in just this way. We shall see examples in Chapter 21.

Another known mechanism of epigenetic regulation is provided by DNA methylation, an example of which we saw in our description of imprinting. DNA methylation is reliably inherited through cell division, as shown in Figure 19-33. Thus, certain DNA methylases can methylate, at low frequency, previously unmodified DNA; but far more efficiently, so-called **maintenance methylases** modify hemimethylated DNA, the very substrate provided by replication of fully methylated DNA. In mammalian cells, DNA methylation may be the primary marker of regions of the genome that are silenced. Initial methylation is likely directed to particular sequences through the actions of DNA-binding proteins of RNAs (Chapter 20). After DNA replication, hemimethylated sites in both daughter cells are



**FIGURE 19-32** Epigenetic control of the maintenance of the lysogenic state.



**FIGURE 19-33** Patterns of DNA methylation can be maintained through cell division. As we saw in Figure 19-26, DNA involved in expression of a vertebrate gene can become methylated and expression of that gene switched off. This initial methylation is performed by a de novo methylase. For the shutdown state to keep a gene off permanently, the methylation state must be inherited through cell division. This figure shows how that is achieved. A DNA sequence is shown in which two cytosines are present on each strand—one methylated, the other not. This pattern is maintained through cell division, because, upon DNA replication, a maintenance methylase recognizes the hemimethylated DNA and adds a methyl group to the unmethylated cytosine within it. The completely unmethylated sequence is not recognized by this enzyme and thus remains unmethylated. Thus, both daughter DNA duplexes end up with the same pattern of methylation as the parent. (Adapted, with permission, from Alberts B. et al. 2002. *Molecular biology of the cell*, 4th ed., p. 431, Fig. 7-81. © Garland Science/Taylor & Francis LLC.)

## MEDICAL CONNECTIONS

### Box 19-6 Transcriptional Repression and Human Disease

Several human diseases are caused by the derepression of silenced genes. Here, we consider two conditions, each caused by the loss of repression of a gene whose mechanism of repression is described in the text: first, Rett syndrome, which is caused by the loss of the repressor protein MeCP2; and second, Beckwith–Wiedemann syndrome, which is caused by loss of binding sites for CTCF in the ICR of the *Igf2* gene.

**Rett syndrome (RTT)** is a severe autism spectrum disorder found in one in 10,000 girls. This condition is characterized by loss of language and motor skills in early childhood, microcephaly, seizures, stereotypical behaviors (such as repetitive hand-wringing), and intermittent hyperventilation. It is a common cause of sporadic mental retardation. RTT is caused by a mutation in the X-linked gene encoding the repressor protein MeCP2. We encountered this transcriptional regulator earlier in the text; it recognizes methylated DNA sequences and silences transcription of nearby genes through recruitment of histone deacetylases (Fig. 19-31). Mice carrying a disrupted *MeCP2* gene have symptoms similar to those of RTT patients, and this is preserved in mice in which *MeCP2* loss is restricted to the brain.

Because the *MeCP2* gene is found on the X chromosome, girls with a defective copy (the RTT patients) are a mosaic: in those cells in which the X chromosome carrying the mutant copy (allele) of the gene is inactivated, wild-type MeCP2 is made; but in those cells in which the wild-type copy of the gene is on the inactive copy of the X chromosome, no MeCP2 is made. Boys carrying the mutant gene on their (single) X chromosome lack MeCP2 in all cells and usually die from respiratory failure within a year or two.

RTT is thought to be a neurodevelopmental condition rather than a neurodegenerative disorder, because patients—and the knockout mice—show abnormal neuronal morphology but not neuronal death. Even so, it was thought likely that the MeCP2 deficiency was critical at a particular point during development, after which even restoring its function would not reverse the phenotype. But recently a mouse was constructed in which MeCP2 expression could be manipulated, allowing the mouse to be grown to adulthood without MeCP2 expression, before switching on expression of that regulator. Remarkably, making MeCP2 in adulthood was sufficient to reverse the effects of its absence throughout earlier development. This exciting finding

makes therapeutic intervention in humans more feasible, if still difficult.

The link between MeCP2 and the symptoms of RTT is not fully understood. As we have seen, MeCP2 is a repressor of gene expression, and one of its target genes encodes brain-derived neurotrophic factor (BDNF). This protein, a growth factor, has roles in brain development and in synaptic changes associated with learning and memory. Recently, it has been found that neural activity leads to phosphorylation of MeCP2, a modification that causes the repressor to dissociate from DNA, presumably allowing expression of its target genes. Disruption of these activities—inappropriate expression of BDNF, for example—have obvious appeal as an explanation for at least some of the cognitive symptoms of RTT.

There is an ongoing search for the links between MeCP2 and BDNF, and between these proteins and the disease, as there is for other possibly relevant genes regulated by MeCP2. The broad array of symptoms—from cognitive impairment to unusual gait—suggests that there are probably several genes whose misexpression is required for the full disease.

**Beckwith–Wiedemann syndrome (BWS)** is a developmental disorder affecting one in 15,000. The condition is characterized by overgrowth (children with this condition are born prematurely and are larger than normal) and increased susceptibility to a variety of childhood cancers (including Wilms' tumor). The syndrome is also associated with disrupted expression of imprinted genes on chromosome 11p15.5, including the Insulin-like growth factor 2 gene (*Igf2*) discussed in the text (Fig. 19-31).

As we have described, the *Igf2* gene is usually expressed mono-allelically; that is, only one of the two alleles (in this case, the paternal allele) is expressed as a result of imprinting of the other. At the same time, the *H19* gene is expressed only from the maternal allele. Many cases of BWS are associated with biallelic expression of *Igf2* and no expression of *H19*, the result of methylation of ICR on both chromosomes. IGF2 is a fetal growth factor, and *H19* is a regulatory RNA (see Chapter 20) believed to be involved in tumor suppression, and thus the phenotype of the condition—overgrowth and tumor sensitivity—makes sense.

As with RTT, the symptoms of BWS are mimicked in suitably manipulated mice: overexpression of *Igf2* produces general overgrowth of mice and the appearance of specific tumors.

remethylated. These can then be recognized by the repressor MeCP2, which, in turn, recruits histone deacetylases and methylases, re-establishing silencing (Fig. 19-30).

Nucleosome modifications could in principle provide the basis for epigenetic inheritance, although no examples of this have yet been found. Consider a gene switched off by a stretch of methylated histones. When that region of the chromosome is replicated during cell division, the methylated histones from the parental DNA molecule end up distributed equally between the two daughter duplexes (see Chapter 8, Fig. 8-43). Thus, each

of the daughter molecules carries some methylated and some unmethylated nucleosomes. The methylated nucleosomes could recruit proteins bearing chromodomains, including the histone methylase itself, which could then methylate the adjacent unmodified nucleosomes. In this way, the state of chromatin modification could be maintained through generations using the same strategy used to achieve spreading. Although this is an appealing model, it is yet to be seen to operate in the absence of direction by DNA methylation, DNA-binding proteins, or regulatory RNAs (Chapter 20) (i.e., in the absence of specific signals determining which cells recruit the modification enzymes).

## SUMMARY

---

As in bacteria, transcription initiation is the most frequently regulated step in gene expression in eukaryotes, despite the additional steps that can be regulated in these organisms. Also as in bacteria, transcription initiation is typically regulated by proteins that bind to specific sequences on DNA near a gene and either switch that gene on (activators) or switch it off (repressors). This conservation of regulatory mechanism holds in the face of several complexities in the organization and transcription of eukaryotic genes not found in bacteria.

The DNA in a eukaryotic cell is wrapped in histones to form nucleosomes. Thus, the DNA sequences to which the transcriptional machinery and the regulatory proteins bind are in many cases occluded. Enzymes that modify histones, by adding (or removing) small chemical groups, alter the histones in two possible ways: changing how tightly the nucleosomes are packed (and thus how accessible the DNA within them is) and forming (or removing) binding sites for other proteins involved in transcribing the gene. Other enzymes “remodel” the nucleosomes: they use the energy from ATP hydrolysis to move the nucleosomes around, influencing which sequences are available. An important mechanism of transcriptional activation is the removal of nucleosomes at the core promoter. It is possible that inactive but bound forms of Pol II (paused or “poised” Pol II) render promoter regions in an “open” or nucleosome-free condition, thereby fostering rapid and efficient induction of gene expression when appropriate signals become available.

Genes of multicellular eukaryotes are typically controlled by more regulatory proteins than their bacterial counterparts, some bound far from the gene. This reflects the larger number of physiological signals that control a typical gene in multicellular organisms.

The enzyme RNA polymerase is largely conserved between bacteria and eukaryotes (Chapter 13). But there are approximately 50 or so additional proteins that bind at the typical eukaryotic promoter along with polymerase. Many of these proteins come to the promoter as large protein complexes.

In eukaryotes, just as we saw in bacteria, activators predominantly work by recruitment. In these organisms, however, the activators do not recruit polymerase directly, or alone. Thus, they recruit the other protein complexes required to initiate transcription of a given gene. RNA polymerase itself

is brought in along with these other complexes. The activator can recruit histone-modifying enzymes as well, and the effects of those modifications may help the transcriptional machinery bind the promoter or initiate efficient transcription.

The activators can interact with one or more of many different components of the transcriptional machinery or the nucleosome modifiers. Gal4, for example, recruits Mediator, SAGA, Swi/Snf, and TFIID to promoters as required. In other cases, factors required for efficient initiation or elongation might be needed after the polymerase has bound—these too can be recruited by activators. This explains how activators can so readily work together in large numbers and various combinations and accounts for the widespread use of signal integration and combinatorial control that we see, particularly in multicellular organisms.

Some activators work from sites far from the gene, requiring that the DNA between their binding sites and the promoter loops out. How loops can form over the very large distances called for in some cases is not clear, but it might involve changes in the chromatin structure between the activator-binding site and the promoter, bringing those two elements closer together, and the use of cohesin to stabilize loops. DNA sequences called insulators bind proteins that interfere with the interaction between activators bound at distant enhancers and their promoters. These could work by inhibiting mechanisms that facilitate looping (such as changes in chromatin structure). Insulators help ensure that activators work only on the correct genes.

Eukaryotic repressors work in various ways. However, the most common mechanism seen in bacteria—repressor binding to a site overlapping the promoter—is not typically seen in eukaryotes. In some cases repressors bind near activators in enhancers and prevent those activators from mediating looping of the enhancer to the promoter. In addition, eukaryotic repressors work by recruiting histone modifiers that reduce transcription. For example, whereas a histone acetylase is typically associated with activation, a histone deacetylase—that is, an enzyme that removes acetyl groups—acts to repress a gene.

In some cases, long stretches of nucleosomal DNA can be kept in a relatively inert state by appropriate nucleosome modification, most notably deacetylation and methylation. In this way, groups of genes can be kept in a “silent” state without the need for specific repressors bound at each individual gene.

In some eukaryotic organisms, such as mammals, silent genes are also associated with methylated DNA. Methylated sequences can either block the binding of the transcription machinery and activators or specifically bind a class of repressors that recruit histone-modifying enzymes that repress nearby genes. DNA methylation can be maintained through cell division, and thus patterns of gene expression controlled by that methylation can be as well.

If expression of a gene is maintained in some state through cell division—in the absence of either a mutation or the

signal that initiated that pattern—it is said to be inherited epigenetically. Auto-regulatory transcription factor loops can achieve this. Also, DNA methylation can affect gene expression and readily be inherited, as we have seen. Epigenetic inheritance through the use of histone modifications is often discussed as a possibility, but the extent to which they fulfill this role in the absence of input from DNA-binding proteins, regulatory RNAs, or DNA methylation remains unclear.

## BIBLIOGRAPHY

### Books

- Allis C.D., Jenuwein T., Reinberg D., and Caparros M.-L. eds. 2007. *Epigenetics*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
- Carey M., Smale S.T., and Pederson C.L. 2008. *Transcriptional regulation in eukaryotes: Concepts, strategies, and techniques*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
- Ptashne M. and Gann A. 2002. *Genes and signals*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

### DNA Recognition

- Garvie C.W. and Wolberger C. 2001. Recognition of specific DNA sequences. *Mol. Cell.* **8**: 937–946.
- Harrison S.C. 1991. A structural taxonomy of DNA-binding domains. *Nature* **353**: 715–719.

### Activation

- Bjorklund S. and Gustafsson C.M. 2005. The yeast Mediator complex and its regulation. *Trends Biochem. Sci.* **30**: 240–244.
- Fuda N.J., Ardehali M.B., and Lis J.T. 2009. Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature* **461**: 186–192.
- Jones K.A. and Kadonaga J.T. 2000. Exploring the transcription–chromatin interface. *Genes Dev.* **14**: 1992–1996.
- Kim Y.J. and Lis J.T. 2005. Interactions between subunits of *Drosophila* Mediator and activator proteins. *Trends Biochem. Sci.* **30**: 245–249.
- Kornberg R.D. 2005. Mediator and the mechanism of transcriptional activation. *Trends Biochem. Sci.* **30**: 235–239.
- Levine M. 2011. Paused RNA polymerase II as a developmental checkpoint. *Cell* **145**: 502–511.
- Luo Z., Lin C., and Shilatifard A. 2012. The super elongation complex (SEC) family in transcriptional control. *Nat. Rev. Mol. Cell Biol.* **13**: 543–547.
- Malik S. and Roeder R.G. 2005. Dynamic regulation of Pol II transcription by the mammalian Mediator complex. *Trends Biochem. Sci.* **30**: 256–263.
- Ptashne M. 2005. Regulation of transcription: From lambda to eukaryotes. *Trends Biochem. Sci.* **30**: 275–279.

### Repression

- Liu Z. and Karmarkar V. 2008. Groucho/Tup1 family co-repressors in plant development. *Trends Plant Sci.* **13**: 137–144.
- Simon J.A. and Kingston R.E. 2009. Mechanisms of Polycomb gene silencing: Knowns and unknowns. *Nat. Rev. Mol. Cell Biol.* **10**: 697–708.

- Smith R.L. and Johnson A.D. 2000. Turning genes off by Ssn6-Tup1: A conserved system of transcriptional repression in eukaryotes. *Trends Biochem. Sci.* **25**: 325–330.

### Nucleosome Modifiers and Transcriptional Regulation

- Hargreaves D.C. and Crabtree G.R. 2011. ATP-dependent chromatin remodeling: Genetics, genomics and mechanisms. *Cell Res.* **21**: 396–420.
- Henikoff S. and Shilatifard A. 2011. Histone modification: Cause or cog? *Trends Genet.* **27**: 389–396.
- Jenuwein T. and Allis C.D. 2001. Translating the histone code. *Science* **293**: 1074–1080.
- Narlikar G.J., Fan H.Y., and Kingston R.E. 2002. Cooperation between complexes that regulate chromatin structure and transcription. *Cell* **108**: 475–487.
- Rando O.J. and Winston F. 2012. Chromatin and transcription in yeast. *Genetics* **190**: 351–387.
- Shahbazian M.D. and Grunstein M. 2007. Functions of site-specific histone acetylation and deacetylation. *Annu. Rev. Biochem.* **76**: 75–100.
- Wang X., Bai L., Bryant G.O., and Ptashne M. 2011. Nucleosomes and the accessibility problem. *Trends Genet.* **27**: 487–492.

### Silencing, Imprinting, and Epigenetics

- Gartenberg M.R. 2000. The Sir proteins of *Saccharomyces cerevisiae*: Mediators of transcriptional silencing and much more. *Curr. Opin. Microbiol.* **3**: 132–137.
- Goldberg A.D., Allis C.D., and Bernstein E. 2007. Epigenetics: A landscape takes shape. *Cell* **128**: 635–638.
- Gottschling D.E. 2004. Summary: Epigenetics—From phenomenon to field. *Cold Spring Harb. Symp. Quant. Biol.* **69**: 507–519.
- Klose R.J. and Bird A.P. 2005. Genomic DNA methylation: The mark and its mediators. *Trends Biochem. Sci.* **31**: 89–97.
- Ptashne M. 2007. On the use of the word “epigenetic”. *Curr. Biol.* **17**: R233–R236.
- Wood A.J. and Oakey R.J. 2006. Genomic imprinting in mammals: Emerging themes and established theories. *PLoS Genet.* **2**: 1677–1685.

### Enhancers, Combinatorial Control, and Synergy

- Arnoldi D.N. and Kulkarni M.M. 2005. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J. Cell Biol.* **94**: 890–898.
- Li H. and Johnson A.D. 2010. Evolution of transcription networks—Lessons from yeasts. *Curr. Biol.* **20**: R746–R753.

- Merika M. and Thanos D. 2001. Enhanceosomes. *Curr. Opin. Genet. Dev.* **11**: 205–208.  
 Rokas A. 2006. Evolution: Different paths to the same end. *Nature* **443**: 401–402.

### Long-Range Interactions

- Buecker C. and Wysocka J. 2012. Enhancers as information integration hubs in development: Lessons from genomics. *Trends Genet.* **28**: 276–284.  
 Dean A. 2006. On a chromosome far, far away: LCRs and gene expression. *Trends Genet.* **22**: 38–45.  
 Dorsett D. and Ström L. 2012. The ancient and evolving roles of cohesin in gene expression and DNA repair. *Curr. Biol.* **22**: R240–R250.  
 Gaszner M. and Felsenfeld G. 2006. Insulators: Exploiting transcriptional and epigenetic mechanisms. *Nat. Rev. Genet.* **7**: 703–713.  
 Li Q., Barkess G., and Qian H. 2006. Chromatin looping and the probability of transcription. *Trends Genet.* **22**: 197–202.

### Signals and Signal Transduction

- Bromberg J.F. 2001. Activation of STAT proteins and growth control. *Bioessays* **23**: 161–169.  
 Brown M.S., Ye J., Rawson R.B., and Goldstein J.L. 2000. Regulated intramembrane proteolysis: A control mechanism conserved from bacteria to humans. *Cell* **100**: 391–398.  
 Pawson T. and Nash P. 2000. Protein–protein interactions define specificity in signal transduction. *Genes Dev.* **14**: 1027–1047.

### Repression and Disease

- Gabellini D., Green M.R., and Tupler R. 2004. When enough is enough: Genetic diseases associated with transcriptional derepression. *Curr. Opin. Genet. Dev.* **14**: 301–307.  
 Kriaucionis S. and Bird A. 2003. DNA methylation and Rett syndrome. *Hum. Mol. Genet.* **12**: R221–R227.  
 Miller G. 2007. Medicine. Rett symptoms reversed in mice. *Science* **315**: 749.

## QUESTIONS

### MasteringBiology®

For instructor-assigned tutorials and problems, go to MasteringBiology.

For answers to even-numbered questions, see Appendix 2: Answers.

**Question 1.** The most common type of regulation of gene expression occurs at the level of transcription. Name other types of regulation for gene expression in eukaryotic cells. Are there any types of regulation unique to eukaryotic cells versus prokaryotic cells?

**Question 2.** Compare and contrast the DNA sequence elements involved in regulating transcription in bacterial cells and eukaryotic cells.

**Question 3.** Explain why *E. coli lac Z* is often used as a reporter gene in *S. cerevisiae* cells but not in *E. coli* cells.

**Question 4.** The presence of a DNA template (e.g., a product from PCR), general transcription factors, and RNA polymerase II allows for the initiation of transcription in vitro. Explain why the initiation of transcription is not possible with only general transcription factors and RNA polymerase II using the genomic DNA as a template.

**Question 5.** Review the principles of the two-hybrid assay in Box 19-1. Where in the cell must the bait and prey interaction take place to observe activation of the reporter gene? Provide one potential drawback to the yeast two-hybrid assay with respect to this requirement. What other potential drawbacks can you imagine?

**Question 6.**

- A. A list of the steps for chromatin immunoprecipitation (ChIP) in the incorrect order follows. Provide the proper order for the steps of ChIP by listing the letter of each step.
- Immunoprecipitate DNA–protein complex.
  - Amplify DNA by PCR.

- Add antibody specific to one protein.
  - Cross-link proteins to DNA fragments.
  - Remove proteins.
- B. A researcher chose to perform a ChIP assay using an antibody to protein X and PCR primers specific to the region including gene Y. Give an example of the question asked by the researcher that led to these choices.

**Question 7.** Recall the use of combinatorial control in regulation of the *S. cerevisiae* mating type specific gene expression. Describe how Mcm1 serves to repress and activate transcription of target genes in haploid *MAT*α cells.

**Question 8.**

- Give examples of the types of covalent modifications to histones discussed in Chapter 19 that influence gene expression.
- Explain why the presence of a given modification is not always associated uniquely with activation or repression.

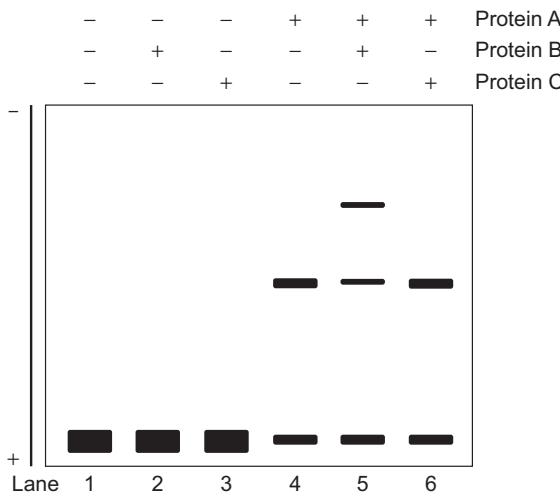
**Question 9.**

- Describe the role of DNA methylation in gene expression in mammalian cells.
- How does the role of DNA methylation in mammalian cells differ from the role of DNA methylation in *E. coli* cells?

**Question 10.** Players in a signal transduction pathway can be classified into the following categories: signal, receptor, relay molecule, and output. Classify each component of the STAT pathway pictured in Figure 19-22 as signal, receptor, relay molecule, and output. You can use a term more than once.

**Question 11.** Hypothesize why imprinting leads to non-Mendelian inheritance (traits segregating in a pattern not fitting a Mendelian pattern). For review of Mendelian genetics, see Chapter 1.

**Question 12.** You decide to study three proteins—Protein A, Protein B, Protein C—with potential roles in transcriptional regulation in mammalian cells. You perform an electrophoretic mobility shift assay (EMSA). The data is shown below. For review of this technique, see Chapter 7. All reactions contained binding buffer and labeled DNA fragment that included the sequence for the DNA-binding site for Protein A, a sequence upstream of a mammalian gene. The added purified protein(s) are indicated above the gel.



- Based on this data, does Protein A bind DNA? Explain your answer.
- Provide an explanation for the result observed in lane 5. Use the data in lane 2 as well in your explanation.
- Propose a model for how Protein B potentially activates transcription.

**Question 13.** You want to make a transgenic line of the nematode *C. elegans*. The xyz gene is expressed in every cell of the embryo. You fuse the promoter (*Pxyz*, 300 bp) for the xyz gene to the gene encoding green fluorescent protein (GFP) (900 bp). You inject the construct (pictured below) into the worm. The construct randomly inserts into the genome and is then stably inherited to all progeny.



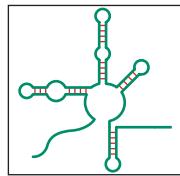
Although xyz should be expressed in every cell of the embryo, the progeny of this injected adult do not express GFP. You are confident in your injection abilities and believe that the construct did, in fact, integrate into the genome of the injected worm.

- You amplify and sequence the integrated construct and flanking regions in both directions. On one side, your construct is flanked by a 5000-bp region of random sequence. On the other side, your construct is flanked by 200 “TAAGGC” repeats. *Based on these results*, propose a hypothesis as to why your construct is not expressed.

You decide to reinject the same construct into another adult worm in an attempt to integrate the construct elsewhere in the genome. You are happy to find that the embryos from this adult worm all show GFP expression in every cell (by fluorescent microscope observations).

- You want to study the subcellular localization of protein XYZ. Suggest a modification to the construct from part A that would allow you to study Protein XYZ localization in the cell. Explain your reasoning.

CHAPTER 20



# Regulatory RNAs

WE DISCUSSED IN THE PREVIOUS TWO CHAPTERS how transcription is regulated in prokaryotes and eukaryotes. We learned that this control is achieved using regulatory proteins—typically, sequence-specific DNA-binding proteins that either activate or repress transcription of nearby genes. The mechanistic details of gene regulation have been studied since François Jacob and Jacques Monod proposed their model of repression more than 50 years ago (Chapter 18, Box 18-2). At that time, they could not say whether the *trans* factors (repressors) were proteins or RNA. It transpired that in the cases they studied (and, indeed, the majority of other cases), the regulators were proteins that worked by binding the operator sites on DNA. But in their original paper they suggested that the regulators could just as easily be RNA molecules—indeed, they favored that possibility.

The idea that RNA molecules might be regulators was largely forgotten as more and more protein regulators were found in both prokaryotes and eukaryotes. But in recent years, there has been an explosion in the study of RNA regulators, particularly in eukaryotes, that operate at the level of transcription and especially translation. This new field emerged from two sources: the discovery of microRNAs, first reported in the early 1990s, and then the discovery of the phenomenon known as RNA interference in the late 1990s. Before we describe these forms of regulation—how they work and the applications they afford researchers—we consider cases of RNA-mediated gene regulation in bacteria.

## REGULATION BY RNAs IN BACTERIA

Small RNAs have been recognized in prokaryotes for many years. Some are involved in regulating the replication of plasmids, and others are involved in regulating gene expression (see the discussion of Tn10 in Chapter 12). Of the latter group, some of these RNAs control transcription—the 6S RNA of *Escherichia coli*, for example. This RNA binds to the  $\sigma^{70}$  subunit of RNA polymerase and down-regulates transcription from many  $\sigma^{70}$  promoters. The 6S RNA accumulates at high levels in stationary phase (the growth-phase bacteria enter as nutrients become depleted and the cells stop dividing). In stationary phase, an alternative  $\sigma$  factor,  $\sigma^S$ , is made. This  $\sigma$  competes with  $\sigma^{70}$  for core polymerase and directs the enzyme to promoters expressing genes for the multiple stress responses needed to survive stationary phase. By down-regulating transcription from  $\sigma^{70}$  promoters, 6S RNA helps this shift in expression to  $\sigma^S$  promoters.

## O U T L I N E

- Regulation by RNAs in Bacteria, 701
- Regulatory RNAs are Widespread in Eukaryotes, 711
- Synthesis and Function of miRNA Molecules, 714
- Silencing Gene Expression by Small RNAs, 718
- Long Non-Coding RNAs and X-Inactivation, 728
- Visit Web Content for Structural Tutorials and Interactive Animations

In recent years, attention has focused on small RNA molecules in bacteria that regulate translation and mRNA degradation. Interest in these small RNAs has been heightened by their similarity to RNAs that regulate gene expression in eukaryotes—the small interfering and microRNAs we discuss in the second half of this chapter.

One class of bacterial regulatory RNAs (called **sRNAs**) acts in *trans* to control translation of target genes, rather as microRNAs do in eukaryotes. They are, however, larger (80–110 nucleotides) than those eukaryotic regulatory RNAs (which range from 21 to 30 nucleotides), and they are not generally formed by processing of larger double-stranded RNA (dsRNA) precursors (as those eukaryotic RNA regulators are); instead, they are encoded in their final form by small genes. Many of these genes have been identified by bioinformatics, with more than 100 sRNAs being uncovered in *E. coli*. Most sRNAs work by base pairing with complementary sequences within target mRNAs and directing destruction of the mRNA, inhibiting its translation or even in some cases *stimulating* translation.

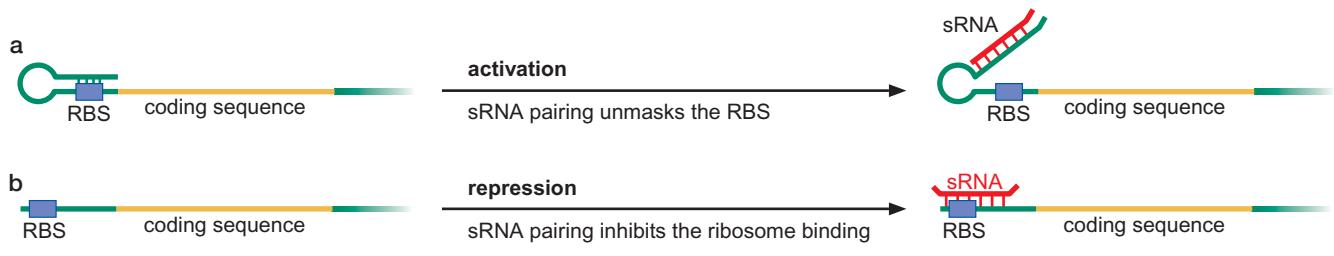
Binding of an sRNA to its target mRNA is in most cases aided by the bacterial protein Hfq. This RNA chaperone is needed because the complementarity between the sRNAs and their target mRNAs is typically imperfect and short, and thus their interaction is weak. Hfq facilitates base pairing. Also, by binding the sRNAs even before they are paired with their targets, Hfq increases the stability of these regulators.

A well-studied sRNA from *E. coli* is the 81-nucleotide RybB RNA. This sRNA binds several target mRNAs and triggers their destruction because the double-strand stretch of heteroduplex formed upon pairing is recognized as a substrate by the nuclease RNase E. Most of the mRNAs targeted by RybB encode iron storage proteins. Free iron is required by the cell under certain circumstances, but high levels are toxic. RybB regulates the levels of free iron by controlling the levels of iron storage proteins. RybB is expressed from a promoter recognized by a special  $\sigma$  factor called  $\sigma^E$  (like  $\sigma^S$ , a stress response  $\sigma$  factor). Expression of the gene encoding  $\sigma^E$  is itself regulated by RybB, and thus this sRNA is part of an autoregulatory loop for  $\sigma^E$ .

The stationary-phase  $\sigma$  factor  $\sigma^S$  mentioned above is encoded by the *rpoS* gene of *E. coli*. Translation of *rpoS* mRNA is stimulated by two sRNAs: DsrA and RprA. Activation is achieved by a switch in alternative RNA base pairing: the small RNAs bind to a region of the mRNA that otherwise would pair with the ribosome-binding site, inhibiting translation. The *rpoS* gene is also acted on negatively by another small RNA, OxyS. Figure 20-1 shows these two mechanisms.

Other examples of regulatory RNAs in bacteria act simply as “antisense” RNAs: they are encoded by the strand opposite the coding strand of a gene and act through homologous base pairing to inhibit expression of the mRNA produced from that gene. These tend to be associated with genes encoding potentially toxic products, and also in regulation of some phage genes (as in  $\lambda$ ) (see Chapter 18). These RNAs are often said to act in *cis* because they act only on the gene from which they are made (in contrast to the *trans*-acting sRNAs described above).

We return to *trans*-acting regulatory RNAs in the second half of this chapter, where we consider their role in regulating gene expression in eukaryotes. But before turning to that topic, we consider other examples in bacteria of gene regulation mediated through alternative RNA pairing that truly operate in *cis*. These are RNA regulatory elements that control expression of the genes *within whose mRNAs they reside*. The most striking examples are the so-called **riboswitches** that control metabolic operons and **attenuation** in biosynthetic operons. The *trp* genes of *E. coli* are the classic



**FIGURE 20-1** Activation and repression of translation by sRNAs. When the ribosome-binding site (RBS) is occluded by base pairing with another RNA molecule (as in part b) or another region of the same RNA molecule (as in part a), translation is inhibited. (Adapted, with permission, from Gottesman S. et al. 2006. *Cold Spring Harbor Symp. Quant. Biol.* 71: 1–11, Fig. 1. © Cold Spring Harbor Laboratory Press.)

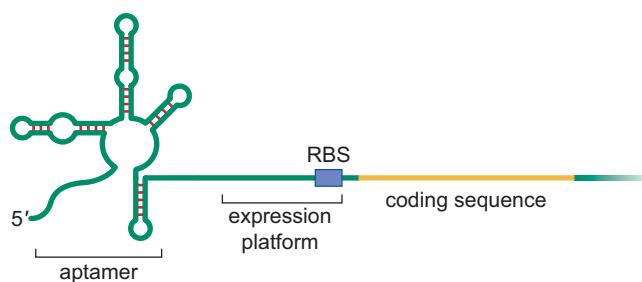
example of the latter mechanism and are where RNA-mediated regulation was discovered (we describe that case in detail in Box 20-1).

### Riboswitches Reside within the Transcripts of Genes Whose Expression They Control through Changes in Secondary Structure

Riboswitches control gene expression in response to changes in the concentrations of small molecules. These regulatory elements are typically found within the 5'-untranslated regions (5'-UTRs) of the genes they control, and they can regulate expression at the level of transcription or translation. They do this through changes in RNA secondary structure, as we shall see.

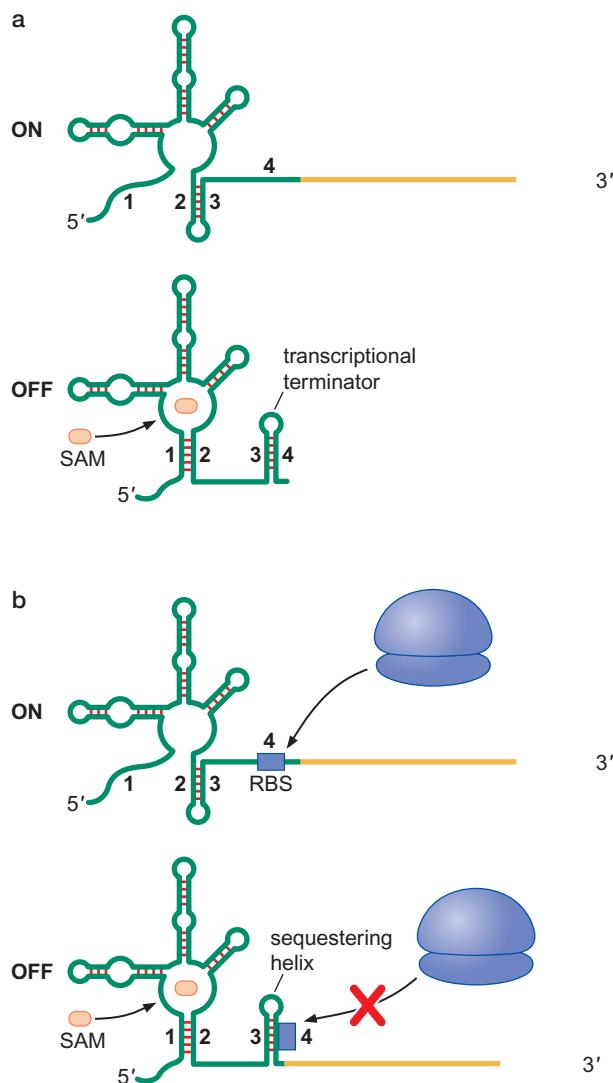
Each riboswitch is made up of two components: the **aptamer** and the **expression platform** (Fig. 20-2). The aptamer binds the small-molecule ligand and, in response, undergoes a conformational change, which, in turn, causes a change in the secondary structure of the adjoining expression platform. These conformational changes alter expression of the associated gene by either terminating transcription or inhibiting the initiation of translation. Both mechanisms are illustrated in the example shown in Figure 20-3, which we now describe.

Riboswitches are, not surprisingly, typically found upstream of genes involved in the synthesis of the metabolite ligand recognized by the riboswitch in question. For example, in *Bacillus subtilis*, many genes involved in the use of the amino acid methionine have a 200-nucleotide-long untranslated leader RNA that acts as a SAM (*S*-adenosylmethionine)-sensing riboswitch. RNA polymerase initiates transcription at the promoter and transcribes through this leader region before entering the coding sequence of the downstream genes. Once transcribed into RNA, the leader region can adopt alternative structures through alternative patterns of intramolecular base



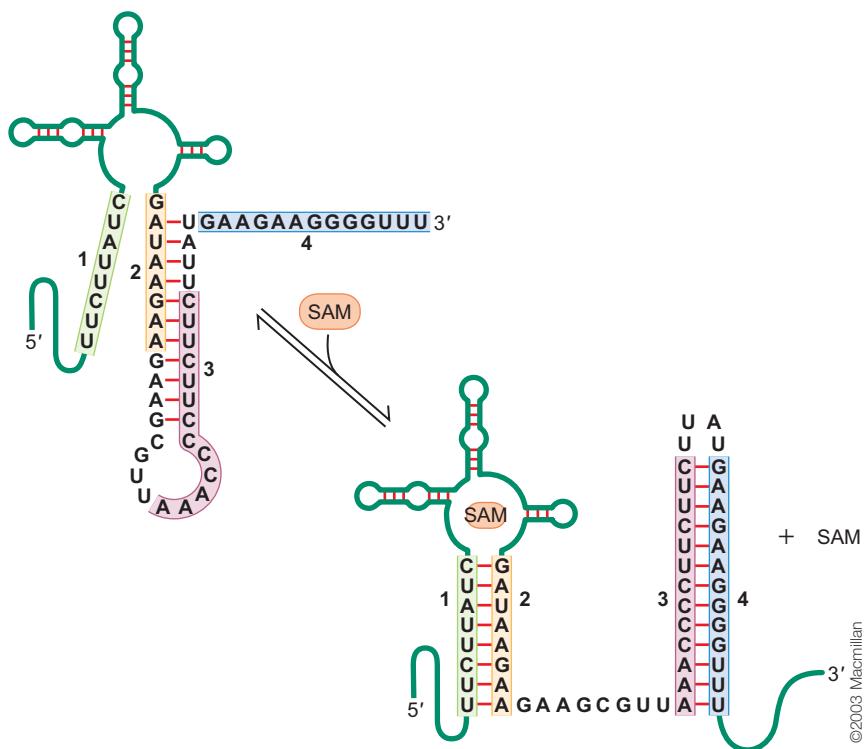
**FIGURE 20-2** Organization of riboswitch RNAs. As described in the text, the aptamer binds the controlling metabolite, causing changes in the structure of the adjoining expression platform. The aptamers identified to date vary in size from 70 to 200 nucleotides; the expression platforms vary more in both size and character.

**FIGURE 20-3** Riboswitches regulate transcription termination or translation initiation. Two examples of a SAM-sensing riboswitch, in one case (a) regulating transcription termination, in the other (b) translation initiation. Numbers 1–4 indicate different sequence elements within the RNA upstream of the coding region (yellow). In the absence of SAM, regions 2 and 3 form a stem-loop; in the presence of SAM, regions 1 and 2 form a stem-loop, and regions 3 and 4 do likewise. The consequence of that change in secondary structure controls transcription or translation as shown. (a) A stem-loop of regions 3 and 4 produces a transcriptional terminator, which triggers RNA polymerase to terminate transcription immediately after transcribing those regions and before entering the downstream coding region. The stem-loop in this case is followed by a stretch of Us in the mRNA, another feature of the transcriptional terminator (Chapter 13, Fig. 13-13). (b) The stem-loop formed by regions 3 and 4 inhibits translation initiation by sequestering the ribosome-binding site, as shown.



pairing (Fig. 20-3a). One arrangement includes a stem-loop transcriptional terminator (see Chapter 13). SAM—the ligand for this riboswitch—binds to the aptamer and stabilizes the secondary structure that includes this transcriptional terminator (as shown in the bottom part of Fig. 20-3a). Under these circumstances, transcription is terminated before polymerase has a chance to transcribe the downstream protein-coding segment of the gene. This form of transcriptional regulation is also called **attenuation**. (Note the mechanistic similarity to the *trp* system described in Box 20-1.) In another case—at another gene—a SAM-sensing riboswitch can work by regulating translation. In that case, as shown in Figure 20-3b, the alternative secondary structure stabilized by SAM binding to the aptamer includes a stem-loop that, although not a transcriptional terminator, does include the ribosome-binding site (RBS; within region 4). This conformational change sequesters the RBS and blocks ribosomes from initiating translation (see Chapter 15). This form of translation inhibition is thus essentially identical to that described for *trans*-acting sRNAs above (Fig. 20-1). The details of the changes in RNA secondary structure induced by SAM binding to a riboswitch are shown in Figure 20-4.

Many riboswitches have been identified, and current whole-genome sequencing results suggest there are probably many hundreds or thousands



**FIGURE 20-4** Changes in secondary structure of a SAM-sensing riboswitch. The sequences of regions 1–4 (described in Fig. 20-3) are here shown in detail and color-coded. The base pairing found in the two alternative secondary structures—that is, with and without SAM bound—are shown. (Adapted, with permission, from Winkler W.C. et al. 2003. *Nat. Struct. Biol.* 10: 701–707, Fig. 5b. © Macmillan.)

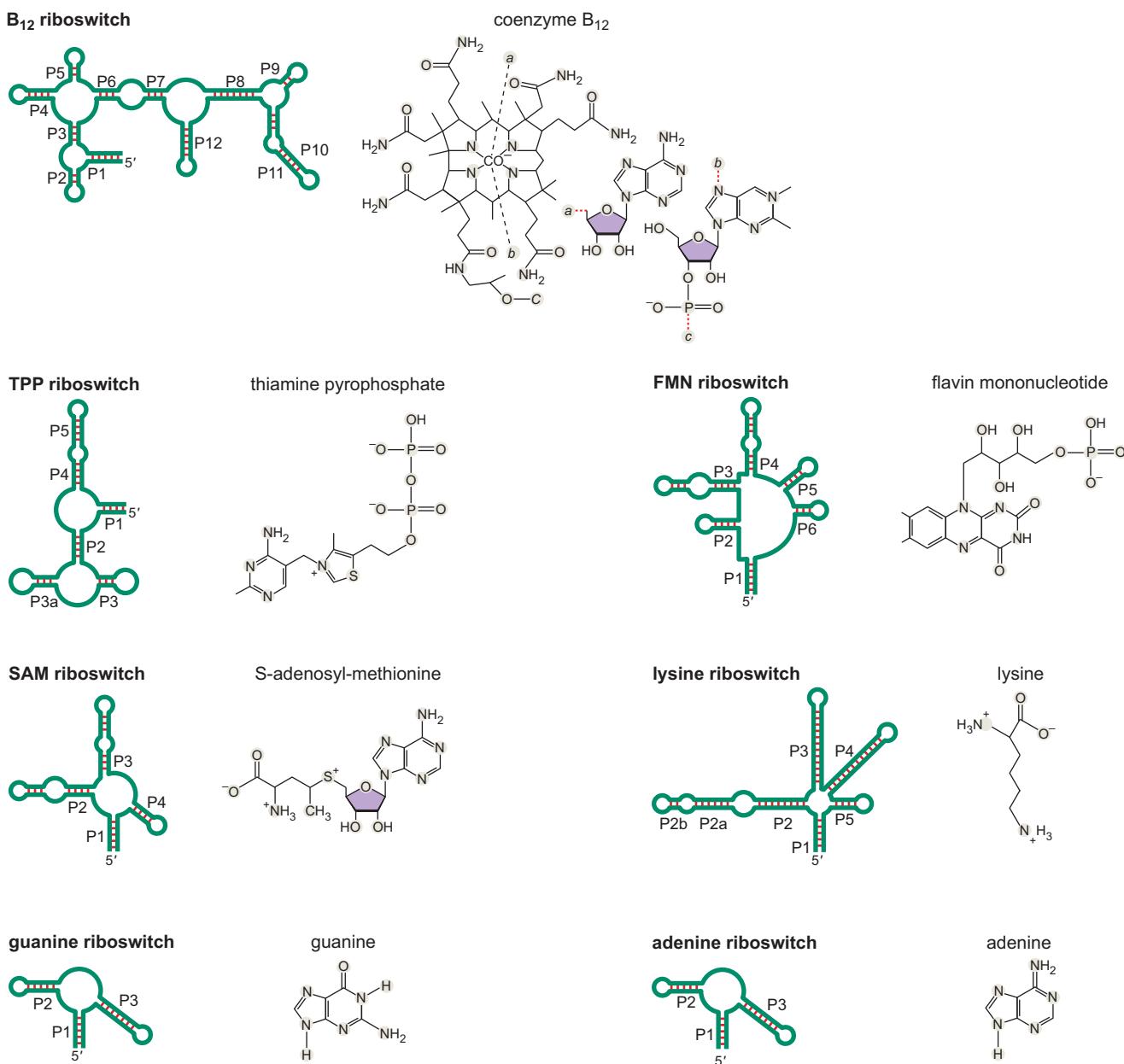
found across bacterial species. Even well-characterized examples respond to a range of different metabolites, including lysine and other amino acids, vitamin B12, coenzyme thiamine pyrophosphate (TPP), flavin mononucleotide (FMN), and guanine (Fig. 20-5).

Another kind of riboswitch responds to uncharged tRNAs, rather than to small-molecule ligands. Thus, certain genes, notably genes for aminoacyl-tRNA synthetases (see Chapter 15), are controlled by attenuation mediated by a 200- to 300-nucleotide-long, untranslated, leader RNA that directly and specifically interacts with the cognate, uncharged tRNA for the synthetase: the charged form of the tRNA does not fit in the binding pocket provided by the RNA secondary structure. Binding of the uncharged tRNA stabilizes the leader RNA in its antitermination structure so that transcription into the adjacent synthetase gene can proceed. Specificity is achieved in part by a “codon–anticodon” interaction between the tRNA and the leader RNA. Because uncharged (but not charged) tRNA can bind to the leader, transcriptional readthrough is only stimulated when the cognate amino acid is in short supply and the level of uncharged tRNA in the cell rises.

Although most prevalent in bacteria, riboswitches are found in other organisms as well, including archaea, fungi, and plants. In some cases in these higher organisms, riboswitches are even involved in controlling alternative splicing (Chapter 14). Thus, for example, in one case described in the fungus *Neurospora crassa*, three TPP aptamers were identified, two of which inhibited, and the third stimulated, expression of genes through regulation of RNA splicing.

### RNAs as Defense Agents in Prokaryotes and Archaea

Before we move on to consider the role of regulatory RNAs in eukaryotes, there is one more system to consider in bacteria. Although it is not strictly an example of gene regulation (it is a system of defense against viruses



**FIGURE 20-5** Riboswitches respond to a range of metabolites. The secondary structure of seven riboswitches and the metabolites they sense are shown here. (Adapted, with permission, from Mandal M. et al. 2003. *Cell* 113: 577–586, Fig. 7A. © Elsevier.)

and other extrachromosomal intruders), the mechanism used is strikingly similar to systems we will encounter in eukaryotes, namely, RNAi.

### CRISPRs Are a Record of Infections Survived and Resistance Gained

Years before any function could be assigned to them, particular stretches of unusual but characteristically organized sequence were noticed in the genomes of several bacteria. The distinctive pattern led to the rather cumbersome name (unmellowed by a helpful function) of Clustered Regularly Interspaced Short Palindromic Repeats (or **CRISPRs**). The generic features are

## ► ADVANCED CONCEPTS

**Box 20-1** Amino Acid Biosynthetic Operons Are Controlled by Attenuation

In *E. coli*, the five contiguous *trp* genes encode enzymes that synthesize the amino acid tryptophan. These genes are expressed efficiently only when tryptophan is limiting (Box 20-1 Fig. 1). The genes are controlled by a repressor, just as the *lac* genes are, although in this case, it is the *absence* of its ligand (tryptophan) that relieves repression.

Even after RNA polymerase has initiated a *trp* mRNA molecule, however, it does not always complete the full transcript. As with riboswitches, the decision to make a complete transcript is controlled by attenuation; in this case, most transcripts are terminated prematurely, before they include even the first *trp* gene (*trpE*). But attenuation is overcome if tryptophan levels are low in the cell; when tryptophan is limiting, polymerase does not terminate and instead transcribes all of the *trp* genes. Whether or not attenuation occurs depends on the ability of RNAs to form alternative secondary structures, just as it did with the riboswitches. In this case, however, the choice between alternative structures formed by the leader RNA is not controlled by binding of ligand directly to that RNA; instead, the choice of alternatives relies on the coupling of transcription and translation in bacteria.

The sequence of the 5' end of the *trp* operon mRNA includes a 161-nucleotide leader sequence upstream of the first codon of *trpE* (Box 20-1 Fig. 2). Near the end of this leader sequence, and before *trpE*, is a transcription terminator, composed of a characteristic hairpin loop in the RNA (made from sequences in regions 3 and 4 of Box 20-1 Fig. 2), followed by eight uridine residues (see Chapter 13, Fig. 13-13). Transcription usually stops after this terminator (and, we might have thought, should always stop), yielding a leader RNA 139 nucleotides long. This is the RNA product seen in the presence of high levels of tryptophan.

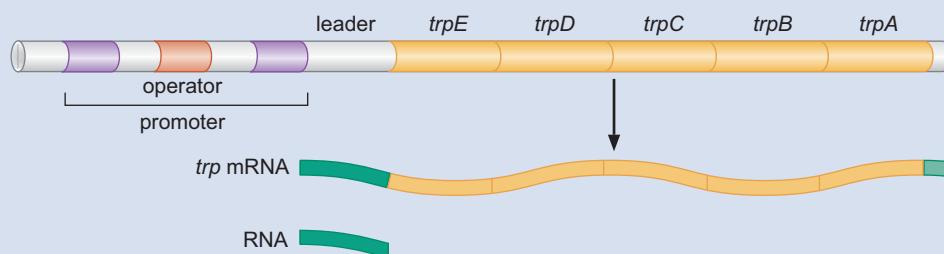
Three features of the leader sequence allow the terminator to be passed by RNA polymerase when the cellular concentration of tryptophan is low. First, there is a second hairpin (besides

the terminator hairpin) that can form between regions 1 and 2 of the leader (see Box 20-1 Fig. 2). Second, region 2 also is complementary to region 3; thus, yet another hairpin consisting of regions 2 and 3 can form, and when it does, it prevents the terminator hairpin (3, 4) from forming. Third, the leader RNA contains an open reading frame encoding a short “leader” peptide of 14 amino acids, and this open reading frame is preceded by a strong ribosome-binding site (see Box 20-1 Fig. 2).

The sequence encoding the leader peptide has a striking feature: two tryptophan codons in a row. When tryptophan is scarce, there is very little charged tryptophan tRNA available, and the ribosome stalls when it reaches the two tryptophan codons. Under these circumstances, RNA around the tryptophan codons is within the ribosome and cannot be part of a hairpin loop. The consequence of this is shown in Box 20-1 Figure 3.

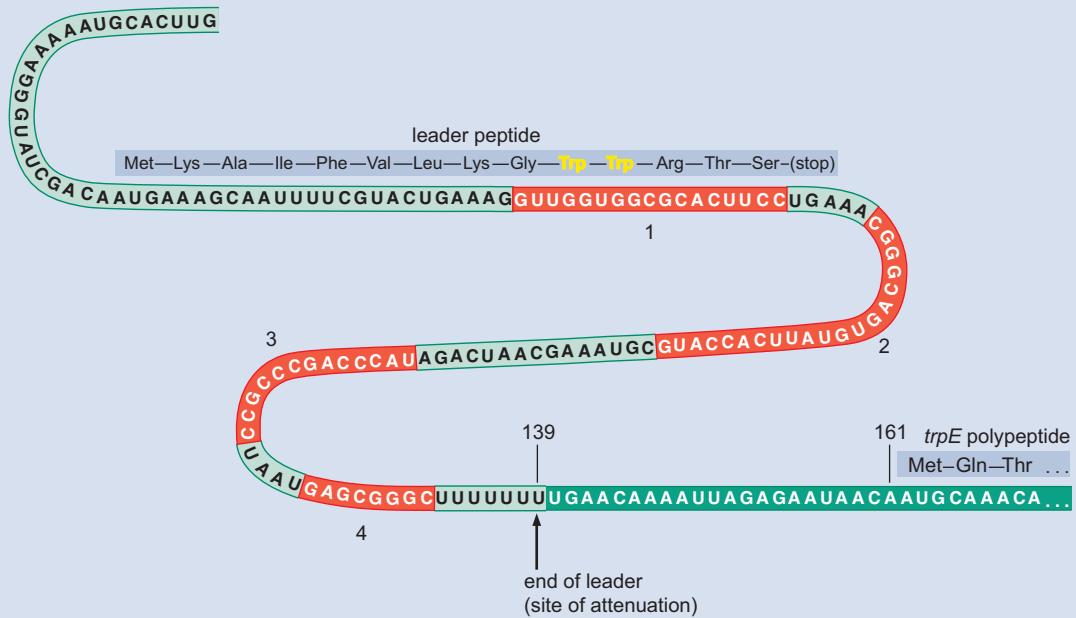
A ribosome caught at the tryptophan codons (part b) masks region 1, leaving region 2 free to pair with region 3; thus, the terminator hairpin (formed by regions 3 and 4) cannot be made, and transcription is not attenuated. If, on the other hand, there is enough tryptophan (and, therefore, enough charged Trp tRNA) for the ribosome to proceed through the tryptophan codons, the ribosome blocks sequence 2 by the time RNA containing regions 3 and 4 has been made. Thus, the terminator forms, attenuating transcription, and the *trp* genes are not transcribed.

The *trp* operon is controlled by repression and attenuation, providing a two-stage response to progressively more stringent tryptophan starvation. But attenuation alone can provide robust regulation: other amino acid operons such as *leu* and *his* rely entirely on attenuation for their control. In the case of the leucine operon, its leader peptide has four adjacent leucine codons, and the histidine operon leader peptide has seven histidine codons in a row.

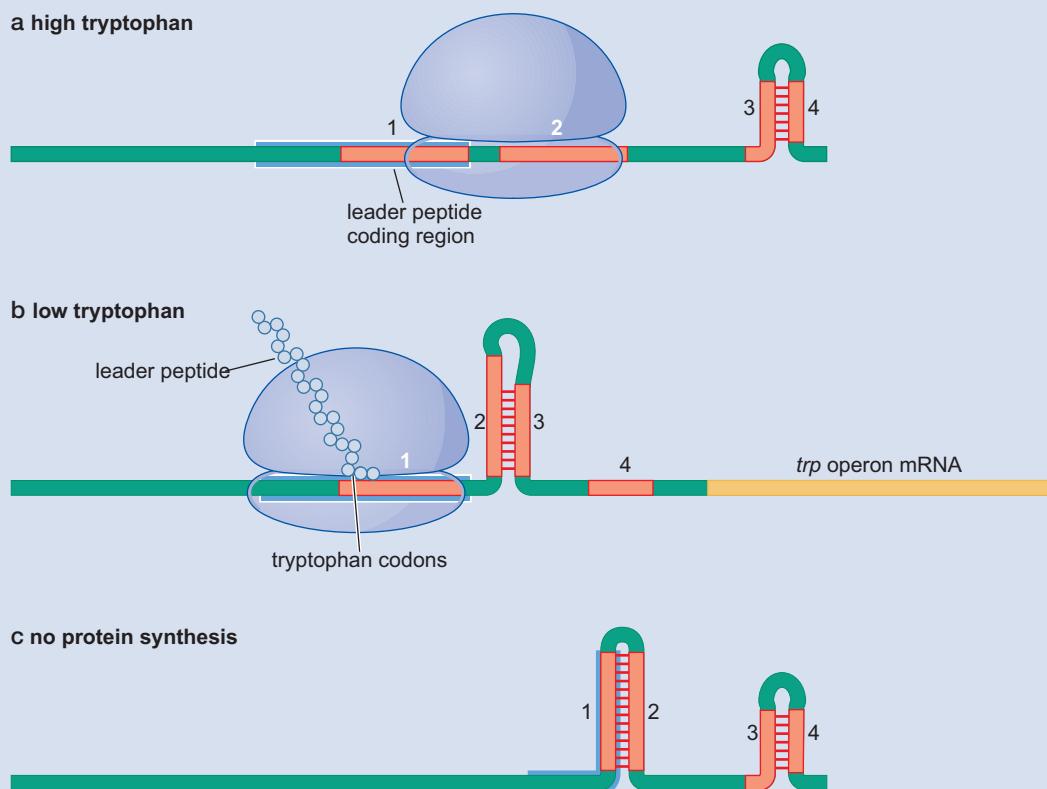


**BOX 20-1 FIGURE 1** The *trp* operon. The tryptophan operon of *E. coli*, showing the relationship of the leader (see text) to the structural genes that code for the Trp enzymes. The gene products are anthranilate synthetase (product of *trpE*), phosphoribosyl anthranilate transferase (*trpD*), phosphoribosyl anthranilate isomerase-indole glycerol phosphate synthetase (*trpC*), tryptophan synthetase  $\beta$  (*trpB*), and tryptophan synthetase  $\alpha$  (*trpA*).

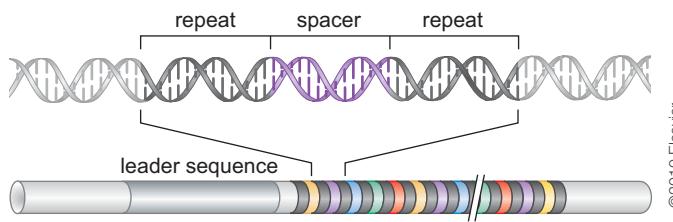
**Box 20-1** (Continued)



**BOX 20-1 FIGURE 2** *trp* operator leader RNA. Features of the nucleotide sequence of the *trp* operon leader RNA.



**BOX 20-1 FIGURE 3** Transcription termination at the *trp* attenuator. Transcription termination at the *trp* operon attenuator is controlled by the availability of tryptophan. The blue box shows the leader peptide-coding region. (a) Conditions of high tryptophan: sequence 3 can pair with sequence 4 to form the transcription termination hairpin. (b) Conditions of low tryptophan: the ribosome stalls at adjacent tryptophan codons, leaving sequence 2 free to pair with sequence 3, thereby preventing formation of the 3–4 termination hairpin. (c) No protein synthesis: if no ribosome begins translation of the leader peptide AUG, the hairpin forms by pairing of sequences 1 and 2, preventing formation of the 2–3 hairpin, and allowing formation of the hairpin at sequences 3–4. The Trp enzymes are not expressed.



**FIGURE 20-6** The organization of the CRISPR locus. The conserved repeat sequences and variable spacer sequences are shown at the top. Underneath is an array of such sequences (the number varies enormously); the proximal leader sequence is also shown. (Adapted, with permission, from Karginov F.V. and Hannon G.J. 2010. *Mol. Cell* 37: 7–19, Fig. 1A,B, p. 8. © Elsevier.)

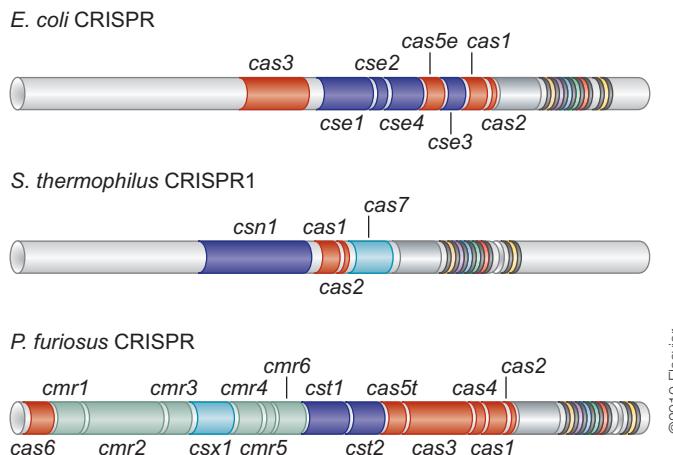
shown in Figure 20-6 and consist of **repeated sequences** (each ~30 bp long and highly conserved within a given cluster) interleaved with **spacer sequences** of similar length but highly divergent sequence. At one end of the array is a so-called leader sequence, often A-T rich and ~500 bp in length.

These clusters are not rare—indeed, CRISPRs have been found in half of all bacterial genomes sequenced, and essentially all genomes of Archaea. In many cases, there is only one cluster per genome, but not uncommonly there are more and the number can range up to 20 or more—and in one case almost 400 were detected in a species of *Chloroflexus*. But how do they arise, and what do they do?

The first clue to their origin came from the striking observation (a purely bioinformatics finding) that a significant number of the spacer sequences were identical to regions of known phage or plasmids. This quickly led to the proposal that these arrays are involved in some sort of defense mechanism against foreign nucleic acids entering the cell.

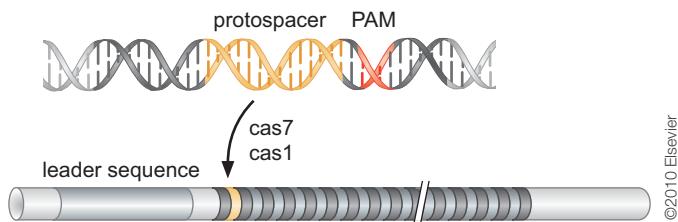
Experimental support for this model came when resistant bacterial cells that arose in populations challenged with a given phage were found to have incorporated spacer sequences derived from that phage. Likewise, reduced sensitivity to a phage could be conferred or revoked by addition or removal of relevant spacer sequences. Furthermore, bacteria were increasingly insensitive to infection by a given phage the more spacer sequences they acquired from that phage. It was also shown that viruses regain the ability to infect these previously resistant cells when they pick up mutations within those regions of their genomes that donated—and thus match—the spacer sequences.

A set of conserved protein-coding genes is tightly associated with the CRISPR sequences. The two most highly conserved members (*cas1* and *cas2*, for “CRISPR associated”) are found in all cases, but other *cas* genes, and more distantly conserved genes, are less so. These genes encode proteins involved in different aspects of CRISPR function, as we discuss later. An example of the complexity seen in a few real cases is shown in Figure 20-7.



**FIGURE 20-7** The organization of *cas* genes at three CRISPR loci. The varying numbers, orientations, and types of *cas* gene are shown at three well-studied CRISPR loci. The core *cas* genes are shown in red. The repeat and spacer sequences shown in Figure 20-6 are here at the right-hand end. (Adapted, with permission, from Karginov F.V. and Hannon G.J. 2010. *Mol. Cell* 37: 7–19, Fig. 1C, p. 8. © Elsevier.)

**FIGURE 20-8** The mechanism of spacer sequence acquisition. Each new spacer sequence is inserted next to the leader sequence, with the consequence that the array is a temporal record of acquisitions past. The sequence destined to become a spacer is, within the phage genome, known as a “proto-spacer” and lies adjacent to a PAM sequence as described in the text. (Adapted, with permission, from Karginov F.V. and Hannon G.J. 2010. *Mol. Cell* 37: 7–19, Fig. 2B, p. 10. © Elsevier.)



©2010 Elsevier

### Spacer Sequences Are Acquired from Infecting Viruses

As we outlined above, acquisition by a cell of spacer regions from a given phage confers decreased sensitivity to further infection by that phage. The basic process is shown in Figure 20-8. The sequence in the virus that will become a new spacer is called the proto-spacer and is found close to a PAM (proto-spacer adjacent motif) sequence. When a new spacer is added to a CRISPR array, it is incorporated at the proximal end, near the leader sequence.

Some of the *cas* genes encode proteins required for this acquisition process. Thus, the antiviral defense mechanism is not impaired by their absence, but the cell cannot acquire resistance to new viruses. The products of the *cas1*, *cas2*, and *cas4* genes fall into this category. Cas1 is a putative integrase, whereas Cas2 is a ribonuclease. In contrast, of other Cas proteins, Cas6 is involved in expression and processing of the CRISPR cluster and Cas3 in the interference of viral infection.

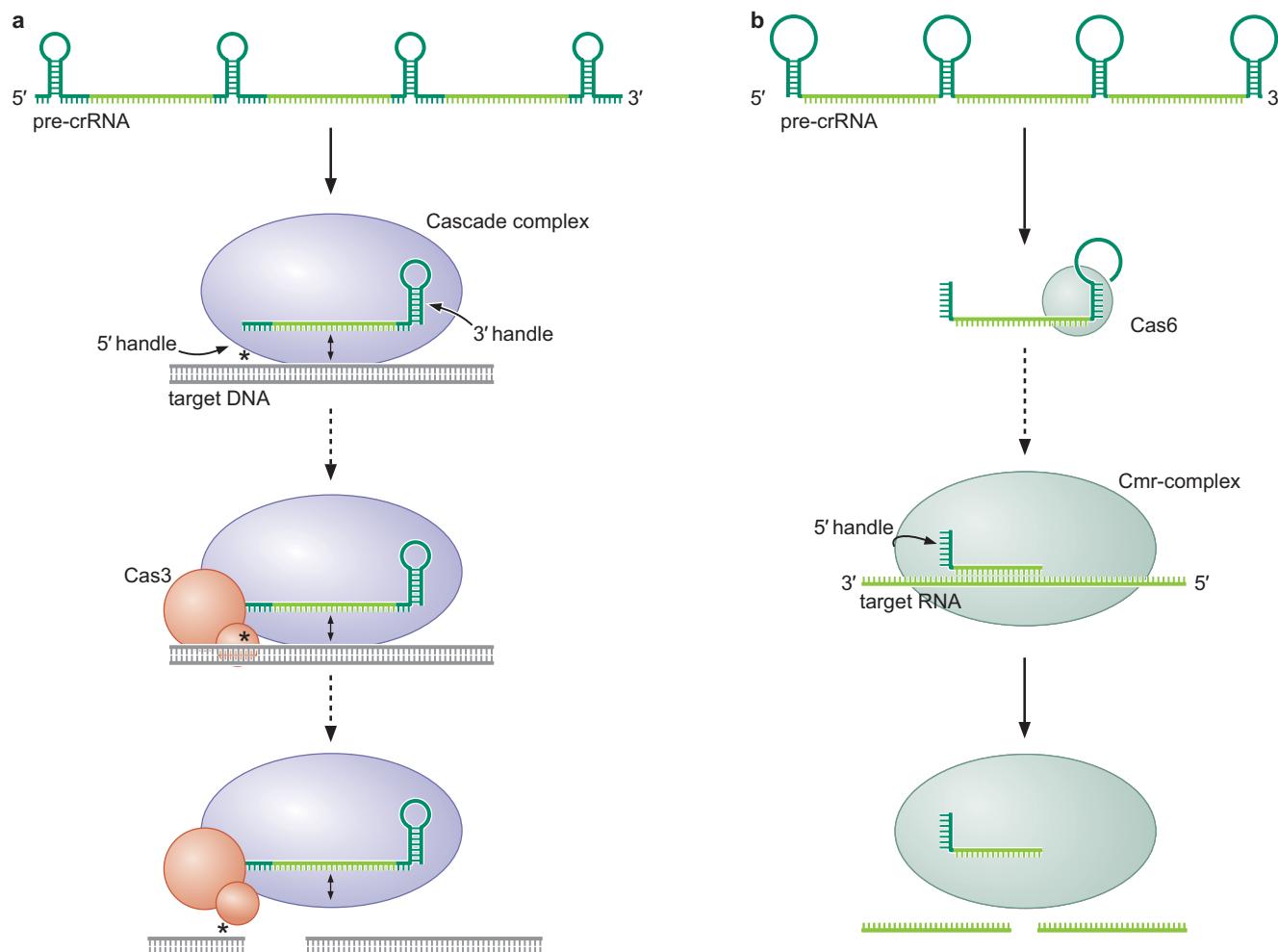
### A CRISPR Is Transcribed as a Single Long RNA, Which Is Then Processed into Shorter RNA Species That Target Destruction of Invading DNA or RNA

Expression of the *E. coli* CRISPR has been studied most extensively. The promoter from which expression is initiated is located within the leader region and generates a single RNA transcript called the pre-crRNA. In the case of *E. coli*, the CRISPR is associated with eight *cas* genes, the products of five of which form a complex called Cascade. This complex includes one subunit that is implicated in the processing of the long transcript into the individual short **crRNAs**, each the length of one spacer and one repeat sequence. These small RNAs remain bound to the Cascade complex and direct it to the DNA genomes of invading foreign DNA (Fig. 20-9a).

Each crRNA contains 8 nucleotides of the 5' repeat followed by the complete spacer region and most of the next repeat. The repeat sections included in the crRNA are called the 5' and 3' “handles,” respectively, and represent conserved parts of every crRNA, and are thus believed to be the regions that bind subunits of the Cascade complex.

In other cases (e.g., *Pyrococcus furiosus*), processing of the pre-crRNA is performed by a different (but structurally closely related) enzyme, and the crRNAs are bound by an alternative protein complex made up of a different collection of Cas proteins. In this case, the crRNAs start out with the 5' and 3' handles, but the 3' handle is trimmed off in a subsequent processing step. In this case, the crRNA–protein complex is directed against foreign RNA rather than DNA.

The mechanisms of the *E. coli* and *P. furiosus* systems are laid out in Figure 20-9. We shall see later how these resemble the RNAi process seen in eukaryotes, although in detail they operate quite differently.



**FIGURE 20-9** The antiviral operation of the CRISPR loci from *E. coli* and *P. furiosus*. The *E. coli* (a) system targets incoming DNA, whereas *P. furiosus* (b) targets RNA. Although similar in many ways, the processing mechanism and final operation of the two systems are different in striking ways as outlined in the text. In a, the pre-crRNA is processed by the CasE subunit of the Cascade complex (CasE is encoded by the *cse3* gene in Fig. 20-7). crRNA and Cascade are then directed to and cleave the target DNA with the help of Cas3 in ways not fully understood. In b, Cas6 (see Fig. 20-7) processes the pre-crRNA, and this, in complex with a complex distinct from Cascade, targets viral RNA in a mechanism strikingly analogous to the RNAi system in eukaryotes. (Adapted from Jore M.M. et al. 2012. *Cold Spring Harb. Perspect. Biol.* 4: a003657. © Cold Spring Harbor Laboratory Press.)

## REGULATORY RNAs ARE WIDESPREAD IN EUKARYOTES

Eukaryotic regulatory RNAs come in many flavors characterized by their size (“long” or “short”), their origin, and the mechanisms by which they are generated and regulate gene expression. It is now believed that between 30% and 70% of genes in higher eukaryotes are regulated to some extent by RNAs, with roles ranging from development (well-studied in the worm *C. elegans* and the plant *Arabidopsis*, described in Appendix 1) to cellular homeostasis and protection of the cell from viruses and transposons. Furthermore, one form of regulation (RNA interference) has been adapted for use as a powerful experimental tool to manipulate gene expression in many organisms.

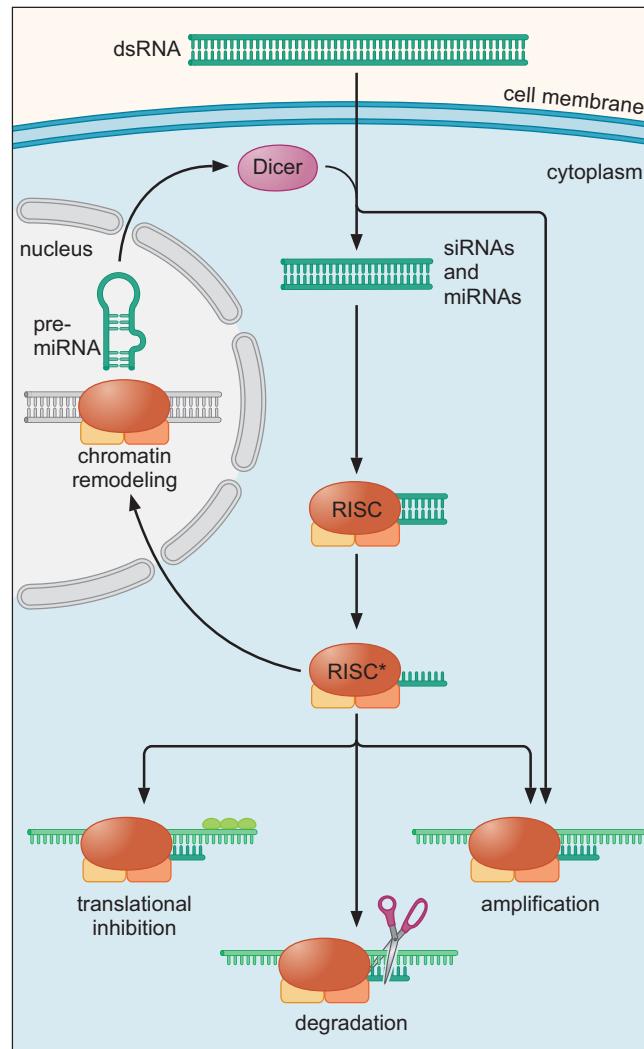
We start this section by looking at various short RNA regulators, and return to longer species at the end of the chapter.

Several types of very short RNAs repress—or silence—expression of genes with homology to those short RNAs. Depending on the origin and context, these RNAs act by inhibiting translation of the mRNA, destruction of the mRNA, or even by transcriptional silencing of the promoter that directs expression of that mRNA. As we shall describe later, these short RNAs are often generated by special enzymes from longer double-stranded RNAs (dsRNAs) of various origins.

### Short RNAs That Silence Genes Are Produced from a Variety of Sources and Direct the Silencing of Genes in Three Different Ways

Before describing aspects of the production and function of these short silencing RNAs in more detail, we first provide an overview of how this type of silencing works (illustrated in Fig. 20-10).

The small RNAs have different names depending on their origin. Those made artificially or produced *in vivo* from dsRNA precursors are typically called **small interfering RNAs (siRNAs)**. Another group of regulatory RNAs is the **microRNAs (miRNAs)**. These miRNAs are derived from pre-



**FIGURE 20-10** Generation of siRNAs and miRNAs, and their mode of action. Processing of dsRNA to make siRNAs and pre-miRNAs to make miRNAs by the enzyme Dicer. Another enzyme involved only in the generation of pre-miRNAs—Drosha—is not shown here but is described later. The siRNAs and miRNAs direct a complex called RISC (RNA-induced silencing complex) to repress genes in three ways. It attacks and digests mRNA that has homology; it interferes with translation of those mRNAs; or it directs chromatin-modifying enzymes to the promoters that direct expression of those mRNAs (Fig. 20-18). By recruiting an RNA-dependent RNA polymerase, siRNAs can generate more double-stranded RNA as fodder for Dicer to make more siRNA. This is the “amplification” step shown on the right and in more detail in Figure 20-11. (Adapted, with permission, from Hannon G.J. 2002. *Nature* 418: 244–251, Fig. 5. © Macmillan.)

cursor RNAs that are encoded by genes expressed in cells where those miRNAs have specific regulatory functions. A third class of short regulatory RNAs is the **piwi-interaction RNAs (piRNAs)**, which are expressed predominantly in the germline and have features distinct from miRNAs.

Both siRNAs and miRNAs are generated from longer RNA molecules by the enzyme **Dicer**, an RNase III–like enzyme that recognizes and digests longer dsRNAs or the stem-loop structures formed by miRNA precursors (see later discussion). The siRNA and miRNA products are typically 21–23 nucleotides long; their production is shown as the Dicer-stimulated step at the top of Figure 20-10. The piRNAs (which are 24–34 nucleotides long) are derived in a manner that does not involve a dsRNA precursor. Instead, the piRNAs are generated by processing long single-stranded transcripts covering so-called piRNA clusters found in the genome. This processing does not require Dicer.

These small RNAs inhibit expression of homologous target genes in three ways: they trigger destruction of the mRNA encoded by the target gene, they inhibit translation of the mRNA, or they induce chromatin modifications within the target gene and thereby silence its transcription. Remarkably, whichever route is used in any given case, much of the same machinery is required. This machinery includes a complex called the **RNA-induced silencing complex (RISC)**. A RISC contains, in addition to the small RNA, various proteins including a member of the **Argonaute** family.

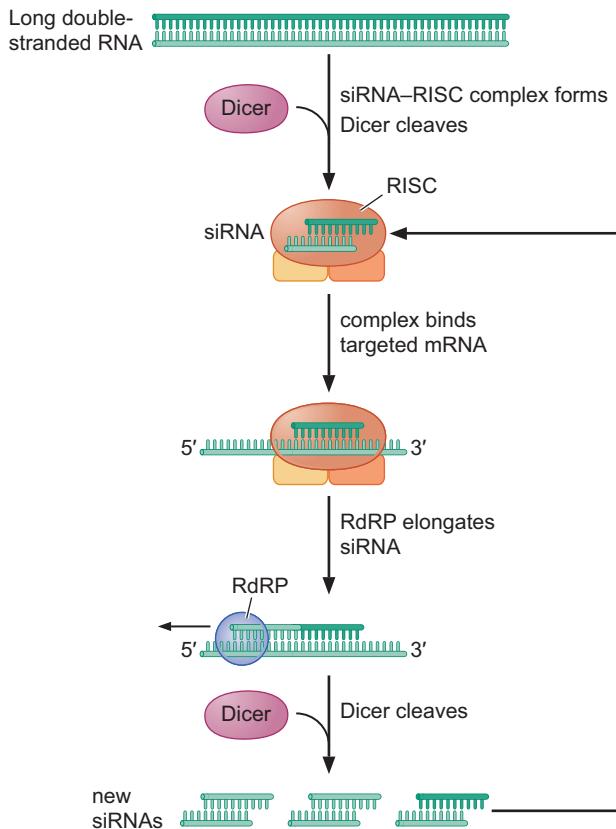
The small RNA must be denatured to give a **guide RNA**—the strand that gives the RISC specificity, as we shall see—and a **passenger RNA**, which usually gets discarded. The resulting complex, the mature RISC, is then directed to target RNAs containing sequences complementary to the guide RNA. These target RNAs are degraded or their translation is inhibited. Typically, the choice depends in part on how closely the guide RNA matches the target mRNA: if the sequences are highly complementary (as is usually the case with siRNAs), the target is degraded; if the match is not as good (i.e., if there are several base-pairing mismatches, as is often the case with miRNAs), the response is more often inhibition of translation. In those cases in which the target RNA is degraded, Argonaute is the catalytic subunit that performs the initial mRNA cleavage; for this reason, Argonaute is often called **Slicer** and mRNA cleavage is called slicing.

A RISC can also be directed into the nucleus, where it recruits other proteins that modify the chromatin around the promoter of the gene complementary to the guide RNA (shown on the left of Fig. 20-10). This modification leads to silencing of transcription (Chapter 19). Establishing silencing in the centromeric regions of the yeast *Schizosaccharomyces pombe*, for example, requires the RNAi machinery, as we shall see later.

A distinction worth making between miRNAs and siRNAs is that the former act like traditional *trans*-acting regulators: they are encoded by a gene but act on other genes (like the sRNAs we encountered in bacterial systems). In contrast, siRNAs are typically generated by transcripts of the regions on which they act (formally like the antisense RNAs we described in bacteria) and are thus described as working in *cis*.

Another feature of RNAi silencing is worth noting—its extreme efficiency. Thus, very small amounts of dsRNA are often enough to induce a near complete shutdown of target gene expression. A factor adding to efficiency could be the action of an **RNA-dependent RNA polymerase (RdRP)**, an additional enzyme required in many cases of RNAi including centromeric silencing in fission yeast. This polymerase can amplify the inhibitory signal: the RdRP generates dsRNA after recruitment to the mRNA by the original siRNA (as indicated on the right of Fig. 20-10 and shown in detail in Fig. 20-11). This feedback process generates large

**FIGURE 20-11** Amplification of the siRNA signal by RdRP. As shown on the right in Figure 20-10, the siRNA signal can be amplified, generating more dsRNA for Dicer to process into more siRNAs. This is achieved because the siRNA–RISC complex can recruit an enzyme, RNA-dependent RNA polymerase to the targeted RNA, and the siRNA acts as a primer for that enzyme to transform the target into dsRNA, which can itself then be acted on by Dicer. RdRPs are found in plants, worms, and the yeast *Schizosaccharomyces pombe* (*Saccharomyces cerevisiae* does not have the RNAi machinery at all), and we will see the importance of this amplification step in the case of centromere silencing in *S. pombe* (Fig. 20-18).



amounts of siRNA. RdRP has not yet been identified in mammalian cells, and high efficiency there likely results from the fact that slicing is catalytic—that is, each RISC can cleave several mRNAs.

Thus, although in the first section of this chapter we saw examples of small RNAs regulating gene expression in bacteria, the mechanisms of both production and action of such RNAs in eukaryotes is very different.

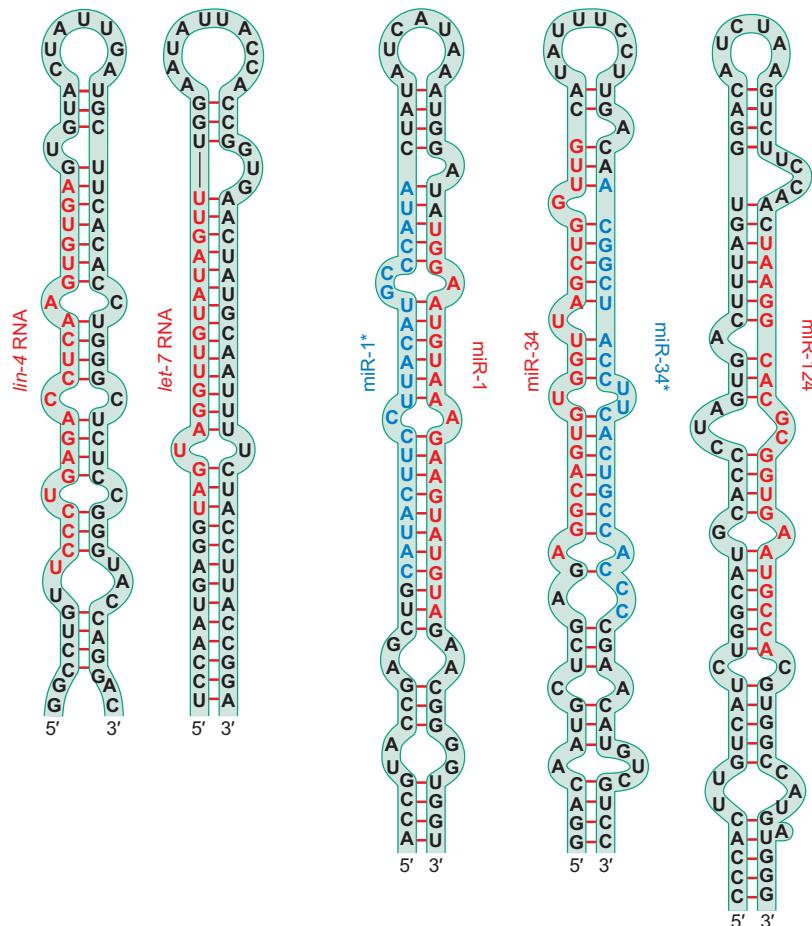
## SYNTHESIS AND FUNCTION OF miRNA MOLECULES

### miRNAs Have a Characteristic Structure That Assists in Identifying Them and Their Target Genes

As mentioned above, miRNAs are encoded in the genome as segments of longer transcripts. Their characteristic structure helps identify them and predict the target genes they might regulate.

The functional form of an miRNA is typically ~21 or 22 nucleotides (it can vary from 19 to 25 nucleotides). These short RNAs are generated by two RNA cleavage reactions from a longer RNA transcript (called a pri-miRNA, for “primary”) that carries a hairpin-shaped secondary structure. The first cleavage liberates the stem-loop, called the pre-miRNA; the second generates the mature miRNA from the pre-miRNA. One of the first identified, and best-characterized, miRNAs is *let-7*, which regulates development at the larval-to-adult transition in the worm *C. elegans* (see Appendix 1). The structures of the pre-miRNAs for *let-7* and some other naturally occurring miRNAs are shown in Figure 20-12.

It was thought initially that one “arm” of pre-miRNA stem-loop structure would be the regulatory miRNA. But numerous examples have been identi-



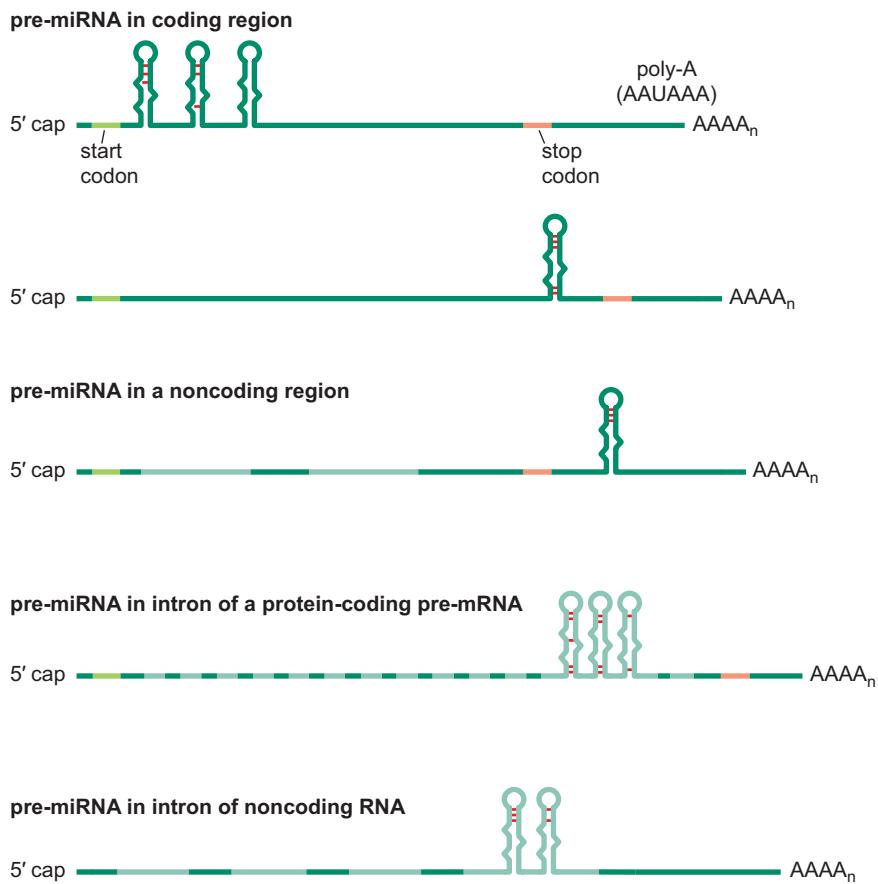
**FIGURE 20-12** Structure of some pre-miRNAs before processing to generate the mature miRNAs. The sequences in red are miRNAs. In some cases, both “arms” of a stem-loop can generate a functional miRNA. In such cases, the second miRNA is shown in blue—for example, miR-1 (red) and miR-1\* (blue), as well as with miR-34 (red) and miR34\* (blue). The miRNAs shown are all from the worm. *lin-4* and *let-7* were identified genetically; those called miR were found by bioinformatics. (Modified, with permission, from Lim L.P. et al. 2003. *Genes Dev.* **17**: 991, Fig. 6. © Cold Spring Harbor Laboratory Press.)

fied in which both arms of the structure give rise to functional miRNAs, each with its own set of target genes (in these cases, the two miRNAs are red and blue in Fig. 20-12). It now appears that having miRNAs produced from both arms is common. The pre-miRNAs can be encoded by any part of a transcript: that is, they might fall within coding regions, within leader regions, or within introns (Fig. 20-13).

The distinctive secondary structure of a primary transcript carrying an miRNA (pri-miRNA) has made it possible to predict their presence based on the calculated secondary-structure fold of the RNA sequence. Furthermore, in many cases, candidates for the regulated target genes can also be predicted, because silencing depends on sequence complementarity between the target and the mature miRNA. The base pairing between miRNA and target RNA is initiated by interactions of so-called seed residues—typically the sequence between bases 2 and 9 of the 22-nucleotide miRNA. This is the region of highest complementarity, and thus it is the region most useful in identifying candidate target genes. Of course, establishing that an miRNA really exists requires that its presence be detected in cells (e.g., by northern blotting) and that gene expression from target mRNAs be affected by its presence.

The two cleavage reactions required to generate the miRNA from these primary transcripts are mediated by two distinct RNases. One is Dicer, which we have already introduced and is required in the generation of siRNAs as well. The other, specifically required for miRNA processing, is **Drosha**. A characteristic of both these enzymes is that they recognize and

**FIGURE 20-13** miRNAs are coded in both introns and exons in RNA. Intronic sequences are shown in light green. Start and stop codons are indicated by lime green and pink, respectively.



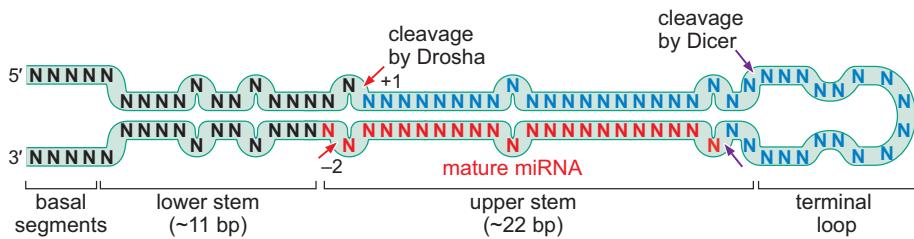
cleave RNAs on the basis of the structure of their substrates rather than their specific sequence. We now turn our attention to how these enzymes work.

### An Active miRNA Is Generated through a Two-Step Nucleolytic Processing

The first step is carried out by the enzyme Drosha, a member of the RNase III family of enzymes. Drosha makes two cleavages that cut the stem-loop region of the RNA (pre-miRNA) out of the primary transcript RNA (pri-miRNA). This enzyme works together with an essential specificity subunit protein (called Pasha in some organisms and DGCR8 in others), and together these two proteins form an active **Microprocessor complex**. The pre-miRNA generated by Drosha is usually ~65–70 nucleotides long. Drosha resides in the cell's nucleus, and the Drosha-catalyzed cleavage event occurs in that cellular compartment.

The base-paired stem in the pri-miRNA is typically ~33 bp in length (three helical turns of dsRNA) and contains only a few mismatches (Fig. 20-14). At the “top” of the stem is a loop of variable size (usually relatively large, ~10 nucleotides); the sequence of this loop region is not critical for the processing reactions. Importantly, for processing by Drosha, single-stranded RNA (ssRNA), lacking significant secondary structure, is needed flanking each side (5' and 3') of the stem-loop. It is the ssRNA–dsRNA junctions that are in large part responsible for determining the cleavage specificity of Drosha.

The stem region can be divided into two functional segments: an ~11-bp lower stem and an ~22-bp upper stem (Fig. 20-15). Drosha cleaves 11 bp



**FIGURE 20-14** Overview of the structure of pri-RNA showing Dicer and Drosha cleavage sites. The region in red becomes the mature miRNA. Note that the basal segments must be single-stranded for proper recognition by the Drosha complex.

away from the dsRNA–ssRNA junctions—that is, between the lower and upper stems in the pri-miRNA. The two cleavages thus generate the ~65-nucleotide pre-miRNA composed of the 22 bp (two helical turns) of dsRNA and the top loop. The RNase III family enzymes are specific for dsRNA and cleave it in a manner that leaves a 2-nucleotide overhang on the 3' ends of the dsRNA product. This 3' overhang is important for recognition of that RNA molecule by the next enzyme in the pathway, Dicer.

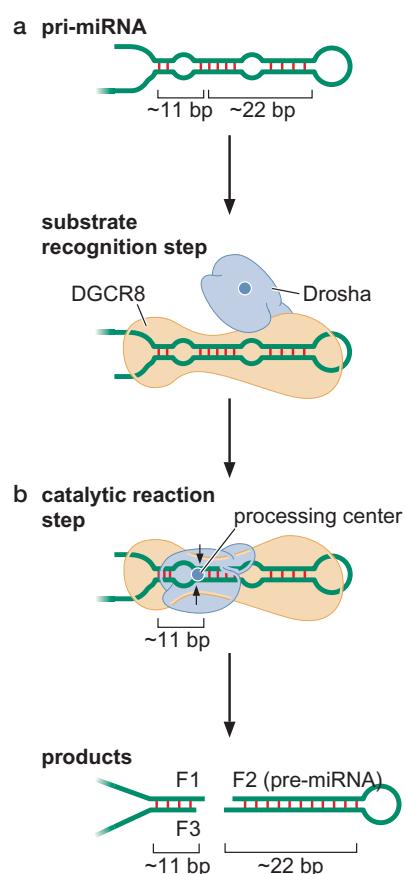
### Dicer Is the Second RNA-Cleaving Enzyme Involved in miRNA Production and the Only One Needed for siRNA Production

The pre-miRNA liberated by Drosha is exported to the cytoplasm, where the second RNA cleavage reaction, performed by Dicer, takes place. As with Drosha, Dicer selects its cleavage sites using a measuring, rather than sequence-specific, mechanism. A high-resolution structure of Dicer provides insight into how this likely occurs.

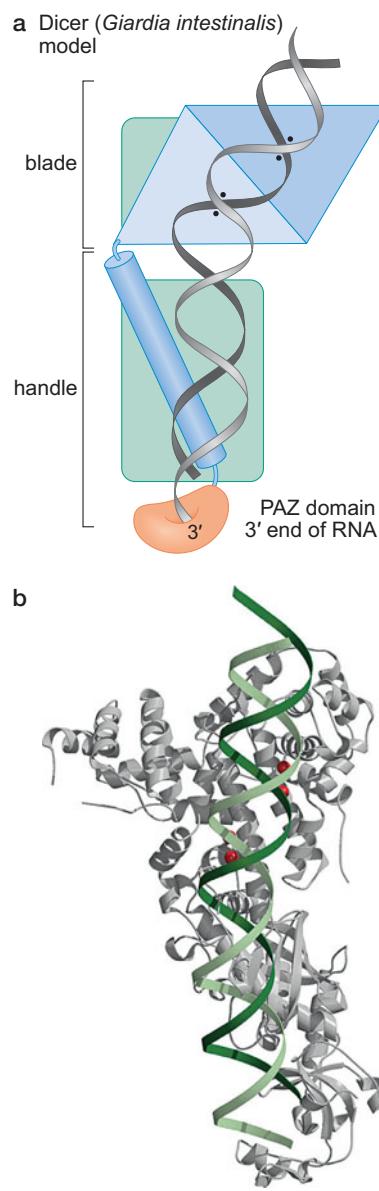
Dicer is constructed of three modules: two RNase III domains and a dsRNA-binding domain called the PAZ domain (named for three proteins that contain this domain: Piwi, Argonaute, and Zwille).

Figure 20-16a shows in cartoon form the organization of the Dicer protein and how it is believed to interact with a dsRNA molecule. In the bottom panel, Figure 20-16b, is shown the structure of Dicer, modeled with a substrate RNA. The protein is shaped like a hatchet. The PAZ domain is at the bottom of the handle, where it forms a binding pocket for the 3' end of the dsRNA substrate. The handle of the hatchet is formed by a linker domain and contains a positively charged binding surface for the RNA molecule. The top “blade” region comprises the two RNase domains, arranged in a symmetrical dimer. Each RNase domain carries an active site and is responsible for cleaving one of the two strands of the substrate RNA. Thus, Dicer will act on any dsRNA, regardless of sequence, and will cleave this molecule 22 nucleotides from its end. The PAZ domain anchors the 3' terminus of the substrate RNA to position the active sites of the enzyme ~22 nucleotides away in a ruler-like fashion (see Fig. 20-16). Indeed, the occurrence of differently sized PAZ domains correlates with the different sizes of Dicer products found in different organisms.

As we have seen, only miRNAs are made from large hairpin precursors. In contrast, the precursor RNA for the siRNA pathway is a longer dsRNA. As a consequence of this different initial substrate, Drosha is not needed for the generation of siRNAs. Cleavage by Dicer is still required, however, and again generates a suitable 21- to 22-nucleotide RNA for incorporation into RISC.



**FIGURE 20-15** Recognition and cleavage of pri-miRNA by the Microprocessor complex. Three fragments are generated by cleavage, labeled F1, F2 (the pre-miRNA), and F3.



**FIGURE 20-16** Dicer structure and organization. (a) The scheme shows Dicer organization. (b) Dicer structure modeled with dsRNA reveals how length is measured. The protein is shown in gray, with nuclease active sites indicated by the red spheres (and as black dots in part a). The RNA is in green. The structure shown contains only the RNase III and PAZ domains. The Dicer protein also contains ATPase and other domains. (b, MacRae I.J. et al. 2006. *Science* **311**: 195–198. PDB Code: 2FFL; note that the RNA was modeled into the structure and was not part of the crystal structure.) Image prepared with MolScript, BobScript, and Raster3D.

In plants, even miRNAs are generated by Dicer alone; it is not clear how they manage to forgo prior action of Drosha.

## SILENCING GENE EXPRESSION BY SMALL RNAs

We have so far seen how the small RNAs are generated, from double-stranded RNA or miRNA precursors. We now turn to how these small RNAs silence expression of their target genes.

### Incorporation of a Guide Strand RNA into RISC Makes the Mature Complex That Is Ready to Silence Gene Expression

The action of Dicer generates the short RNA molecule that will determine what target RNA is acted upon. The active form of the regulatory RNA is the single-strand form—at this stage called the **guide RNA**—incorporated into a RISC protein complex. Within this complex, the RNA guide strand recruits RISC to a target RNA. It has been argued that the length of ~22 nucleotides is just long enough to specify a single target sequence uniquely in the large genomes of complex eukaryotes using RNA–RNA base pairing.

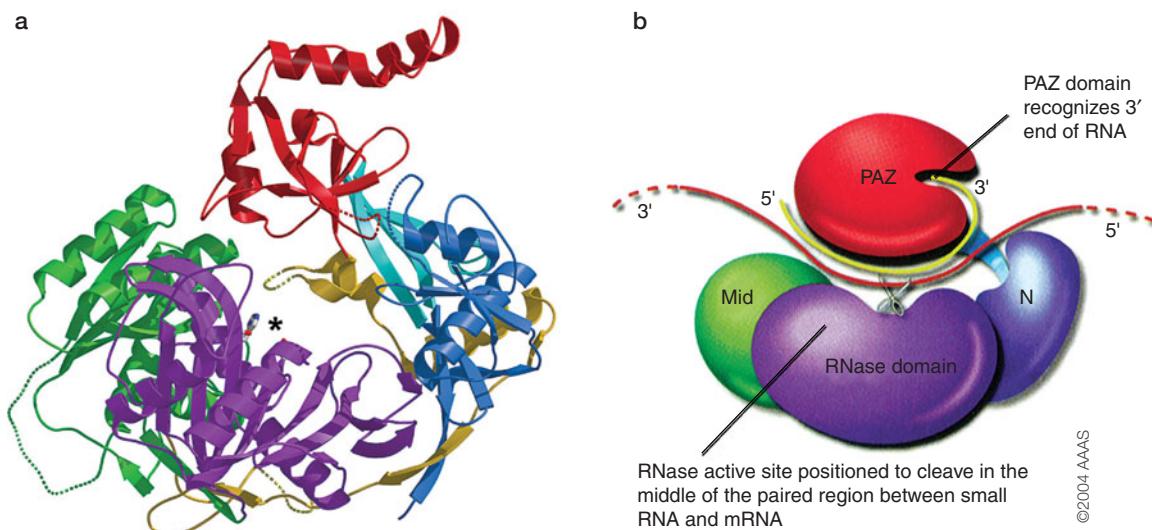
The central component of RISC is a protein called **Argonaute**, which is in many cases an RNA-cleaving enzyme. The best understood mechanism of gene silencing is RISC-mediated cleavage—or slicing—of the target mRNA. However, many organisms have multiple members of the Argonaute protein family. For example, there are eight distinct Argonautes in humans, but not all of these Argonautes, when incorporated into a RISC complex, have slicer activity. RISCs containing other Argonautes must silence gene expression using non-slicer-dependent mechanisms, such as repression of translation. The piRNAs we encountered above, which are not generated by Dicer, nonetheless bind an Argonaut-related protein called PIWI within a RISC-like complex.

Generation of the active RISC and slicing, under the guidance of an siRNA, occurs as follows. The short dsRNA generated by Dicer is incorporated into RISC, where it is denatured to provide the guide strand and the passenger strand (which is discarded). The resulting RISC—called mature RISC—with its single-stranded guide RNA is now ready to recognize and slice the target mRNA.

As we saw with Dicer, the structure of an Argonaute protein provides a framework for understanding the mechanism of target RNA recognition and cleavage by RISC (Fig. 20-17; see also Structural Tutorial 20-1). Like Dicer, Argonaute has both a PAZ domain and an RNase domain. The PAZ domain specifically recognizes the 3' end of the guide RNA. The bound guide RNA is base-paired to the target RNA, and the architecture of the complex is such that this binding positions the active site of the RNase domain appropriately to cleave the target RNA strand. Cleavage occurs nearly in the middle of the guide RNA–target RNA duplex, between the 10th and 11th nucleotides from the 5' end of the guide RNA.

As we have already mentioned, in some cases, mature RISC can inhibit translation rather than slicing mRNA, and, indeed, this is how miRNAs are thought most commonly to work.

The mechanism of this translational repression is still under scrutiny, and there is much debate over the order of events. Thus, although translation is certainly inhibited, the mRNA also decays. It is therefore difficult to prove



**FIGURE 20-17** Argonaute structure, showing RNA-binding regions and an RNase H-like nuclease domain. (a) Crystal structure of Argonaute. The domains are colored as in part b, with the blue domain being the amino-terminal part of the protein, and the green domain in the middle. (b) Cartoon of the Argonaute domains. The arrow shows the RNase active site positioned to cleave in the middle of the paired region between small RNA and miRNA. (a,b, Adapted, with permission, from Song J.J. et al. 2004. *Science* 305: 1434–1437, Fig. 4C. PDB Code: Iu04. © AAAS.) Images prepared with MolScript, BobScript, and Raster3D.

which is the result of the other: if mRNA decays, there will, of course, be no translation; likewise, when translation of an mRNA is inhibited (by any mechanism), the cell tends to destroy that mRNA molecule. Either way, this is distinct from the active slicing mechanism triggered by siRNAs, as described above.

Translation initiation is an elaborate process involving a lot of factors (see Chapter 15) and affording many opportunities for interference. Whatever the mechanism of translation inhibition, it appears that miRNAs lead, in some cases, to the sequestration of mRNA in so-called processing bodies (P-bodies) within the cytoplasm where translation is repressed.

### Small RNAs Can Transcriptionally Silence Genes by Directing Chromatin Modification

We have now seen how miRNAs and siRNAs can silence genes by inhibiting the translation of target mRNAs or directing their destruction. Regulatory RNAs can also act at the level of transcription, switching off expression of target genes by directing histone modification of the promoter. This mechanism has been most extensively studied in centromeric silencing in the fission yeast *S. pombe*.

We noted in Chapter 19 that in yeast, genes placed in certain regions of the genome are typically silenced. In the case we described in detail in that chapter, genes placed near the telomeres in *Saccharomyces cerevisiae* were silenced. Genes in the mating-type locus of that yeast and of *S. pombe* are also silenced. In *S. pombe*, the centromeres are another silenced region of the genome. In both organisms, silencing involves histone modifications. But unlike cases of silencing in *S. cerevisiae*, which lacks the RNAi machinery, centromeric silencing in *S. pombe* requires that pathway.

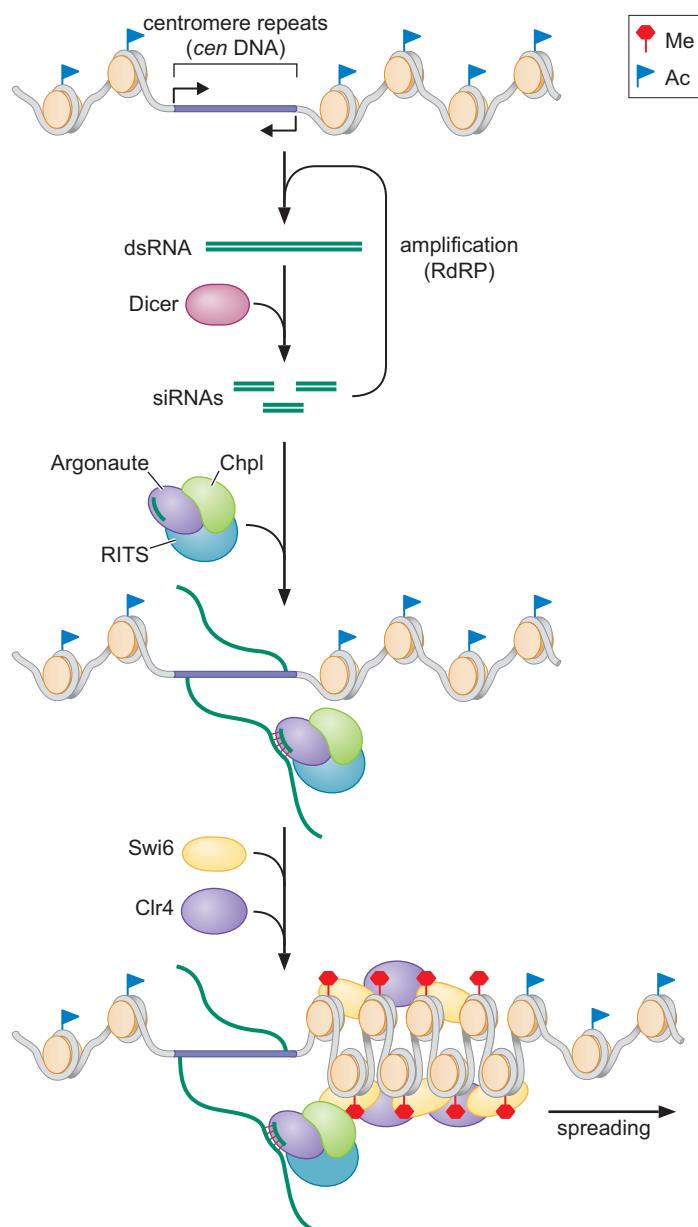
The centromeres of *S. pombe* have a sequence organization more like that of higher eukaryotes (e.g., flies and humans) than that of *S. cerevisiae* (see Chapter 8, Fig. 8-8). Each centromere has a central region, of largely unique sequence, flanked by a series of repeats common to all centromeres. The repeats are important to function and contribute to the formation of heterochromatin and the transcriptional silencing associated with the region, as we shall see later. Histones within the heterochromatin carry repressing markers: low levels of acetylation and methylation on lysine 9 of the histone H3 tail (H3K9).

*S. pombe* has only a single gene for each of the major components of the RNAi pathway—Dicer and Argonaute. Higher organisms have multiple Dicer and Argonaute genes with partially redundant functions, making genetic manipulation of the pathway more difficult. In addition, unlike the situation in flies and worms, loss of the RNAi pathway is not lethal to *S. pombe*, although it does make the cells grow poorly by, for example, disrupting chromosome segregation. It was a surprise, however, to discover that loss of any component of the RNAi pathway led to loss of histone H3K9 methylation and loss of gene silencing at the centromeres, particularly because this silencing was known to be transcriptional. Until this discovery, RNAi had been thought to act only post-transcriptionally.

The key to understanding this transcriptional silencing seems to be the centromeric repeats themselves; these sequence elements are transcribed from both strands by RNA polymerase II, producing complementary transcripts that can hybridize to form dsRNAs—a process that is amplified by RdRP (as we saw in Fig. 20-11). The RNAs are, in turn, acted on by the RNAi machinery to generate siRNAs that somehow—and quite how remains unclear—direct an Argonaute-containing RISC-like complex (called RNA-induced transcriptional silencing [RITS] complex) to the centromeres. The siRNAs could in theory do this by recognizing DNA at the centromeres, through sequence-specific base pairing directly with the DNA template. But more likely are models in which the siRNAs recruit RITS to transcripts tethered to the centromere by RNA polymerase II. Recruitment results in slicing of centromic transcripts, which is, in turn, required for spreading of the histone modification apparatus along the centromere (Fig. 20-18). Thus transcription itself may spread silencing when transcripts are targeted by RNAi.

As we mentioned above, the mating-type loci of *S. pombe* are also transcriptionally silenced, and here the silencing is not lost in mutant strains defective for RNAi. It is believed that RNAi acts in this case as well, but only in initially establishing the silenced state—it is not required for maintaining silencing once it is established. Other, protein-based mechanisms sustain the repressed state—just as they do in *S. cerevisiae* (Chapter 19). RNAi is also believed to play a part in heterochromatin silencing in other organisms, ranging from flies to plants. Silencing of unwanted transcription from transposons also appears to be RNA-mediated, as we describe presently.

A fascinating series of observations and experiments led to our current understanding of small regulatory RNAs in eukaryotes. These began in the late 1980s with the seemingly mystifying results of attempts to overexpress pigment genes in petunias (to make them a deeper purple, but ending up with the flowers turning white). Next was the surprising discovery of regulatory genes from worms whose products turned out to be miRNAs, and then experiments showing that introducing dsRNAs into worms silenced complementary gene expression. This story is described in Box 20-2, Discovery of miRNAs and RNAi.



**FIGURE 20-18** A model for RITS recruitment and the silencing of centromeres. Shown at the top are nucleosomes around the repeat sequences at a centromere in *S. pombe*. The repeat sequences (cen DNA) are transcribed by RNA polymerase II, generating dsRNA that is a substrate for Dicer. The siRNAs thereby produced are loaded into the Argonaute-containing complex RITS. As shown in the middle, the siRNA-containing RITS is then recruited to the Pol II–tethered transcripts being generated by continued transcription of the centromeric repeats, through complementarity between the siRNA and the transcript. This complex then recruits a number of other complexes: RDRC, which allows production of further dsRNA by RdRP (see Fig. 20-11), and Clr4 and Swi6, which locally modify nucleosomes by adding the H3K9 silencing markers. Another subunit of RITS, Chp1, contains a chromodomain (Chapter 8, Fig. 8-41), which, by binding to the methylated nucleosomes, likely stabilizes the binding of RITS. Although not shown in the figure, “slicing” of the transcripts by Argonaute (within RITS) generates substrate RNAs for the RdRP, which synthesizes a complementary strand and thus generates further substrate for Dicer. This process is required for the nucleosome modification—and thus the region silenced—to spread. (Redrawn, with permission, from Martienssen R. and Moazed D. 2007. *Epigenetics* [ed. D. Allis et al.], p. 157, Fig. 4. © Cold Spring Harbor Laboratory Press.)

### RNAi Is a Defense Mechanism That Protects against Viruses and Transposons

The RNAi machinery is widespread in eukaryotes, although not ubiquitous. It does not occur in *S. cerevisiae*, for example, as we just noted. It is believed, however, that at least the basic system existed in the most recent common ancestor to all eukaryotes but was subsequently lost in some lineages.

But what does RNAi do, biologically? There are miRNAs, of course—and the RNAi machinery is required to produce and use those regulators—but some organisms have the RNAi machinery and no miRNAs (including *S. pombe*). It is, in fact, believed that miRNAs evolved to take advantage of the existence of the RNAi machinery rather than being the reason that machinery exists. One ancient function the RNAi machinery might have served (and still serves) is protecting organisms from transposons and viruses.

## ► KEY EXPERIMENTS

### Box 20-2 Discovery of miRNAs and RNAi

In 1989, Richard Jorgensen, working at the biotech company Advanced Genetic Sciences in Oakland, California, was trying to make petunia plants with more deeply purple flowers than existing strains. The strategy seemed straightforward: he would introduce into the plants an additional copy of the pigment gene (encoding chalcone synthase) under the control of a strong promoter. These plants would make more chalcone synthase and the flowers would be more purple. What he actually got were plants with varying degrees of paler flowers, many that were sectored—with purple and white regions—and even some that were completely white (Box 20-2 Fig. 1).

Although disappointing, these results were intriguing. In attempting to understand what was going on, Jorgensen uncovered various features of the phenomenon, called **cosuppression** (because expression of both the transgene and the endogenous gene is repressed). The greater the expression of the transgene, the lower is the level of chalcone synthase; this was true whether increased expression resulted from multiple copies of the transgene or from use of stronger promoters driving the transgene. It was also noted that some plants had variegated patterns of pigmentation, and that different variegation patterns could be found in different flowers on the same plant. These patterns were sometimes inherited, but on other occasions apparently altered at random. These observations suggested to Jorgensen and others (particularly Marjori Matzke, who also was investigating this phenomenon) that they were dealing with an epigenetic phenomenon.

Other investigators were trying to make plants resistant to viral infection. One approach was to overexpress in plants a dominant-negative derivative of a common viral replication factor: this protein was expected to block replication of any infecting virus that used this common replicative mechanism. Although the dominant-negative viral product blocked replication of the potato virus from which it was derived, its specificity of action was surprisingly tightly restricted to that virus. It was also shown that the protein itself was not even needed—just the RNA.



**BOX 20-2 FIGURE 1** Petunia flower. An example of the effects of overexpressing the pigment gene chalcone synthase in what would otherwise be a completely purple petunia flower. (Courtesy of Richard A. Jorgensen, University of Arizona.)

Meanwhile other researchers were using antisense RNA to knock down expression of the *par-1* gene in worms. Their intention was to prove that this gene was responsible for a particular developmental phenotype. Antisense RNA produced *in vitro* and injected into the developing worm induced the phenotype predicted for the loss of *par-1* expression. But it was found that sense RNA had the same effect. This was only included in the experiment as a negative control, of course; it was not expected to have any effect on expression. RNAs unrelated to the *par-1* gene had no effect.

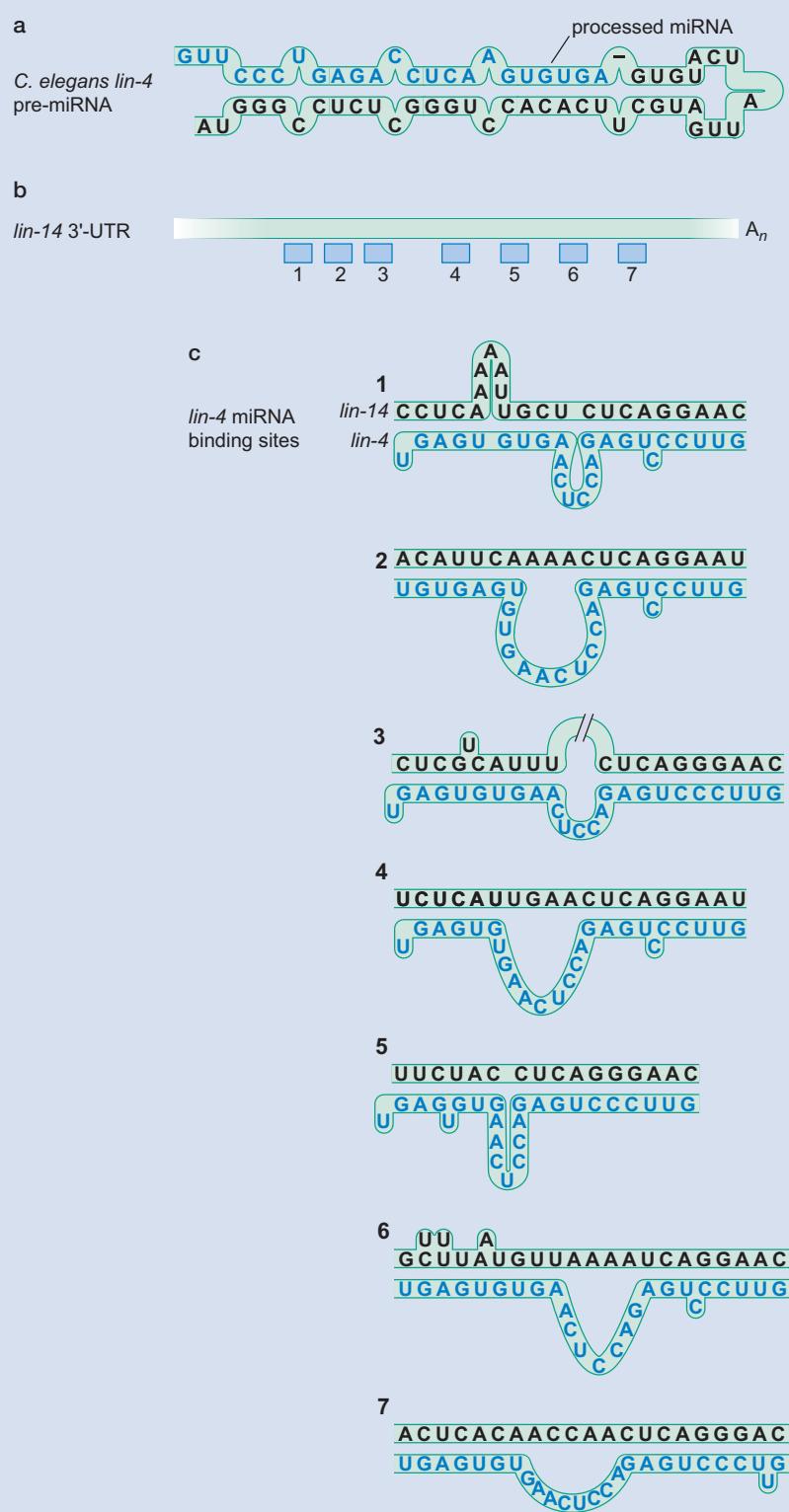
The explanation for this RNA-dependent gene repression was provided by Andrew Fire and Craig Mello in experiments that earned them the 2006 Nobel Prize in Physiology or Medicine. They showed that it was, in fact, neither sense nor antisense RNA that silenced the gene—it was the dsRNA produced by a mix of the two. It turned out that the RNA preparations of sense or antisense were both contaminated with small amounts of the opposite strand, and it was the resulting double-strand population that caused silencing. When dsRNAs were deliberately prepared, they were shown to be very potent in eliminating expression of the target gene. Hence, the phenomenon of RNAi had been discovered, a finding published in 1998.

Mechanistic insights came thick and fast from several laboratories. First, dsRNAs were shown to trigger degradation of homologous mRNAs in extracts from *Drosophila* cells, an assay that led to the identification of RISC. The identification of siRNAs—the species that directs RISC to the target genes—was reported in plants in 1999. Dicer, the nuclease that creates them, was described in 2001. And the final major component of the pathway, Slicer, was identified in 2005 when the crystal structure of Argonaute revealed the protein to be an RNase.

In addition to being needed to generate the siRNAs, Dicer was shown also to be required for miRNAs to function during development. The first miRNA and its target had been described in 1993, by Victor Ambros and Gary Ruvkun, respectively. At the time, this observation was seen as a neat but eccentric oddity; the *lin-4* gene encoded a small RNA that acted on a target gene, *lin-14*, by virtue of sequence complementarity between the miRNA and regions of the 3'-UTR of the target genes (Box 20-2 Fig. 2). Subsequently, other miRNAs were found in worms, some with homology to similar genes in animals and plants, suggesting that this mechanism of regulation was more widespread. Thus, the picture emerged of a world of tiny RNAs involved in gene regulation—some exogenously supplied, others built in as part of the gene regulatory programs used during development. The field developed very rapidly, as the dates in this account reveal, moving from obscure phenomenology to a Nobel Prize and demanding of its own chapter in textbooks, in just 15 years. The accelerated progress was perhaps largely a consequence of the range of species (yeast, plant, and worm) studied and approaches (genetics, biochemistry, structural studies, and bioinformatics) used.

**Box 20-2 (Continued)**

**BOX 20-2 FIGURE 2** microRNA *lin-4* binds within the 3'-UTR of its target gene *lin-14*. (a) The *lin-4* pre-miRNA before processing by Dicer. The sequence of the miRNA is shown in blue. (b) The seven sequences within the *lin-4* 3'-UTR that can base-pair with the *lin-4* miRNA to various extents, as shown in part c. The biology behind the regulatory events controlled by *lin-4*/*lin-14* is rather intriguing. We shall see (in Chapter 21) examples of how gene expression regulates developmental decisions. Famous examples include the Hox genes, whose expression defines spatial identity; hence, limbs and other structures form where they should along the body axis, and not in the wrong place—for example, a leg does not grow out of a head. But that is, in fact, what can happen in mutants lacking certain Hox genes—thus, the famous *Drosophila* homeotic mutant *antennapedia*, where a leg grows out of the head in place of an antenna. The *lin-14* protein is an example of a regulator that controls temporal differential identity. Thus, mutations in these so-called heterochronic genes result in temporal transformations rather than spatial ones. That is, cells adopt fates normally specific to cells at earlier or later stages in development. Expression of the *lin-14* protein is, of course, regulated by the microRNA *lin-4*, as described in the text. (Modified from Ha I. 1996. *Genes Dev.* **10**: 3041–3050, Fig. 1. © Cold Spring Harbor Laboratory Press.)



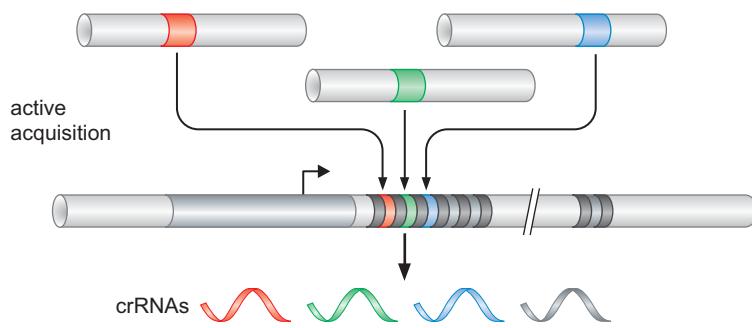
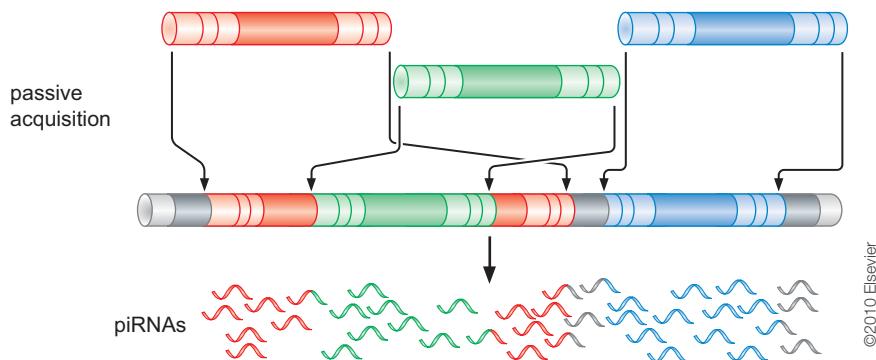
We earlier described the CRISPR–*cas* system in bacteria (see earlier Fig. 20-9). In that case, DNA sequences derived from foreign DNA (phage or plasmids) are accumulated in regions of the genome that can subsequently be expressed in the form of small RNA molecules that destroy homologous nucleic acids should they invade the cell again. Although the CRISPR system has no components in common with the eukaryotic RNAi machinery (other than using small RNAs to guide protein complexes to destroy targeted nucleic acids), in many ways the function and logic of the two systems are strikingly alike. This is especially true of the piRNA system.

As we have seen, piRNAs are the third class of small regulatory RNAs (after siRNAs and miRNAs); they arise from long, single-stranded transcripts of **piRNA clusters** in the genome, without the need for Dicer action.

Like CRISPR, piRNAs seem to target nucleic acid parasites—but whereas CRISPR's main targets are infecting phage, for an animal genome, the main threats are **transposons**. Transposons are found in essentially all eukaryotes and, in some cases, make up a substantial amount of a genome (see Chapters 8 and 12). In humans, for example, ~45% of our genome is made up of sequences that were once transposons. Transposons are often transcriptionally silent and packaged into heterochromatin. In some RNAi mutants, however, the histone modifications associated with transposon silencing are lost. In addition, in plants and worms, several siRNAs have been identified that correspond to transposons. And in some cases in both these organisms, the loss of RNAi reactivates transposons, causes them to jump, and leads to high levels of spontaneous mutagenesis. Not as many transposons are reactivated as are known to generate siRNAs, however. This might reflect a situation similar to that described above for mating-type silencing in *S. pombe*: RNAi might be essential for initiating silencing of some transposons, but the silencing then becomes self-sustaining without further need for the siRNAs.

The piRNAs seem particularly dedicated to the task of protecting cells from transposons and are predominantly expressed in the germline (the cells whose protection is most important). The piRNA clusters contain bits and pieces of transposons—they have been described as “transposon graveyards”—and thus the piRNAs generated from transcripts of these regions often target transcripts made by active transposons, inhibiting their action. As we saw, the CRISPR system includes a mechanism to actively acquire specimens of invading foreign DNA and uses this to prime the CRISPR arrays to make small RNAs that will target those very same sequences should they turn up in the cell again. The piRNA system appears to lack this feature, but DNA sequences that end up in piRNA clusters presumably get there by moving themselves into that region, thus the clusters will be enriched for transposon sequences (see Fig. 20-19).

In plants, RNAi is needed to control transposons, as we have mentioned, and viral infection as well. The protective effect of RNAi on viral infection of plants has been widely observed. Indeed, the effects were recognized long before RNAi was known to be an underlying mechanism. When one leaf on a plant is infected by a virus, a factor able to silence replication of the virus is spread systematically throughout the whole plant. This factor does not protect that originally infected leaf, but it does stop the infection from spreading. In plants mutated in the Argonaute or Dicer genes, infection spreads unhindered and viral replication is much higher. The protecting signal comprises siRNAs generated from the viral genome itself. Viruses have retaliated: they often carry genes whose function is to protect the infecting virus from host RNAi. One example is HcPro from potato virus Y, which acts to reduce production or stability of siRNAs. Other viral products affect other steps in the defense mechanism, including the systemic spreading of siRNAs.

**a CRISPR system****b piRNA system**

**FIGURE 20-19 Comparison of the bacterial CRISPR (a) and animal piRNA (b) defense systems.** Although many features are analogous, the molecular components are not conserved. In addition, as discussed in the text, whereas CRISPR actively acquires new spacer sequences from infecting phage (see Fig. 20-8), the transposon sequences that prime the piRNA system arrive passively within the piRNA clusters. (Adapted from Karginov F.V. and Hannon G.J. 2010. *Mol. Cell* 37: 7–19, Fig. 5, p. 16. © Elsevier.)

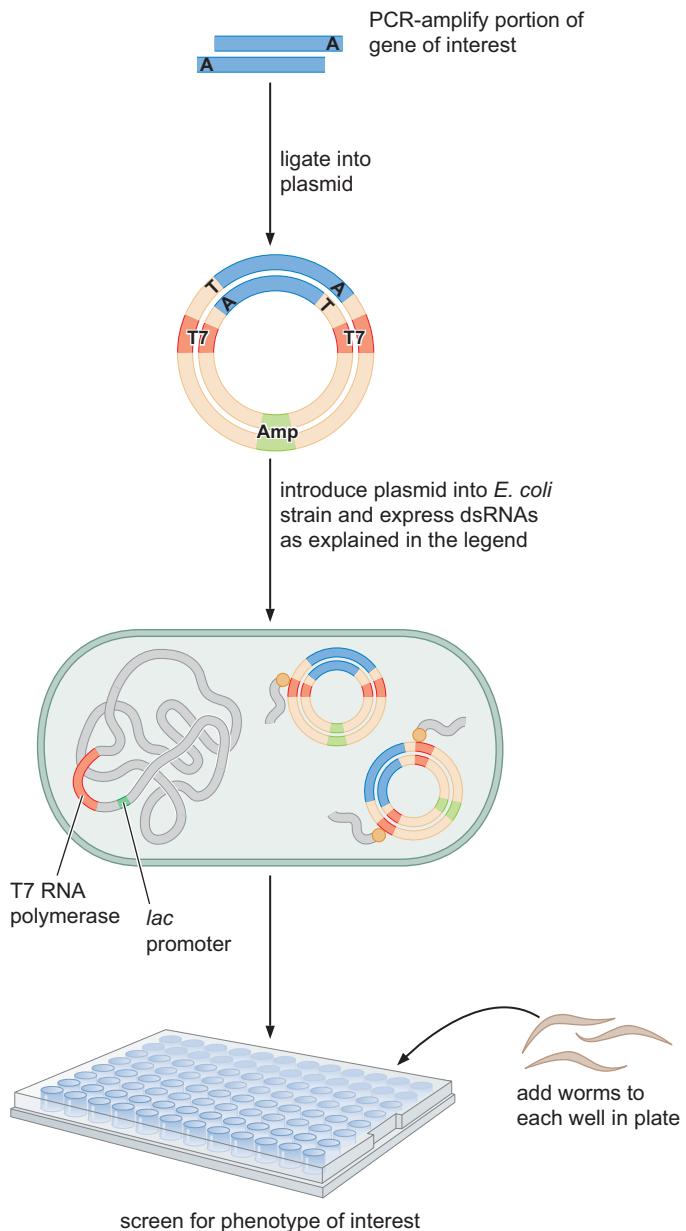
Links between miRNAs and human disease are described in Box 20-3, microRNAs and Human Disease.

### RNAi Has Become a Powerful Tool for Manipulating Gene Expression

The discovery of RNAi arose from observations made by investigators attempting to manipulate gene expression (see Box 20-2). In the case of both cosuppression in plants and antisense RNA in worms, it was attempts to understand unexpected blips in those manipulations that led to the discovery of RNAi. It was therefore perhaps not surprising that once understood, RNAi was quickly exploited as a tool for manipulating gene expression. In worms, this was soon routinely done. Libraries that encode dsRNAs can target any gene in the worm genome and so can be used to screen worms for the consequences of inhibiting expression of any given gene. The general way in which this is done is shown in Figure 20-20. Worms feed on bacteria. In the laboratory, they are fed *E. coli*, and it turns out that the quickest route to a worm's genome is through its stomach: any desired dsRNA can be expressed in the *E. coli* on which the worms feed, and this delivers enough substrate for the RNAi response to be triggered in cells of the worm, switching off genes homologous to the original dsRNA.

It would, of course, be of great benefit to screen for genes in this way in mammalian cells, where traditional genetic screens are not feasible. It was established that artificially synthesized siRNAs, made *in vitro* and introduced into mammalian cells in culture, trigger an RNAi response and down-regulate appropriate target genes, but the efficiency of transfection (getting the RNAs into the cells) was low. Longer dsRNA molecules are also problematic because they trigger a response that shuts down all

**FIGURE 20-20** RNA interference can be induced in worms by feeding bacteria expressing dsRNAs. See Chapter 7 for details of the molecular manipulations required in the first steps of this scheme. The dsRNA expression from the plasmid is under the control of two promoters, in opposite orientations, recognized by a single-subunit RNA polymerase from a phage called T7. The gene for that polymerase is expressed artificially in the cells used in this scheme, under control of the *lac* promoter (Chapter 18). Thus, production of the dsRNA can be controlled using an inducer of the *lac* promoter.



translation in the cell, a response evolved to block viral replication because many viruses have RNA genomes.

Instead of trying to deliver dsRNA into cells, investigators found it more fruitful to mimic miRNAs. To this end, libraries have been generated in which short genes are synthesized as oligonucleotides and cloned in plasmids. Each short gene is designed to give a transcript that will fold into a stem-loop. These are processed by Dicer in the cell to generate an siRNA that will direct silencing of its target genes. These short synthetic genes are called **short hairpin RNA genes (shRNAs)**. By using an appropriately designed shRNA, any individual gene in the genome can be targeted. Or, with a suitable library, a genetic screen can be performed. In such a library, for example, each plasmid would encode an shRNA directed against a different gene. The whole library is transfected into cells such that each cell receives a different shRNA. Cells with a particular phenotype are chosen, and the gene whose repression led to that phenotype can be identified.

## ► MEDICAL CONNECTIONS

### Box 20-3 microRNAs and Human Disease

#### Cancer

A general decrease in levels of many miRNAs is often seen in cancers. This decrease has been taken to indicate that those miRNAs usually have a tumor-suppressing effect. Despite this general trend, other specific miRNAs are up-regulated in some cancers. Analogous to protein-coding genes implicated in cancer, the miRNAs in question are described as being tumor suppressors (if their absence increases cancer) or oncogenic (if their increased expression leads to cancer). Their targets tend to be genes involved in cell cycle progression (proliferation) or apoptosis.

Of the several hundred miRNAs identified in humans, more than one-half are located in regions of the genome regularly disrupted in cancers. Thus, in many cancers, the genes for these miRNAs are deleted or amplified, depending on the nature of the chromosomal rearrangement. Therefore, for example, the miRNAs *miR-15* and *miR-16* induce apoptosis of cells by down-regulating the *BCL2* gene (see Box 20-3 Fig. 1). The most common form of adult leukemia in the Western Hemisphere is chronic lymphocytic leukemia (CLL), a disease associated with deletions in a region of chromosome 13 (13q14). This region of the genome contains the miRNA *miR-15a* and *miR-16a* genes; indeed, these are the only two genes included in the smallest deletions associated with CLL. Thus, when these genes are deleted, apoptosis is down-regulated and tumors can more readily arise and develop.

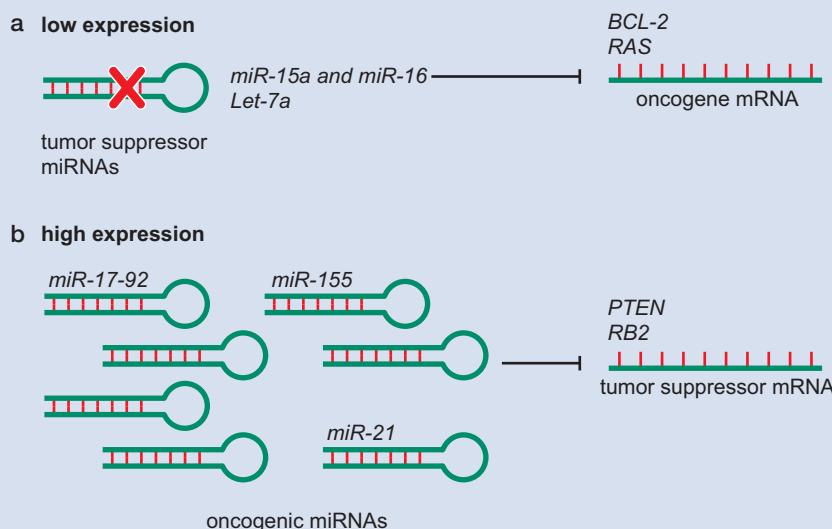
In another region of chromosome 13 (13q31) is found *miR-17-92*, an oncogenic miRNA. Compared with normal tis-

sue, expression of this gene is significantly increased in many cancers, including lung cancer, especially in its most aggressive forms (e.g., small-cell lung cancer). In addition, overexpression of this miRNA in transgenic mice drives tumorigenesis. Among the many predicted targets of *miR-17-92* are two tumor-suppressor genes, *PTEN* and *RB2*. One definite target is the cell cycle progression regulator E2F1. Both these and other examples of miRNAs in cancer are shown in Box 20-3 Figure 1.

#### Fragile X Mental Retardation

Through biochemical analysis of the RISC complex, several associated proteins have been identified. One of these is the Fragile X mental retardation protein (FMRP). The gene encoding this protein (*FMR1*) is X-linked, and its mutation is the cause of the most common inherited form of mental retardation. FMRP is an RNA-binding protein involved in gene regulation; it is known to interact with a number of miRNAs associated with neuroplasticity. Patients lacking FMRP have a range of developmental defects, as well as the mental retardation, due to disrupted gene expression.

*Drosophila* has an FMRP homolog. In flies deficient for this gene, unusual synaptic connections between neurons and muscles were observed. One of the *Drosophila* Argonaute proteins was found to be associated with FMRP, whereas separate studies of Argonaute also found that FMRP was bound to that component of the RNAi machinery. Similar findings followed in human cells as well, indicating an intriguing connection between the Fragile X condition and miRNA maturation and function.



**BOX 20-3 FIGURE 1** miRNAs as tumor suppressors of oncogenes. (a) In this model, an miRNA that normally down-regulates an oncogene can function as a tumor-suppressor gene. The loss of function of the miRNA by mutation or deletion, for example, might result in an abnormal expression of the target oncogene, which would then contribute to tumor formation. (b) Here, the amplification or overexpression of an miRNA that down-regulates a tumor suppressor or other important genes involved in differentiation may contribute to tumor formation by stimulating proliferation, angiogenesis, and invasion. (Redrawn, with permission, from Garzon R. et al. 2006. *Trends Mol. Med.* 12: 580–587, Fig. 2. © Elsevier.)

## LONG NON-CODING RNAs AND X-INACTIVATION

### Long Non-Coding RNAs Have Many Roles in Gene Regulation, Including *Cis* and *Trans* Effects on Transcription

In recent years, high-throughput sequencing techniques (described in Chapter 7) have revealed the presence of many non-coding RNAs expressed in animal and plant cells. So far we have discussed short ones, but there is also a class of non-coding RNAs longer than 200 nucleotides known as **long non-coding RNAs (lncRNAs)**. These serve many roles in processes from translation and splicing to transcriptional regulation. The best-studied case of the latter is the RNA *Xist* involved in the process of mammalian X-inactivation, which we discuss presently. But before that, we consider a couple of the other regulatory lncRNAs with roles in development.

There is much debate regarding exactly how prevalent lncRNAs are in the cell. As just mentioned, most are detected using powerful high-throughput methods that can detect even very rare species. One must remain cautious about assigning biological significance to these in the absence of direct experimentation. But some specific lncRNAs (in addition to *Xist*) have clearly been shown to have specific regulatory functions, and we consider some of those here.

*HOTAIR* is a lncRNA whose gene is found in the *HoxC* cluster in humans, but it acts by regulating expression of the *HoxD* genes on another chromosome (in *trans*) by recruiting to that locus Polycomb Repressive Complex 2 (PRC2; see Chapter 19). PRC2 adds trimethyl groups to lysine 27 of histone H3, a mark (H3K27me3) associated with repressed gene expression (see Fig. 19-29). *HOTAIR* also recruits a second complex that removes histone modifications typically associated with active genes. These two protein complexes bind to separate regions of the *HOTAIR* RNA molecule and are presumably targeted to the specific *HoxD* locus through yet a third region. As well as being involved in development, *HOTAIR* has also been found to be up-regulated in some cancers. Deleting the PRC2-binding domain of *HOTAIR* destroys its regulatory functions in both development and cancer. Although this suggests a critical biological role, it is intriguing to note that *HOTAIR* is poorly conserved in mouse, and its deletion has no apparent phenotype. This would suggest that it might have evolved rapidly within mammals.

Other lncRNAs work not in *trans*, but in *cis*. As we shall see, *Xist* is one such case. But others include RNAs involved in imprinting. We discussed imprinting in Chapter 19 (see Fig. 19-31). The Igf2/H19 locus described there also includes a transcription unit that makes the lncRNA *AIR*. *AIR* is itself imprinted and is only expressed from the paternal allele; its expression is required for repression of several imprinted genes on the paternal chromosome from which it is expressed. The RNA somehow remains associated with the chromosome and recruits a protein complex that trimethylates H3K9, leading to repression of transcription at those targeted promoters.

### X-Inactivation Creates Mosaic Individuals

We now look in more detail at the function of the lncRNA called *Xist* and the process of X-inactivation. Female mammals have two X chromosomes, whereas males have only one X and a Y chromosome. Although this is the basis of sex determination—what enables males to differ from females—it also poses a problem: any gene encoded by the X chromosome would, if left unchecked, be expressed at twice the level in females as in males.

This imbalance would potentially cause disruption to metabolic and other cellular processes. Avoiding such problems requires what is called **dosage compensation**. In mammals this is achieved by females **inactivating** one of their two X chromosomes. This action results in none of the genes on that copy of the chromosome being expressed. In placental mammals, inactivation occurs at the 32- to 64-cell stage, and the choice of which X chromosome to inactivate—the maternal or paternal copy—is apparently made at random in each cell at that time. Once selected in each cell, the same copy remains inactivated in all of the descendants of that cell.

A consequence of inactivation being random in each cell is that females are mosaics—some of their cells express the paternal and others the maternal X chromosome. This is usually of little consequence, although it can influence the severity of symptoms of X-linked diseases, depending on the proportion of cells in which the mutated gene is expressed or silenced. A more familiar example is the calico (or tortoiseshell) cat (Fig. 20-21). In cats, one gene on the X chromosome influences whether fur is orange or black—one allele of that gene gives rise to orange fur, and another allele gives black. In cats heterozygous for this gene, the different patches of black and orange fur reveal regions made up of cells in which one or the other X chromosome was inactivated. This observation also explains why all calico cats are female. The white comes from effects of an autosomal gene.

### Xist Is a Long Non-Coding RNA That Inactivates a Single X Chromosome in Female Mammals

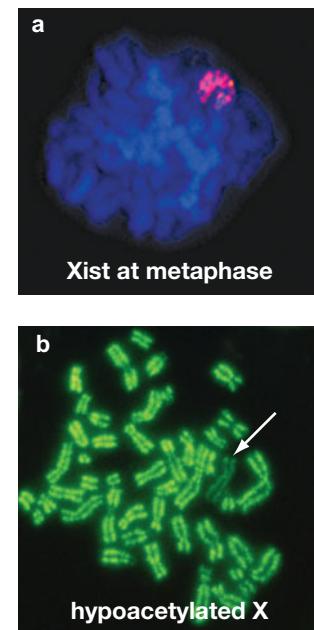
How is an X chromosome inactivated, and how is inactivation inherited through the remainder of development? The initiating regulator is an RNA molecule called *Xist*. This RNA is encoded within the locus known to be vital for X-inactivation, the *Xic* (X-inactivation center) on the X chromosome. *Xist* RNA coats the X chromosome from which it is expressed. This is shown in the *in situ* hybridization result in Figure 20-22a. It is not clear what causes this coating nor how it is restricted to one X chromosome (i.e., why it acts only in *cis*). It is, however, known that the action of *Xist* is central to inactivation and does not require other X-chromosomal sequences beyond *Xic*: when expressed ectopically from an autosomal location (i.e., from a non-sex chromosome), *Xist* can, to varying extents, silence genes along that chromosome. That is, it “inactivates” the autosome from which it is expressed.

*Xist* RNA itself does not cause silencing, but it recruits other factors that modify and condense chromatin (PRC2, etc.) and perhaps methylate DNA as well (just as we saw in other examples of mammalian silencing in Chapter 19). Later, there is an accumulation of a rare histone variant called MacroH2A, which is typically associated with compact, silent chromatin. It is these modifications that cause silencing and ensure that it is inherited: once firmly established, *Xist* itself is no longer required. One characteristic of histone modification of the inactivated X chromosome is shown in Figure 20-22b. There, the single inactivated X chromosome is much less acetylated than is the rest of the genome. As we saw in earlier chapters (Chapters 8 and 19), deacetylated histones are associated with regions of the genome that are not transcribed.

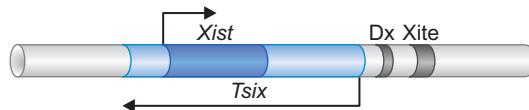
How does a cell choose which X chromosome to inactivate? The answer is still proving elusive, but another RNA regulator might be key. This other RNA is also encoded by the *Xic* locus but on the opposite strand and overlapping the *Xist* gene. It is called *Tsix* (*Xist* spelled backward) and acts as a negative regulator of *Xist* (Fig. 20-23). Indeed, if *Tsix* is mutated on a given



**FIGURE 20-21** Visualizing X-inactivation: the calico cat. The patches of orange and black fur provide an indirect visualization of X-chromosome inactivation, as described in the text. (Courtesy VG.)



**FIGURE 20-22** Visualizing X-inactivation: molecular markers. (a) Localization of *Xist* RNA along the inactive X chromosome is shown by *in situ* hybridization in metaphase cells. (b) Chromosomes are stained for acetylation on histone H4. The arrow points to the inactivated X chromosome, which has much lower levels of acetylation than the other chromosomes. (Reprinted, with permission, from Brockdorff N. and Turner B.M. 2007. *Epigenetics* [ed. D. Allis et al.], p. 327. © Cold Spring Harbor Laboratory Press.)



**FIGURE 20-23** *Tsix* antagonizes expression and action of *Xist*. *Tsix* (shown in light blue shading) is expressed as an antisense RNA of *Xist* (in dark blue shading) and is longer than *Xist*. The degree of overlap is indicated in the dark blue region. Xite and DxPas34 (Dx) are regulatory elements that control expression of the genes. At the start of inactivation, both *Xist* and *Tsix* are expressed from both X chromosomes, but after awhile, the chromosome that will become inactivated increases expression of *Xist*, whereas expression from that chromosome destined to remain active decreases *Xist* expression. How this change in *Xist* levels is regulated by *Tsix* is still not clear, but if *Tsix* is deleted from either chromosome, it is always that copy that becomes inactivated.

X chromosome, it is that chromosome that will be chosen for inactivation. Thus, a balance between the production and stability of the *Xist* and *Tsix* RNAs may tilt the outcome one way or the other in each cell.

Dosage compensation is necessary in all animals (e.g., worms and flies) just as it is in mammals. But in each case, the mechanisms for achieving compensation are different. For example, in *Drosophila*, it is achieved by *increasing* expression of X-linked genes in the male (rather than decreasing them in the female). But there, too, the mechanism involves non-coding regulatory RNAs. In this case, the RNAs (called *roX1* and *roX2*) are involved in recruiting chromatin-modifying complexes to genes on the X chromosome in males, where they help activate transcription.

## SUMMARY

Despite it being proposed as long ago as 1961 that RNA molecules were likely agents of gene regulation, it is only in the last decade that their widespread occurrence and significance in that role have come to light. Before that, attenuation of the *trp* operon in *E. coli* was a rare case in which RNA sequences in the 5' region of an mRNA were known to control expression of the downstream genes. In that case, alternative patterns of intramolecular base pairing within that region of the RNA gives rise to alternative secondary structures that communicate different outcomes to the genes. In one conformation, transcription is terminated before it enters the coding region of the downstream genes, whereas in another conformation, it allows that transcription to continue, and the genes are expressed.

Riboswitches control genes in a similar way: alternative secondary structures in the 5'-UTRs of genes determine whether transcription of those genes continues (or, in other cases, whether translation is initiated). With riboswitches, the choice of alternative secondary structure depends on the direct binding to the RNA of the ligands that control the given gene.

*E. coli* also encodes small RNAs (sRNAs) that act in *trans* to regulate genes. Thus, small genes encode short RNAs that base-pair with mRNAs bearing complementary sequences. This situation either inhibits translation of those target mRNAs, triggers their destruction, or even, in some cases, stimulates their translation. The actions of bacterial small RNAs are similar in many regard to those of sRNAs that

regulate genes in eukaryotic cells, although the machinery used to produce these eukaryotic RNA regulators and the machinery used in achieving their effects on target genes are quite different.

One bacterial system more closely mirrors what we see in eukaryotes, and this is the CRISPR system. Clusters of characteristic repeating DNA sequences (different in each cluster) give rise to a special class of sRNAs that direct a protein machinery to destroy infecting phage and plasmids—indeed, any foreign DNA that finds its way into the cell. The ability to distinguish “self” from “non-self” stems from the fact that the CRISPR regions accumulate fragments of phage genomes from earlier infections and these determine the sequence of the targeting RNAs.

The closest equivalent of the CRISPR in eukaryotes are the so-called piRNA clusters. These regions contain fragments of transposons (and are sometimes called “transposon graveyards”). The short piRNAs generated from these regions silence homologous transposons within the genome, particularly in the germline, ensuring they remain inactive. Small interfering RNAs (siRNAs) and microRNAs (miRNAs) are two other classes of short regulatory RNAs found in eukaryotes. Unlike piRNAs, siRNAs and miRNAs are generated from regions of double-stranded RNA through processing by an enzyme called Dicer. The siRNAs are derived directly from endogenous or exogenous dsRNA in a single step. The miRNAs are encoded within the genome and their double-stranded character derives from regions of secondary

structure that are recognized first by an enzyme called Drosha, which processes them to the stage where Dicer can then act. In both cases the active regulatory RNA species generated by Dicer are 19–25 nucleotides long. Both Drosha and Dicer have RNase domains and cut their substrate RNAs on the basis of size and structure rather than sequence.

Once produced, siRNAs and miRNAs act in essentially the same way. They are incorporated into a machine called RISC, where one of the RNA strands is selected as the so-called guide RNA and directs the mature RISC complex to target RNAs with complementarity to that guide RNA. Once there, RISC either “slices” the RNA (through its catalytic subunit Argonaut, which includes an RNase H–related domain) or inhibits the translation of the mRNA. Which route to silencing is chosen depends largely on how good the base-pairing match is between the guide RNA and the target—the higher the match, the more likely it is to trigger slicing. The guide RNA can also direct RISC with associated histone-modifying complexes to promoter regions, where it silences genes transcriptionally by modifying their promoters. Even in these cases, recruitment to the promoter is through base pairing between the guide RNA and an mRNA, but in this instance, one still being made and thus tethered by RNA polymerase II at the gene.

miRNAs are encoded by genes within organisms where they typically act as regulators of genes involved in development—those from worms and plants are well-studied exam-

ples. miRNAs have also been associated with cancer, with some miRNAs being classified as tumor suppressors and others as oncogenes. The dsRNAs that give rise to siRNAs can arise from various sources ranging from infecting viruses, to transcribed repeat regions (centromeres or transposons), to dsRNA introduced into a cell deliberately by an experimenter who wants to down-regulate expression of a specific gene. This latter use of RNAi has become a regular tool and is particularly useful in systems where traditional genetics is not feasible.

Animal and plant cells also contain “long” regulatory RNAs (of 200 nucleotides or more). These have roles in development, as we saw for the cases of HOTAIR and AIR, and although their detailed mechanisms of action are not known, they appear to recruit protein complexes that modify the chromatin in the vicinity of the genes they regulate. Some of these long RNAs work in *trans* and others in *cis*. The best-understood example of the latter is *Xist*, which directs inactivation of an X-chromosome in mammalian dosage compensation. Female animals have two X chromosomes, whereas males have just one (and a Y chromosome). To ensure that both sexes express comparable amounts of X-chromosome gene products, a mechanism of dosage compensation must correct for this unequal chromosome number. Mammals do this by inactivating one of the X chromosomes in females. An RNA molecule (*Xist*) encoded on the X chromosome regulates this process.

## BIBLIOGRAPHY

### Books

Stillman B. and Stewart D., eds. 2006. *Regulatory RNAs*. Cold Spring Harbor Symposia on Quantitative Biology, Vol. 71. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

### Bacterial Small RNAs

Gottesman S. and Storz G. 2011. Bacterial small RNA regulators: Versatile roles and rapidly evolving variations. *Cold Spring Harb. Perspect. Biol.* **3**: a003798.  
 Storz G., Vogel J., and Wassarman K.M. 2011. Regulation by small RNAs in bacteria: Expanding frontiers. *Mol. Cell* **43**: 880–891.  
 Waters L.S. and Storz G. 2009. Regulatory RNAs in bacteria. *Cell* **136**: 615–628.

### Riboswitches and Attenuation

Bastet L., Dubé A., Massé E., and Lafontaine D. A. 2011. New insights into riboswitch regulation mechanisms. *Mol. Microbiol.* **80**: 1148–1154.  
 Breaker R. R. 2011. Prospects for riboswitch discovery and analysis. *Mol. Cell* **43**: 867–879.  
 Gollnick P., Babitzke P., Antson A., and Yanofsky C. 2005. Complexity in regulation of tryptophan biosynthesis in *Bacillus subtilis*. *Annu. Rev. Genet.* **39**: 47–68.  
 Winkler W.C. 2005. Riboswitches and the role of noncoding RNAs in bacterial metabolic control. *Curr. Opin. Chem. Biol.* **9**: 594–602.  
 Winkler W.C. and Breaker R.R. 2005. Regulation of bacterial gene expression by riboswitches. *Annu. Rev. Microbiol.* **59**: 487–517.

Yanofsky C. 2000. Transcription attenuation: Once viewed as a novel regulatory strategy. *J. Bacteriol.* **182**: 1–8.

### CRISPR

Bhaya D., Davison M., and Barrangou R. 2011. CRISPR-Cas systems in bacteria and archaea: Versatile small RNAs for adaptive defense and regulation. *Annu. Rev. Genet.* **45**: 273–297.  
 Karginov F.V. and Hannon G.J. 2010. The CRISPR system: Small RNA-guided defense in bacteria and archaea. *Mol. Cell* **37**: 7–19.  
 Jore M.M., Brouns S.J.J., and van der Oost J. 2012. RNA in defense: CRISPRs protect prokaryotes against mobile genetic elements. *Cold Spring Harb. Perspect. Biol.* **4**: a003657.  
 Wiedenheft B., Sternberg S.H., and Doudna J.A. 2012. RNA-guided genetic silencing systems in bacteria and archaea. *Nature* **482**: 331–338.

### Mechanisms of RNAi

Baulcombe D. 2005. RNA silencing. *Trends Biochem. Sci.* **30**: 290–293.  
 Czech B. and Hannon G. J. 2011. Small RNA sorting: Matchmaking for Argonautes. *Nat. Rev. Genet.* **12**: 19–31.  
 Farazi T.A., Juranek S.A., and Tuschl T. 2008. The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members. *Development* **135**: 1201–1214.  
 Joshua-Tor L. and Hannon G.J. 2011. Ancestral roles of small RNAs: An Ago-centric perspective. *Cold Spring Harb. Perspect. Biol.* **3**: a003772.

- Liu Q. and Paroo Z. 2010. Biochemical principles of small RNA pathways. *Annu. Rev. Biochem.* **79**: 295–319.
- Molnar A., Melnyk C., and Baulcombe D. C. 2011. Silencing signals in plants: A long journey for small RNAs. *Genome Biol.* **12**: 215.
- Peters L. and Meister G. 2007. Argonaute proteins: Mediators of RNA silencing. *Mol. Cell* **26**: 611–623.
- Tolia N.H. and Joshua-Tor L. 2007. Slicer and the Argonautes. *Nat. Chem. Biol.* **3**: 36–43.
- Volpe T. and Martienssen R. A. 2011. RNA interference and heterochromatin assembly. *Cold Spring Harb. Perspect. Biol.* **3**: a003731.
- Zaratiegui M., Irvine D.V., and Martienssen R.A. 2007. Noncoding RNAs and gene silencing. *Cell* **128**: 763–776.

### siRNAs, miRNAs, and piRNAs

- Ambros V. 2011. MicroRNAs and developmental timing. *Curr. Opin. Genet. Dev.* **21**: 511–517.
- Banisch T.U., Goudarzi M., and Raz E. 2012. Small RNAs in germ cell development. *Curr. Top. Dev. Biol.* **99**: 79–113.
- Bushati N. and Cohen S.M. 2007. microRNA functions. *Annu. Rev. Cell Dev. Biol.* **23**: 175–205.
- Ebert M.S. and Sharp P.A. 2012. Roles for microRNAs in conferring robustness to biological processes. *Cell* **149**: 515–524.
- Esteller M. 2011. Non-coding RNAs in human disease. *Nat. Rev. Genet.* **12**: 861–874.
- Fabian M.R. and Sonenberg N. 2012. The mechanics of miRNA-mediated gene silencing: A look under the hood of miRISC. *Nat. Struct. Mol. Biol.* **19**: 586–593.

## QUESTIONS

### MasteringBiology®

*For instructor-assigned tutorials and problems, go to MasteringBiology.*

For answers to even-numbered questions, see Appendix 2: Answers.

**Question 1.** Using examples from bacteria, explain the difference between a regulatory RNA that acts in *cis* and one that acts in *trans*.

**Question 2.** Predict the level of transcription of the *trp* operon genes (low or high) in the presence of low levels of tryptophan in *Escherichia coli* cells bearing an in-frame deletion of the two *trp* codons that are usually found in the leader peptide. Explain your answer.

**Question 3.** Describe how bacterial genes are regulated by ribo-switches that respond to metabolites like S-adenosylmethionine (SAM).

**Question 4.** Aside from regulation of gene expression, what other purpose do regulatory RNAs found in prokaryotes and Archaea serve?

**Question 5.** Generally describe three mechanisms for how short RNAs in eukaryotes (siRNAs, miRNAs, and piRNAs) silence expression.

**Question 6.** How does the source and generation of miRNAs differ from those of siRNAs?

**Question 7.** List the steps whereby miRNAs are generated and act to silence gene expression, and name the primary enzymes involved at each stage.

- Malone C.D. and Hannon G.J. 2009. Small RNAs as guardians of the genome. *Cell* **136**: 656–668.
- Ruvkun G. 2008. The perfect storm of tiny RNAs. *Nat. Med.* **14**: 1041–1045.

### Long Non-Coding RNAs

- Alexander M.K. and Panning B. 2005. Counting chromosomes: Not as easy as 1, 2, 3. *Curr. Biol.* **15**: R834–R836.
- Augui S., Nora E.P., and Heard E. 2011. Regulation of X-chromosome inactivation by the X-inactivation centre. *Nat. Rev. Genet.* **12**: 429–442.
- Brockdorff N. 2011. Chromosome silencing mechanisms in X-chromosome inactivation: Unknown unknowns. *Development* **138**: 5057–5065.
- Deng X. and Meller V.K. 2006. Non-coding RNA in fly dosage compensation. *Trends Biochem. Sci.* **31**: 526–532.
- Lee J.T. 2011. Gracefully ageing at 50, X-chromosome inactivation becomes a paradigm for RNA and chromatin control. *Nat. Rev. Mol. Cell Biol.* **12**: 815–826.
- Ng K., Pullirsch D., Leeb M., and Wutz A. 2006. *Xist* and the order of silencing. *EMBO Rep.* **8**: 34–39.
- Pauler F.M., Koerner M.V., and Barlow D.P. 2007. Silencing by imprinted noncoding RNAs: Is transcription the answer? *Trends Genet.* **23**: 284–292.
- Ponting C.P., Oliver P.L., and Reik W. 2009. Evolution and functions of long noncoding RNAs. *Cell* **136**: 629–641.
- Rinn J.L. and Chang H.Y. 2012. Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* **81**: 145–166.

**Question 8.** Label the statement as true or false. Pre-miRNAs are present only in introns. Explain your answer.

**Question 9.** Describe the key features of piwi-interacting RNAs (piRNA) found in eukaryotes.

**Question 10.** To use RNAi as an experimental tool, initially researchers delivered the dsRNA to worms by feeding the worms *E. coli* engineered to express the dsRNA. Name two reasons why directly delivering the dsRNA does not work as efficiently in mammalian cells and describe how researchers overcome this problem.

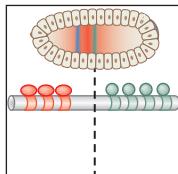
**Question 11.** In the scheme of feeding bacteria to worms to induce an RNAi response in worms, explain the purpose of the following components: *lac* promoter incorporated into the genome of *E. coli*, T7 gene incorporated into the genome of *E. coli*, and T7 promoter on the plasmid introduced into *E. coli*.

**Question 12.** How does the expression of the lncRNA *Xist* cause silencing of one X chromosome in female mammals?

**Question 13.** You want to silence the expression of a target gene using an shRNA.

- Name several possible assays to test expression of the target gene.
- Describe appropriate controls for an assay testing protein levels of the target gene product and proving that down-regulation is due to the shRNA involved.

CHAPTER 21



# Gene Regulation in Development and Evolution

## OUTLINE

**A**NIMAL DEVELOPMENT DEPENDS ON THE differential expression of a constant genome to produce diverse cell types during embryogenesis. A typical animal genome contains approximately 20,000 genes. This is not only true for comparatively simple creatures such as nematode worms, but also pertains to the “crown and summit” of animal evolution, the human genome.

**Differential gene expression** can be defined as the synthesis of a protein (or RNA in the case of non-coding genes) in a subset of the cells comprising an embryo. Differential expression most commonly hinges on de novo transcription. Thus, the  $\beta$ -globin gene is selectively expressed in developing red blood cells, but not other tissues, because the gene is transcribed only in blood cells. However, there are examples of post-transcriptional mechanisms of differential gene expression. For example, mRNAs transcribed from the segmentation gene *hunchback* are distributed throughout the early *Drosophila* embryo, but are translated to produce functional proteins only in anterior (head and thorax), but not posterior (abdomen), regions.

How do we know that differential expression of an invariant genome is the key to animal development? A variety of classical and contemporary studies showed that different cell types contain the same genome. The first conclusive evidence came from the cloning of the frog *Xenopus laevis* in the 1960s and 1970s. These studies culminated with the replacement of the egg nucleus with the nucleus of a gut cell of an adult frog. The gut nucleus was able to sustain embryogenesis, the formation of a *Xenopus* tadpole, and its metamorphosis into an adult frog. The resulting frog is said to be a “clone” of the one that donated its gut cell because the two are genetically identical. Subsequent studies in the late 1990s and early 2000s extended cloning to sheep (Dolly), and it is now possible, at least in principle, to clone most animals.

The most spectacular demonstration of “genetic equivalence” among the different tissues of a developing animal is the transformation of virtually any cell type into an induced pluripotent stem (iPS) cell. Most mammalian embryos, including the human fetus, contain a small group of cells, the inner cell mass (ICM), which form all of the tissues and organs of the adult.

Three Strategies by Which Cells Are Instructed to Express Specific Sets of Genes during Development, 735

Examples of the Three Strategies for Establishing Differential Gene Expression, 738

The Molecular Biology of *Drosophila* Embryogenesis, 746

Homeotic Genes: An Important Class of Developmental Regulators, 762

Genome Evolution and Human Origins, 769

Visit Web Content for Structural Tutorials and Interactive Animations

The ICM cells are said to be “pluripotent” because they can produce many different cell types. The formation of ICM cells depends on the activities of three sequence-specific transcription factors—Oct4, Sox2, and Nanog. The forced expression of these three factors in a differentiated cell type, such as a fibroblast cell (connective tissue), is sufficient to transform them into iPS cells, which have the properties of ICM cells (see Box 21-1, Formation of iPS Cells). Indeed, iPS cells can be used to replace ICM cells within an

### MEDICAL CONNECTIONS

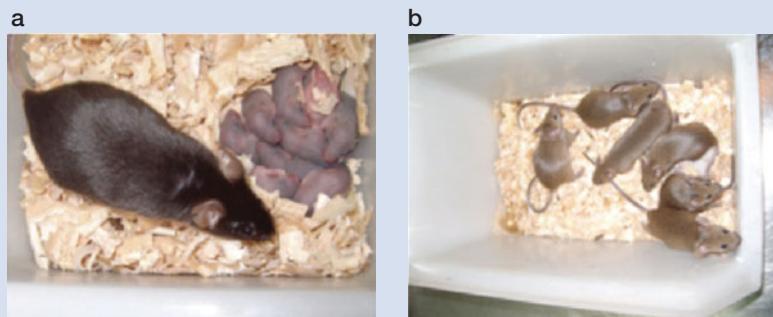
#### Box 21-1 Formation of iPS Cells

The ICM cells of mammalian embryos undergo diverse pathways of differentiation and produce all of the tissues and organs of adults.

In the early 2000s, when stem cell fever gripped the biomedical research community, it was thought that the isolation of ICM cells would be the rate-limiting step for the use of stem cells in regenerative medicine. For example, insulin-dependent diabetics lack  $\beta$ -cells, secretory cells in the pancreas that produce insulin in response to increases in blood sugar levels after a meal. There is the hope that it will be one day possible to replace these  $\beta$ -cells with those produced in laboratory culture using stem cells. But the isolation of ICM cells from human fetuses presented a dizzying maze of technical and ethical challenges. This controversy, which became quite heated and political, has dissipated into obscurity because of a remarkable series of experiments conducted by Takahashi and Yamanaka in 2006. As a postdoctoral fellow, Yamanaka had identified a gene that is selectively expressed in ICM cells. He inserted *lacZ* into this gene and used it as a “marker” for identifying mouse fibroblasts that had been converted into stem cells (these converted cells are called induced pluripotent stem [iPS] cells). The marker gene is not normally expressed in fibroblasts but is activated when the cells are transformed into iPS cells. A variety of research groups had identified about 30 different transcription factors (TFs) that showed expression in cultured ICM stem cells. Takahashi and

Yamanaka systematically forced the expression of these different TFs in fibroblasts, resulting in the induction of the *lacZ* marker gene. They then coexpressed different combinations of the TFs and found that three of these factors—Oct4, Sox2, and Nanog—were particularly potent in converting or reprogramming fibroblasts into iPS cells. These reprogrammed cells have most or all of the properties of bona fide ICM cells. The iPS cells can be induced to form just about any cell type, such as cardiomyocytes (heart muscle). In a further remarkable experiment, Yamanaka and coworkers showed that it was possible to obtain adult mice from iPS cells injected into embryos. The results revealed in Box 21-1 Figure 1 show that the characteristics associated with the iPS cells are transmitted in the germline of the resulting offspring (Box 21-1 Fig. 1).

The competence of different adult tissues to be transformed into iPS cells, which, in turn, can be induced to produce any tissue, is a clear demonstration of genetic equivalence. These studies also raise the possibility of “replacement medicine,” whereby skin fibroblasts from a sick individual can be used to produce iPS cells, which are subsequently programmed to generate the missing tissues causing illness, for example, the dopaminergic neurons for sufferers of Parkinson’s or the  $\beta$ -cells for diabetics. Gurdon and Yamanaka were awarded the Nobel Prize in Physiology or Medicine in 2012 for their discovery that differentiated animal cells can be reprogrammed into any tissue.



**BOX 21-1 FIGURE 1** The developmental potential of iPS cells. (a) Reprogrammed (iPS) cells, derived from a mouse of black coat color, were injected into the blastocyst (early-stage embryo) of a female mouse of white coat color, producing the black adult mouse (a male) shown here. Next to the adult are its progeny, newborn pups resulting from mating the iPS male with a white female. (b) The newborn pups in panel a have developed into young mice with a brown coat color, which is the typical result seen when a black male is crossed with a white female. (Reproduced, with permission, from Zhao X.Y. et al. 2009. *Nature* **461**: 86. © Macmillan. a and b are Fig. 2f and 2g, respectively.)

embryo and produce adult mice whose tissues are derived solely from the iPS cells.

In this chapter, we consider the different mechanisms for achieving differential gene expression in animal development. In the first half of this chapter, we describe how cells communicate with each other during development to ensure that each expresses a particular set of genes required for their proper development. Simple examples of each of these strategies are then described. In the next part, we describe how these strategies are used in combination with the transcriptional regulatory mechanisms described in Chapter 19 to control the development of an entire organism—in this case, the fruit fly. In the final part of the chapter, we discuss how changes in gene regulation can cause diversity of animal morphology during evolution. A particularly important class of developmental control genes, the homeotic genes, is described.

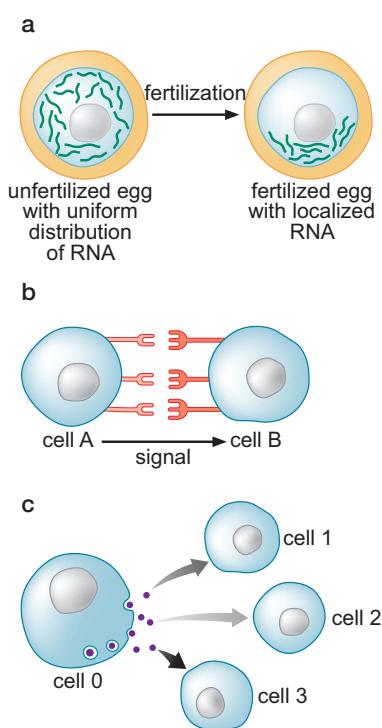
### THREE STRATEGIES BY WHICH CELLS ARE INSTRUCTED TO EXPRESS SPECIFIC SETS OF GENES DURING DEVELOPMENT

We have already seen how gene expression can be controlled by “signals” received by a cell from its environment. For example, the sugar lactose activates the transcription of the *lac* operon in *Escherichia coli*, whereas viral infection activates the expression of the  $\beta$ -interferon gene in mammals. In this chapter, we focus on the strategies that are used to instruct genetically identical cells to express distinct sets of genes and thereby differentiate into diverse cell types. The three major strategies are **mRNA localization**, **cell-to-cell contact**, and **signaling through the diffusion of a secreted signaling molecule** (Fig. 21-1). Each of these strategies is introduced briefly in the following sections.

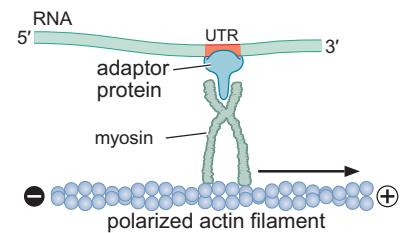
#### Some mRNAs Become Localized within Eggs and Embryos Because of an Intrinsic Polarity in the Cytoskeleton

One strategy to establish differences between two genetically identical cells is to distribute a critical regulatory molecule asymmetrically during cell division, thereby ensuring that the daughter cells inherit different amounts of that regulator and thus follow different pathways of development. Typically, the asymmetrically distributed molecule is an mRNA. These mRNAs can encode RNA-binding proteins or cell-signaling molecules, but most often they encode transcriptional activators or repressors. Despite this diversity in the function of their protein products, a common mechanism exists for localizing mRNAs. Typically, they are transported along elements of the cytoskeleton, actin filaments, or microtubules. The asymmetry in this process is provided by the intrinsic asymmetry of these elements.

Actin filaments and microtubules undergo directed growth at the + ends (Fig. 21-2). An mRNA molecule can be transported from one end of a cell to the other end by means of an “adaptor” protein, which binds to a specific sequence within the non-coding 3' **untranslated trailer region (3' UTR)** of an mRNA. Adaptor proteins contain two domains. One recognizes the 3' UTR of the mRNA, whereas the other associates with a specific component of the cytoskeleton, such as myosin. Depending on the specific adaptor used, the mRNA–adaptor complex either “crawls” along an actin



**FIGURE 21-1** The three strategies for initiating differential gene activity during development. (a) In some animals, certain “maternal” RNAs present in the egg become localized either before or after fertilization. In this example, a specific mRNA (green squiggles) becomes localized to vegetal (bottom) regions after fertilization. (b) Cell A must physically interact with cell B to stimulate the receptor present on the surface of cell B. This is because the “ligand” produced by cell A is tethered to the plasma membrane. (c) In this example of long-range cell signaling, cell 0 secretes a signaling molecule that diffuses through the extracellular matrix. Different cells (1, 2, 3) receive the signal and ultimately undergo changes in gene activity.



**FIGURE 21-2** An adaptor protein binds to specific sequences within the 3' UTR of the mRNA. The adaptor also binds to myosin, which “crawls” along the actin filament in a directed fashion, from the “−” end to the growing “+” end of the filament.

filament or directly moves with the + end of a growing microtubule. We will see how this basic process is used to localize mRNA determinants within the egg or to restrict a determinant to a single daughter cell after mitosis.

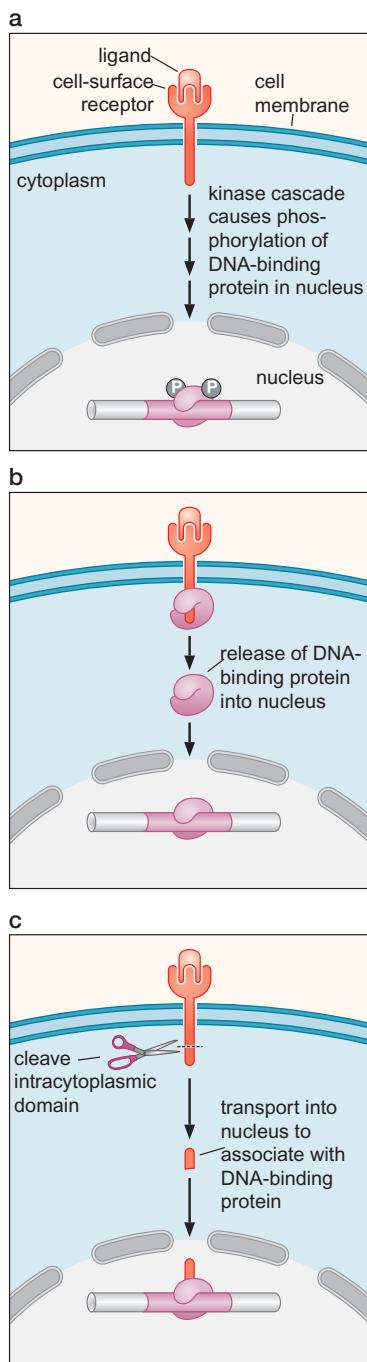
### Cell-to-Cell Contact and Secreted Cell-Signaling Molecules Both Elicit Changes in Gene Expression in Neighboring Cells

A cell can influence which genes are expressed in neighboring cells by producing extracellular signaling proteins. These proteins are synthesized in the first cell and then either deposited in the plasma membrane of that cell or secreted into the extracellular matrix. Because these two approaches have features in common, we consider them together here. We then see how secreted signals can be used in other ways.

A given signal (of either sort) is generally recognized by a specific receptor on the surface of recipient cells. When that receptor binds to the signaling molecule, it triggers changes in gene expression in the recipient cell. This communication from the cell-surface receptor to the nucleus often involves **signal transduction pathways** of the sort we considered in Chapter 19. Here, we summarize a few basic features of these pathways.

Sometimes, ligand–receptor interactions induce an enzymatic cascade that ultimately modifies regulatory proteins already present in the nucleus (Fig. 21-3a). In other cases, activated receptors cause the release of DNA-binding proteins from the cell surface or cytoplasm into the nucleus (Fig. 21-3b). These regulatory proteins bind to specific DNA-recognition sequences and either activate or repress gene expression. Ligand binding can also cause proteolytic cleavage of the receptor. Upon cleavage, the intracytoplasmic domain of the receptor is released from the cell surface and enters the nucleus, where it associates with DNA-binding proteins and influences how those proteins regulate transcription of the associated genes (Fig. 21-3c). For example, the transported protein might convert what was a transcriptional repressor into an activator. In this case, target genes that were formerly repressed before signaling are now induced. We consider examples of each of these variations in cell signaling in this chapter.

Signaling molecules that remain on the surface control gene expression only in those cells that are in direct, physical contact with the signaling cell. We refer to this process as **cell-to-cell contact**. In contrast, signaling molecules that are secreted into the extracellular matrix can work over greater distances. Some travel over a distance of just one or two cell diameters, whereas others can act over a range of 50 cells or more. Long-range signaling molecules are sometimes responsible for positional information, which is discussed in the next section.

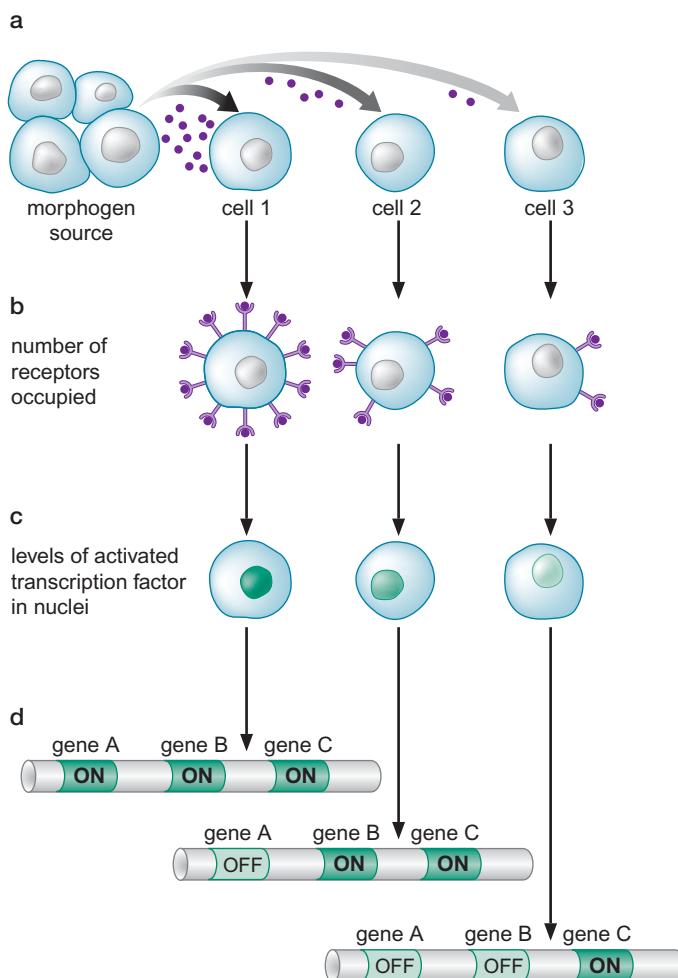


**FIGURE 21-3** Different mechanisms of signal transduction. A ligand (or “signaling molecule”) binds to a cell-surface receptor. (a) The activated receptor induces latent cellular kinases that ultimately cause the phosphorylation of DNA-binding proteins within the nucleus. This phosphorylation causes the regulatory protein to activate (or repress) the transcription of specific genes. (b) The activated receptor releases a dormant DNA-binding protein from the cytoplasm so that it can now enter the nucleus. Once in the nucleus, the regulatory protein activates (or represses) the transcription of specific genes. (c) The activated receptor is cleaved by cellular proteases that cause a carboxy-terminal portion of the receptor to enter the nucleus and interact with specific DNA-binding proteins. The resulting protein complex activates the transcription of specific genes.

### Gradients of Secreted Signaling Molecules Can Instruct Cells to Follow Different Pathways of Development Based on Their Location

A recurring theme in development is the importance of a cell's position within a developing embryo or organ in determining what it will become. Cells located at the front of a fruit fly embryo (i.e., in **anterior** regions) will form portions of the adult head such as the antenna or brain but will not develop into **posterior** structures such as the abdomen or genitalia. Cells located on the top, or **dorsal**, surface of a frog embryo can develop into portions of the backbone in the tadpole or adult but do not form **ventral**, or "belly," tissues such as the gut. These examples illustrate the fact that the fate of a cell—what it will become in the adult—is constrained by its location in the developing embryo. The influence of location on development is called **positional information**.

The most common way of establishing positional information involves a simple extension of one of the strategies we have already encountered in Chapter 19—the use of secreted signaling molecules (Fig. 21-4). A small group of cells synthesizes and secretes a signaling molecule that becomes distributed in an **extracellular gradient** (Fig. 21-4a). Cells located near the "source" receive high concentrations of the secreted protein and develop into a particular cell type. Those cells located at progressively farther distances follow different pathways of development as a result of receiving lower



**FIGURE 21-4** A cluster of cells produces a signaling molecule, or morphogen, that diffuses through the extracellular matrix. (a) Cells 1, 2, and 3 receive progressively lower amounts of the signaling molecule because they are located progressively farther from the source. (b) Cells 1, 2, and 3 contain progressively lower numbers of activated surface receptors. (c) The three cells contain different levels of one or more regulatory proteins. In the simplest scenario, there is a linear correlation between the number of activated cell-surface receptors and the numbers of activated regulatory proteins in the nucleus. (d) The different levels of the regulatory factor lead to the expression of different sets of genes. Cell 1 expresses genes A, B, and C because it contains the highest levels of the regulatory factor. Cell 2 expresses genes B and C, but not A, because it contains intermediate levels of the regulatory factor. These levels are not sufficient to activate gene A. Finally, cell 3 contains the lowest levels of the regulatory factor and expresses only gene C because expression of genes A and B requires higher levels.

concentrations of the signaling molecule. Signaling molecules that control position information are sometimes called **morphogens**.

Cells located near the source of the morphogen receive high concentrations of the signaling molecule and therefore experience peak activation of the specific cell-surface receptors that bind it. In contrast, cells located far from the source receive low levels of the signal, and, consequently, only a small fraction of their cell-surface receptors are activated. Consider a row of three cells adjacent to a source of a secreted morphogen. Something like 1000 receptors are activated in the first cell, whereas only 500 receptors are activated in the next cell, and just 200 in the next (Fig. 21-4b). These different levels of receptor occupancy are directly responsible for differential gene expression in the responding cells.

As we have seen, binding of signaling molecules to cell-surface receptors leads (in one way or another) to an increase in the concentration of specific transcriptional regulators, in an active form, in the nucleus of the cell. Each receptor controls a specific transcriptional regulator (or regulators), and this controls expression of particular genes. The number of cell-surface receptors that are activated by the binding of a morphogen determines how many molecules of the particular regulatory protein appear in the nucleus. The cell closest to the morphogen source—containing 1000 activated receptors—will possess high concentrations of the transcriptional activator in its nucleus (Fig. 21-4c). In contrast, the cells located farther from the source contain intermediate and low levels of the activator, respectively. Thus, there is a correlation between the number of activated receptors on the cell surface and the amount of transcriptional regulator present in the nucleus. How are these different levels of the same transcriptional regulator able to trigger different patterns of gene expression in these different cells?

In Chapter 18, we learned that a small change in the level of the  $\lambda$  repressor determines whether an infected bacterial cell is lysed or lysogenized. Similarly, small changes in the amount of morphogen, and hence small differences in the levels of a transcriptional regulator within the nucleus, determine cell identity. Cells that contain high concentrations of a given transcriptional regulator express a variety of target genes that are inactive in cells containing intermediate or low levels of the regulator (Fig. 21-4d). The differential regulation of gene expression by different concentrations of a regulatory protein is one of the most important and pervasive mechanisms encountered in developmental biology. We consider several examples in the course of this chapter.

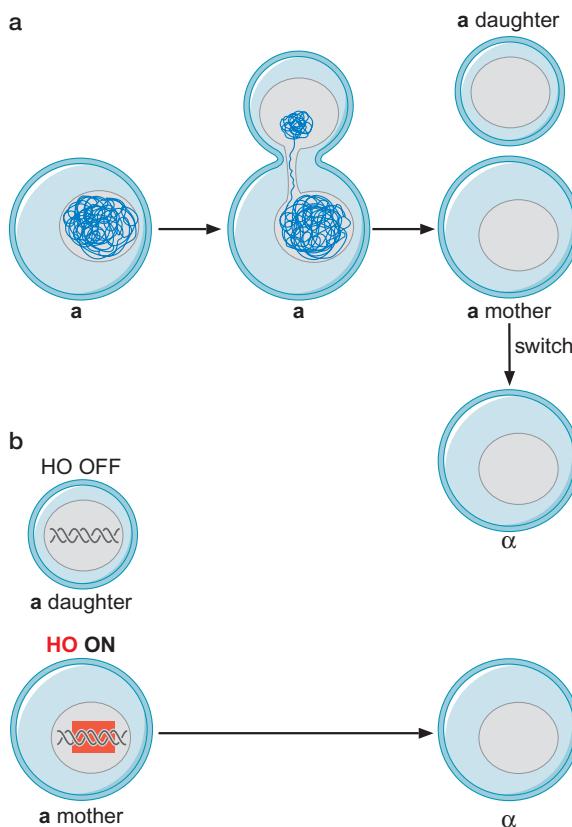
## EXAMPLES OF THE THREE STRATEGIES FOR ESTABLISHING DIFFERENTIAL GENE EXPRESSION

---

### The Localized Ash1 Repressor Controls Mating Type in Yeast by Silencing the *HO* Gene

Before describing mRNA localization in animal embryos, we first consider a case from a relatively simple single-cell eukaryote, the yeast *Saccharomyces cerevisiae*. This yeast can grow as haploid cells that divide by budding (Fig. 21-5). Replicated chromosomes are distributed between two asymmetric cells—the larger progenitor cell, or mother cell, and a smaller bud, or daughter cell (Fig. 21-5a). These cells can exist as either of two mating types, called  $\alpha$  and  $\alpha$ , as discussed in Chapters 11 and 19.

A mother cell and its daughter cell can show different mating types. This difference arises by a process called **mating-type switching**. After budding



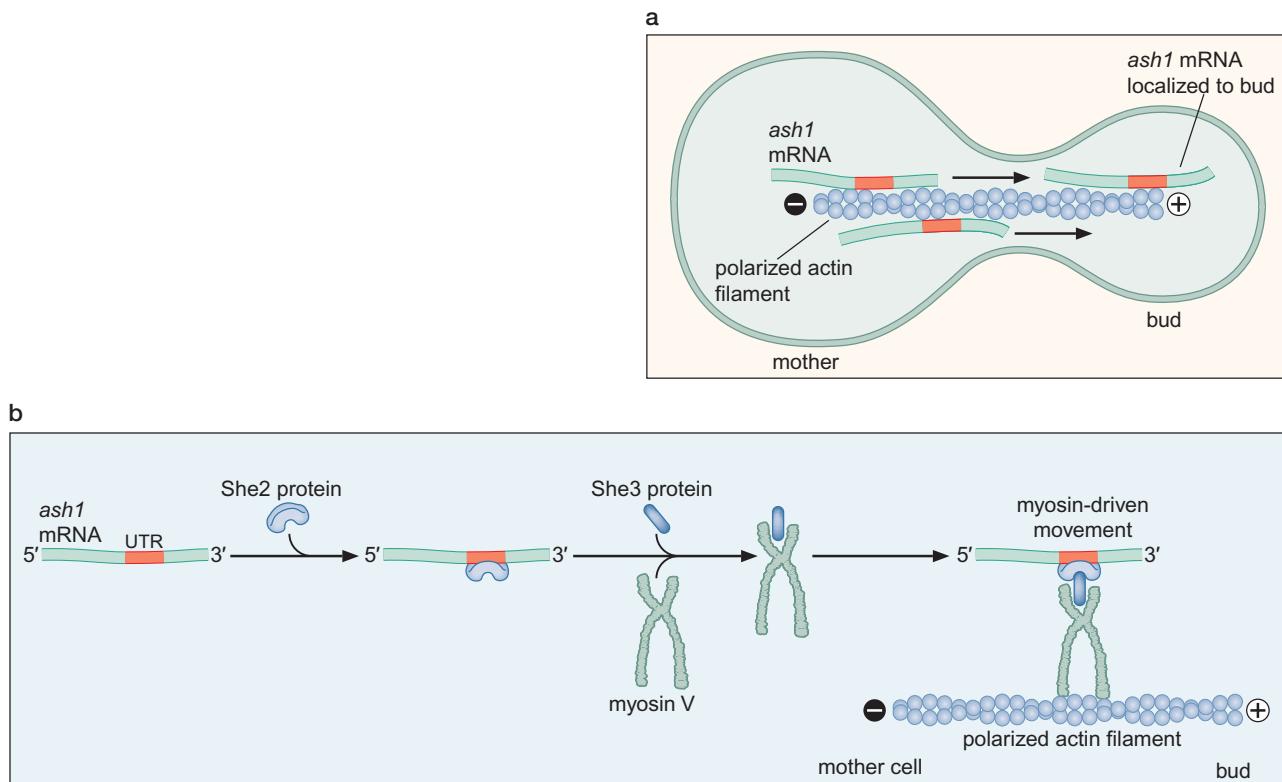
**FIGURE 21-5** A haploid yeast cell of mating type *a* undergoes budding to produce a mother cell and a smaller daughter cell. (a) Initially, both cells are mating type *a*, but sometimes the mother cell can undergo switching to the  $\alpha$  type. (b) Because of the localized *Ash1* transcriptional repressor, the daughter cell is unable to express the *HO* gene and thus cannot undergo switching. In contrast, the mother cell can switch because it lacks *Ash1* and is able to express *HO*.

to produce a daughter, a mother cell can “switch” mating type with, for example, an *a* cell giving rise to an *a* daughter, but subsequently switching to the  $\alpha$  mating type (Fig. 21-5b).

Switching is controlled by the product of the *HO* gene. We saw in Chapter 11 that the *HO* protein is a sequence-specific endonuclease. *HO* triggers gene conversion within the mating-type locus by creating a double-strand break at one of the two silent mating-type cassettes. We also saw in Chapter 19 how *HO* is activated in the mother cell. It is kept silent in the daughter cell because of the selective expression of a repressor called *Ash1* (Fig. 21-6), and this is why the daughter cell does not switch mating type. The *ash1* gene is transcribed in the mother cell before budding, but the encoded RNA becomes localized within the daughter cell through the following process. During budding, the *ash1* mRNA attaches to the growing ends of microtubules. Several proteins function as “adaptors” that bind the 3' UTR of the *ash1* mRNA and also to the microtubules. The microtubules extend from the nucleus of the mother cell to the site of budding, and in this way, the *ash1* mRNA is transported to the daughter cell. Once localized within the daughter cell, the *ash1* mRNA is translated into a repressor protein that binds to, and inhibits the transcription of, the *HO* gene. This silencing of *HO* expression in the daughter cell prevents that cell from undergoing mating-type switching.

In the second half of this chapter, we will see the localization of mRNAs used in the development of the *Drosophila* embryo. Once again, this localization is mediated by adaptor proteins that bind to the mRNAs, specifically, to sequences found in their 3' UTRs (see Box 21-2, Review of Cytoskeleton: Asymmetry and Growth).

A second general principle that emerges from studies on yeast mating-type switching is seen again when we consider *Drosophila* development:



**FIGURE 21-6** Localization of *ash1* mRNA during budding. (a) The *ash1* gene is transcribed in the mother cell during budding. The encoded mRNA moves from the mother cell into the bud by sliding along polarized actin filaments. Movement is directed and begins at the “−” ends of the filament and extends with the growing “+” ends. (b) The *ash1* mRNA transport depends on the binding of the She2 and She3 adaptor proteins to specific sequences contained within the 3' UTR. These adaptor proteins bind myosin, which “crawls” along the actin filament and brings the *ash1* mRNA along for the ride. (Adapted, with permission, from Alberts B. et al. 2002. *Molecular biology of the cell*, 4th ed., p. 971, Fig. 16-84a. © Garland Science/Taylor & Francis LLC.)

the interplay between broadly distributed activators and localized repressors to establish precise patterns of gene expression within individual cells. In yeast, the SWI5 protein is responsible for activating expression of the *HO* gene (see Chapter 19). This activator is present in both the mother cell and the daughter cell during budding, but its ability to turn on *HO* is restricted to the mother cell because of the presence of the Ash1 repressor in the daughter cell. In other words, Ash1 keeps the *HO* gene off in the daughter cell despite the presence of SWI5.

### A Localized mRNA Initiates Muscle Differentiation in the Sea Squirt Embryo

Localized mRNAs can establish differential gene expression among the genetically identical cells of a developing embryo. Just as the fate of the daughter cell is constrained by its inheritance of the *ash1* mRNA in yeast, the cells in a developing embryo can be instructed to follow specific pathways of development through the inheritance of localized mRNAs.

As an example, we consider muscle differentiation in sea squirts. Macho-1 is a major determinant for programming cells to form tail muscles in early sea squirt embryos.

Macho-1 mRNA is initially distributed throughout the cytoplasm of unfertilized eggs but becomes restricted to the vegetal (bottom) cytoplasm shortly after fertilization (Fig. 21-7). It is ultimately inherited by just two of the cells in eight-cell embryos, and as a result these two cells go on to form the tail muscles.

The Macho-1 mRNA encodes a zinc finger DNA-binding protein that is believed to activate the transcription of muscle-specific genes, such as tropomysin. Thus, these genes are expressed only in muscles because Macho-1 is made only in those cells. In the second part of this chapter, we see how regulatory proteins synthesized from localized mRNAs in the *Drosophila* embryo activate and repress gene expression and control the formation of different cell types.

#### ► ADVANCED CONCEPTS

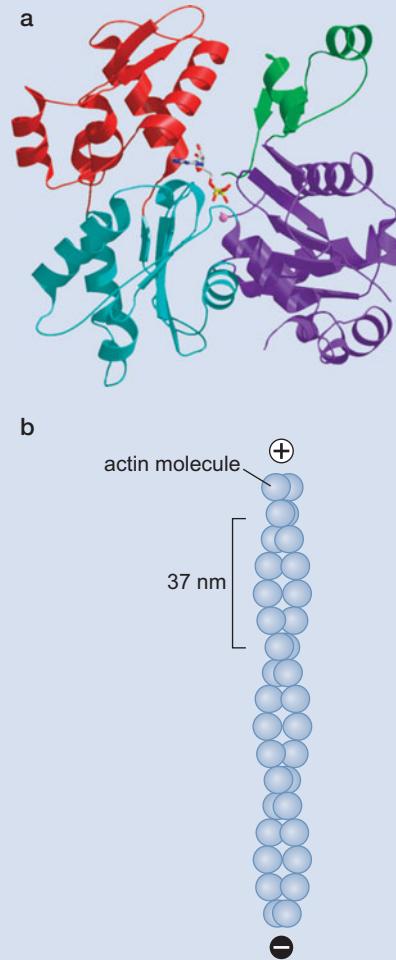
##### Box 21-2 Review of Cytoskeleton: Asymmetry and Growth

The cytoskeleton is composed of three types of filaments: intermediate filaments, actin filaments, and microtubules. Actin filaments and microtubules are used to localize specific mRNAs in a variety of different cell types, including budding yeast and *Drosophila* oocytes. Actin filaments are composed of polymers of actin. The actin polymers are organized as two parallel helices that form a complete twist every 37 nm. Each actin monomer is located in the same orientation within the polymer, and as a result, actin filaments contain a clear polarity. The plus (+) end grows more rapidly than the minus (-) end, and consequently, mRNAs slated for localization move along with the growing + end (Box 21-2 Fig. 1).

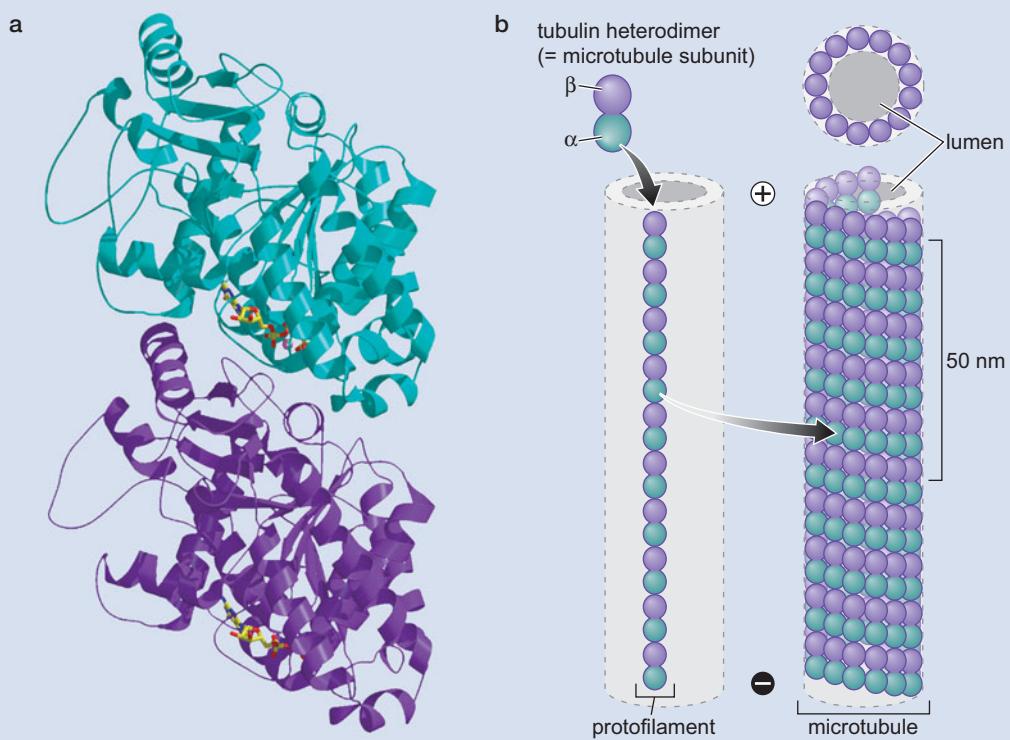
Microtubules are composed of polymers of a protein called **tubulin**, which is a heterodimer composed of related  $\alpha$  and  $\beta$  chains. Tubulin heterodimers form extended, asymmetric protofilaments. Each tubulin heterodimer is located in the same orientation within the protofilament. Thirteen different protofilaments associate to form a cylindrical microtubule, and all of the protofilaments are aligned in parallel. Thus, as seen for actin filaments, there is an intrinsic polarity in microtubules, with a rapidly growing "+" end and more stable "-" end (Box 21-2 Fig. 2).

Both actin and tubulin function as enzymes. Actin catalyzes the hydrolysis of ATP to ADP, whereas tubulin hydrolyzes GTP to GDP. These enzymatic activities are responsible for the dynamic growth, or "treadmilling," seen for actin filaments and microtubules. Typically, it is the actin or tubulin subunits at the "-" end of the filament that mediate the hydrolysis of ATP or GTP, and as a result, these subunits are somewhat unstable and lost from the "-" end. In contrast, newly added subunits at the "+" end have not hydrolyzed ATP or GTP, and this causes them to be more stable components of the filament.

Directed growth of actin filaments or microtubules at the "+" ends depends on a variety of proteins that associate with the cytoskeleton. One such protein is called profilin, which interacts with actin monomers and augments their incorporation into the "+" ends of growing actin filaments. Other proteins have been shown to enhance the growth of tubulin protofilaments at the "+" ends of microtubules.

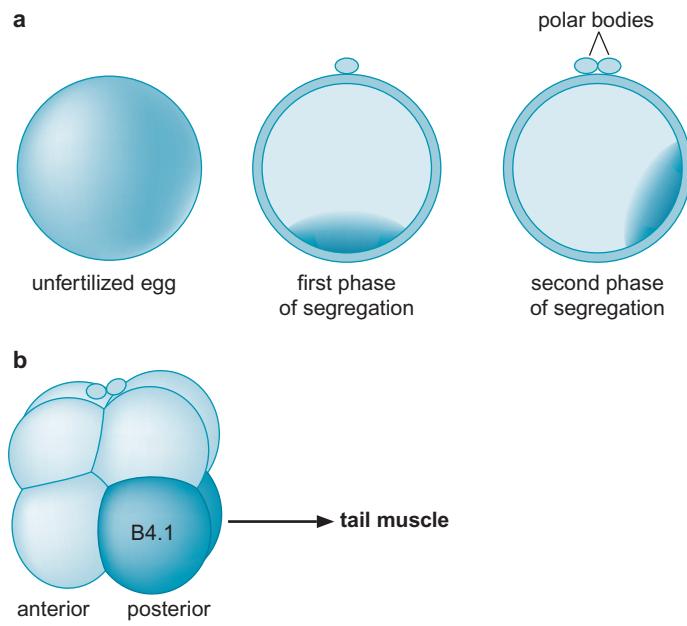


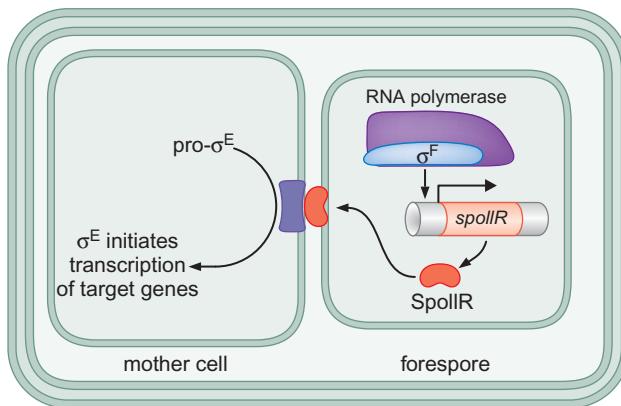
**BOX 21-2 FIGURE 1** Structures of the actin monomer and filament. Crystal structure of the actin monomer. (a) The four domains of the monomer are shown, in different colors, with ATP (in red and yellow) in the center. The "-" end of the monomer is at the top, and the "+" end is at the bottom. (Otterbein L.R. et al. 2001. *Science* 293: 708–711. Image prepared with MolScript, BobScript, and Raster3D.) (b) The monomers are assembled, as a single helix, into a filament.

**Box 21-2 (Continued)**

**BOX 21-2 FIGURE 2** Structures of the tubulin monomer and filament. (a) The crystal structure of the tubulin monomer shows the  $\alpha$  subunit (turquoise) and the  $\beta$  subunit (purple). The GTP molecules in each subunit are red and yellow. (From Lowe J. et al. 2001. *J. Mol. Biol.* 313: 1045–1057. Image prepared with MolScript, BobScript, and Raster3D.) (b) The protofilament of tubulin consists of adjacent monomers assembled in the same orientation.

**FIGURE 21-7** The Macho-1 mRNA becomes localized in the fertilized egg of a sea squirt. (a) The mRNA is initially distributed throughout the cytoplasm of unfertilized eggs. At fertilization, the egg is induced to undergo a highly asymmetric division to produce a small polar body (top). At this time, the Macho-1 mRNA becomes localized to bottom (vegetal) regions. Shortly thereafter, and well before the first division of the one-cell embryo, the Macho-1 mRNA undergoes a second wave of localization. This occurs during the second highly asymmetric meiotic division of the egg. (b) The Macho-1 mRNA becomes localized to a specific quadrant of the one-cell embryo that corresponds to the future B4.1 blastomeres. These are the cells that generate the tail muscles. (a, Adapted, with permission, from Nishida H. and Sawada K. 2001. *Nature* 409: 725, Fig. 1c–e. © Macmillan.)





**FIGURE 21-8** Asymmetric gene activity in the mother cell and forespore of *B. subtilis* depends on the activation of different classes of  $\sigma$  factors. The *spoII R* gene is activated by  $\sigma^F$  in the forespore. The encoded SpoII R protein becomes associated with the septum separating the mother cell (on the left) and forespore (on the right). It triggers the proteolytic processing of an inactive form of  $\sigma^E$  (pro- $\sigma^E$ ) in the mother cell. The activated  $\sigma^E$  protein leads to the recruitment of RNA polymerase and the activation of specific genes in the mother cell. (Redrawn, with permission, from Stragier P. and Losick R. 1996. *Annu. Rev. Genet.* **30**: 297–341, Fig. 3a. © Annual Reviews.)

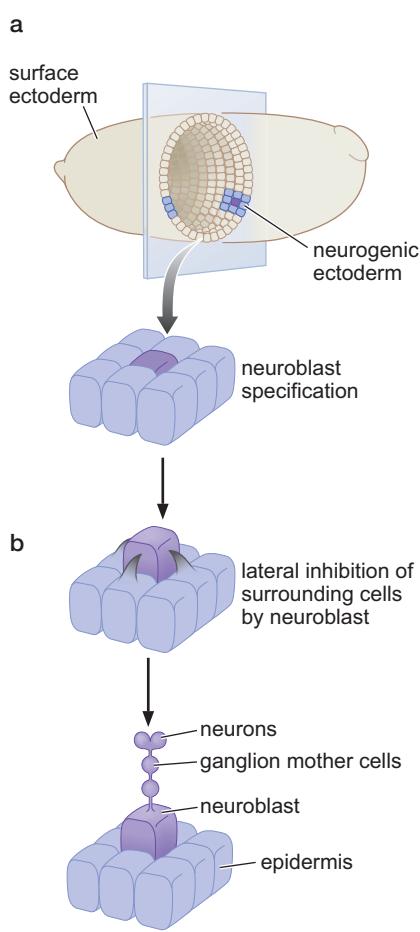
### Cell-to-Cell Contact Elicits Differential Gene Expression in the Sporulating Bacterium, *Bacillus subtilis*

The second major strategy for establishing differential gene expression is cell-to-cell contact. Again, we begin our discussion with a relatively simple case, this one from the bacterium *Bacillus subtilis*. Under adverse conditions, *B. subtilis* can form spores. The first step in this process is the formation of a septum at an asymmetric location within the sporangium, the progenitor of the spore. The septum produces two cells of differing sizes that remain attached through abutting membranes. The smaller cell is called the **forespore**; it ultimately forms the spore. The larger cell is called the mother cell; it aids the development of the spore (Fig. 21-8). The forespore influences the expression of genes in the neighboring mother cell, as described later.

The forespore contains an active form of a specific  $\sigma$  factor,  $\sigma^F$ , that is inactive in the mother cell. In Chapter 18, we saw how  $\sigma$  factors associate with RNA polymerase and select specific target promoters for expression.  $\sigma^F$  activates the *spoII R* gene, which encodes a secreted signaling protein. SpoII R is secreted into the space between the abutting membranes of the mother cell and the forespore, where it triggers the proteolytic processing of pro- $\sigma^E$  in the mother cell. Pro- $\sigma^E$  is an inactive precursor of the  $\sigma^E$  factor. The pro- $\sigma^E$  protein contains an amino-terminal inhibitory domain that blocks  $\sigma^E$  activity and tethers the protein to the membrane of the mother cell (Fig. 21-8). SpoII R induces the proteolytic cleavage of the amino-terminal peptide and the release of the mature and active form of  $\sigma^E$  from the membrane.  $\sigma^E$  activates a set of genes in the mother cell that is distinct from those expressed in the forespore. In this example, SpoII R functions as a signaling molecule that acts at the interface between the forespore and the mother cell and elicits differential gene expression in the abutting mother cell through the processing of  $\sigma^E$ . Induction requires cell-to-cell contact because the forespore produces small quantities of SpoII R that can interact with the abutting mother cell but are insufficient to elicit the processing of  $\sigma^E$  in the other cells of the population.

### A Skin–Nerve Regulatory Switch Is Controlled by Notch Signaling in the Insect Central Nervous System

We now turn to an example of cell-to-cell contact in an animal embryo that is surprisingly similar to the one just described in *B. subtilis*. In that earlier example, SpoII R causes the proteolytic activation of  $\sigma^E$ , which, in its active state, directs RNA polymerase to the promoter sequences of specific genes. In the following example, a cell-surface receptor is cleaved, and the intracytoplasmic



**FIGURE 21-9** The neurogenic ectoderm forms two major cell types: neurons and skin cells (or epidermis). (a) Cells in the early neurogenic ectoderm can form either type of cell. However, once one of the cells begins to form a neuron or “neuroblast” (dark cell in the center of the grid of cells), it inhibits all of the neighboring cells that it directly touches. (b) This inhibition causes most of the cells to remain on the surface of the embryo and form skin cells. In contrast, the developing neuron moves into the embryo cavity and forms neurons.

domain moves to the nucleus, where it binds a sequence-specific DNA-binding protein that activates the transcription of selected genes.

For this example, we must first briefly describe the development of the ventral nerve cord in insect embryos (Fig. 21-9). This nerve cord functions in a manner that is roughly comparable to the spinal cord of humans. It arises from a sheet of cells called the **neurogenic ectoderm**. This tissue is subdivided into two cell populations: one group remains on the surface of the embryo and forms ventral skin (or epidermis), whereas the other population moves inside the embryo to form the neurons of the ventral nerve cord (Fig. 21-9a). This decision whether to become skin or neuron is reinforced by signaling between the two populations.

The developing neurons contain a signaling molecule on their surface called **Delta**, which binds to a receptor on the skin cells called **Notch** (Fig. 21-9b). The activation of the Notch receptor on skin cells by Delta renders them incapable of developing into neurons, as follows. Activation causes the intracytoplasmic domain of Notch ( $\text{Notch}^{\text{IC}}$ ) to be released from the cell membrane and enter nuclei, where it associates with a DNA-binding protein called **Su(H)**. The resulting **Su(H)**– $\text{Notch}^{\text{IC}}$  complex activates genes that encode transcriptional repressors that block the development of neurons.

Notch signaling does not cause a simple induction of the Su(H) activator protein but instead triggers an on/off regulatory switch. In the absence of signaling, Su(H) is associated with several corepressor proteins, including Hairless, CtBP, and Groucho (Fig. 21-10). Su(H) complexed with any of these proteins actively represses Notch target genes. When  $\text{Notch}^{\text{IC}}$  enters the nucleus, it displaces the repressor proteins in complex with Su(H), turning that protein into an activator instead. Thus, Su(H) now activates the very same genes that it formerly repressed.

Delta–Notch signaling depends on cell-to-cell contact. The cells that present the Delta ligand (neuronal precursors) must be in direct physical contact with the cells that contain the Notch receptor (epidermis) in order to activate Notch signaling and inhibit neuronal differentiation. In the next section, we see an example of a secreted signaling molecule that influences gene expression in cells located far from those that send the signal.

### A Gradient of the Sonic Hedgehog Morphogen Controls the Formation of Different Neurons in the Vertebrate Neural Tube

We now turn to an example of a long-range signaling molecule, a **morphogen**, that imposes positional information on a developing organ. For this example, we continue our discussion of neuronal differentiation, but this time, we consider the neural tube of vertebrates. In all vertebrate embryos, there is a stage when cells located along the future back—the dorsal ectoderm—move in a coordinated fashion toward internal regions of the embryo and form the neural tube, the forerunner of the adult spinal cord.

Cells located in the ventralmost region of the neural tube form a specialized structure called the **floorplate**. The floorplate is the site of expression of a secreted cell-signaling molecule called **Sonic hedgehog (Shh)**, which is thought to function as a gradient morphogen.

Shh is secreted from the floorplate and forms an extracellular gradient in the ventral half of the neural tube. Neurons develop within the neural tube into different cell types based on the amount of Shh protein they receive. This is determined by their location relative to the floorplate; cells located near the floorplate receive the highest concentrations of Shh, and those located farther away receive lower levels. The extracellular Shh gradient leads to the specification of three neuronal cell types: V3, MN, and V2. These cells are located

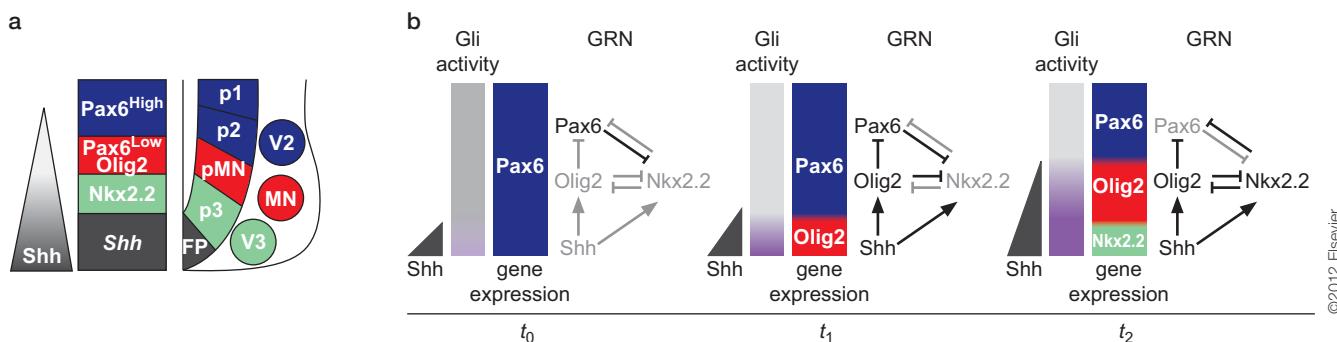
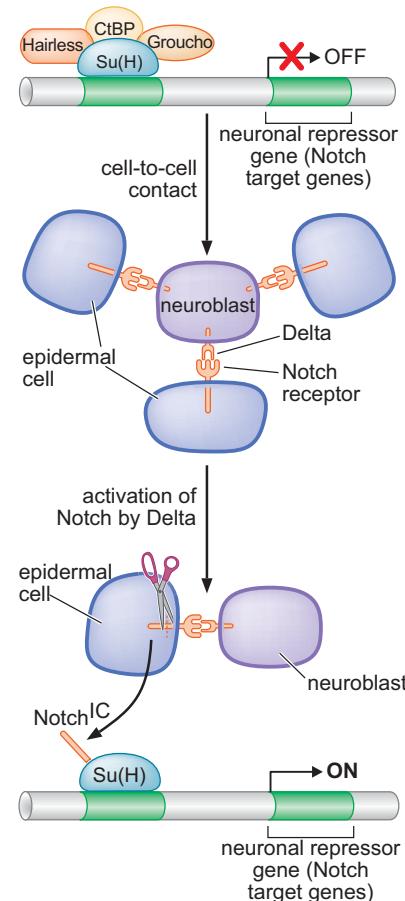
**FIGURE 21-10 Notch–Su(H) regulatory switch.** The developing neuron (neuroblast) does not express neuronal repressor genes (top). These genes are kept off by a DNA-binding protein called Su(H) and associated corepressor proteins (Hairless, CtBP, Groucho). The neuroblast expresses a signaling molecule, called Delta, that is tethered to the cell surface. Delta binds to the Notch receptor in neighboring cells that are in direct physical contact with the neuron. Delta–Notch interactions cause the Notch receptor to be activated in the neighboring cells, which differentiate into epidermis. The activated Notch receptor is cleaved by cellular proteases (scissors), and the intracytoplasmic region of the receptor is released into the nucleus. This piece of the Notch protein causes the Su(H) regulatory protein to function as an activator rather than a repressor. As a result, the neuronal repressor genes are activated in the epidermal cells so that they cannot develop into neurons.

progressively farther from the floorplate and differentially express three regulatory genes: Nkx2.2, Olig2, and Pax6, respectively (see Fig. 21-11a).

Initially, Pax6 is expressed throughout the presumptive neural tube (time  $t_0$ , Fig. 21-11b). Cells located near the floorplate—those receiving the highest concentrations of Shh—acquire the highest activity of Gli, the transcriptional effector of Shh signaling. At early stages, time  $t_1$ , the initial concentrations of the Gli activator are sufficient to induce Olig2 expression (Fig. 21-11b). At subsequent stages, sustained Shh induction raises the levels of Gli in the ventral neural tube, leading to the activation of Nkx2.2 (time  $t_2$ , Fig. 21-11b). Cross-repressive interactions maintain sequential expression of Nkx2.2, Olig2, and Pax6, leading to the specification of the V3, MN, and V2 neurons.

According to a simple “gradient affinity” model, the regulatory DNAs of the Olig2 and Nkx2.2 genes might be expected to contain Gli-binding sites with differing affinities. For example, Olig2 might be activated before Nkx2.2 because it contains high-affinity Gli-binding sites that are occupied by low levels of the Gli activator. In contrast, Nkx2.2 might be regulated by low-affinity Gli-binding sites, requiring higher, sustained levels of Shh and the Gli activator.

Recent studies suggest an alternative view: namely, differential expression of Olig2 and Nkx2.2 is controlled by a network of gene interactions underlying the patterning of the neural tube (see Fig. 21-11b). Once Olig2 is activated in ventral regions, it represses Pax6, thereby creating a “window” for the induction of Nkx2.2. Pax6 is a potent repressor of Nkx2.2 but not Olig2. Differential repression by Pax6 might be a critical mechanism for sequential expression of Olig2 and Nkx2.2. Perhaps Nkx2.2 regulatory DNAs contain Pax6 repressor binding sites, whereas Olig2 regulatory sequences lack such sites.



**FIGURE 21-11 The extracellular Shh gradient leads to the specification of three neuronal cell types.** (a) Shh forms a gradient in the neural tube. (b) Model for Shh signal-mediated patterning as development progresses. (Adapted, with permission, from Balaskas N. et al. 2012. *Cell* **148**: 273–284; part a is Fig. 1A, p. 274; part b is Fig. 7A, p. 281. © Elsevier.)

## THE MOLECULAR BIOLOGY OF *DROSOPHILA* EMBRYOGENESIS



In this section, we focus on the early embryonic development of the fruit fly, *Drosophila melanogaster* (see Interactive Animation 21-1). The molecular details of how development is regulated are better understood in this system than in any other animal embryo. The various mechanisms of cell communication discussed in the first half of this chapter and those of gene regulation discussed in the previous chapters are brought together in this example.

Localized mRNAs and cell-signaling pathways are both used to establish positional information that results in gradients of regulatory proteins that pattern the anteroposterior (head–tail) and dorsoventral (back–belly) body axes. These regulatory proteins—activators and repressors—control the expression of genes whose products define different regions of the embryo. A recurring theme is the use of complex regulatory DNAs—particularly complex enhancers—for the combinatorial control of sharp on/off patterns of gene expression.

### An Overview of *Drosophila* Embryogenesis

Life begins for the fruit fly as it does for humans: adult males inseminate females. A single sperm cell enters a mature egg, and the haploid sperm and egg nuclei fuse to form a diploid, “zygotic” nucleus. This nucleus undergoes a series of nearly synchronous divisions within the central regions of the egg. Because there are no plasma membranes separating the nuclei, the embryo now becomes what is called a **syncytium**—that is, a single cell with multiple nuclei. With the next series of divisions, the nuclei begin to migrate toward the cortex, or periphery, of the egg. Once located in the cortex, the nuclei undergo another three divisions leading to the formation of a monolayer of approximately 6000 nuclei surrounding the central yolk. During a 1-h period, from 2 to 3 h after fertilization, cell membranes form between adjacent nuclei.

The rapid nuclear divisions that occur during the first 2 h of *Drosophila* embryogenesis preclude precocious expression of critical patterning genes. Consider the short gastrulation (*sog*) gene as an example. The *sog* gene encodes an inhibitor of BMP signaling that is important for the patterning of the dorsal ectoderm during 2.5–3 h after fertilization. The *sog* transcription unit is 20 kb in length. RNA polymerase II (Pol II) has a remarkably slow rate of elongation—just 20 bp per second. As a result, it takes nearly 20 min for “tip-to-toe” transcription of *sog* and the synthesis of full-length, mature mRNAs. The first 11 rounds of nuclear divisions occur at a frequency of just 6–8 min, and consequently, there is no time for Pol II to complete transcription of the *sog* gene during the brief interphase periods of these division cycles. During mitosis, Pol II is released from the chromatin template and must reinitiate transcription at the onset of the subsequent division cycle. As a result, no meaningful *sog* mRNAs can be synthesized during the first 2 h of embryogenesis. In this example, the size of the *sog* transcription unit helps ensure that *sog* products are not synthesized until they are needed, at 2.5 h after fertilization. There are additional examples of gene size and large introns in the timing of gene expression during *Drosophila* embryogenesis. For example, the homeotic gene *Ubx* is >80 kb in length, and its expression is delayed for more than an hour relative to coexpressed segmentation genes containing small introns.

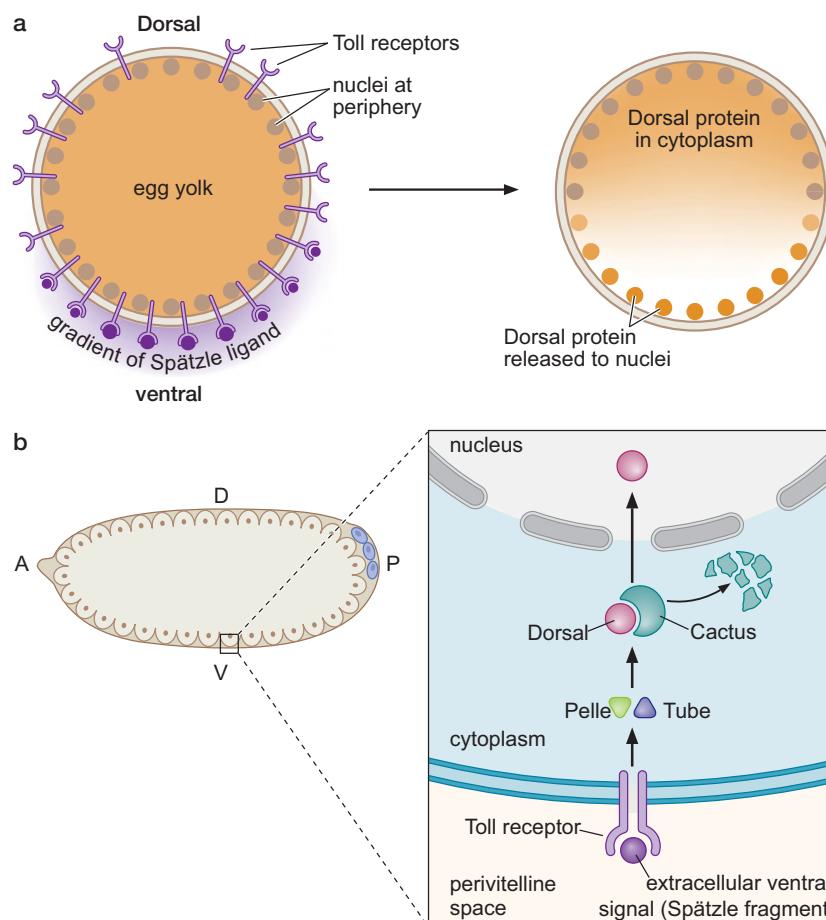
Before the formation of cell membranes, the nuclei are **totipotent** or uncommitted: they have not yet taken on an identity and can still give rise to any cell type. Just after cellularization, however, nuclei have become

irreversibly “**determined**” to differentiate into specific tissues in the adult fly. This process is described in Box 21-3, Overview of *Drosophila* Development. The molecular mechanisms responsible for this dramatic process of determination are described in the following sections of this chapter.

### A Regulatory Gradient Controls Dorsoventral Patterning of the *Drosophila* Embryo

The dorsoventral patterning of the early *Drosophila* embryo is controlled by a regulatory protein called Dorsal, which is initially distributed throughout the cytoplasm of the unfertilized egg. After fertilization, and after the nuclei reach the cortex of the embryo, the Dorsal protein enters nuclei in the ventral and lateral regions but remains in the cytoplasm in dorsal regions (Fig. 21-12). The formation of this Dorsal gradient in nuclei across the embryo is very similar, in principle, to the formation of the Gli activator gradient within the vertebrate neural tube (see Fig. 21-11).

Regulated nuclear transport of the Dorsal protein is controlled by a cell-signaling molecule called **Spätzle**. This signal is distributed in a ventral-to-dorsal gradient within the extracellular matrix present between the plasma membrane of the unfertilized egg and the outer egg shell. After fertilization, Spätzle binds to the cell-surface Toll receptor. Depending on the concentration of Spätzle, and thus the degree of receptor occupancy in a given region of the syncytial embryo, Toll is activated to a greater or lesser extent. There is peak activation of Toll receptors in ventral regions—where the Spätzle concentration is highest—and progressively lower activation in



**FIGURE 21-12** Spätzle-Toll and Dorsal gradient. (a) The circles represent cross sections through early *Drosophila* embryos. The Toll receptor is uniformly distributed throughout the plasma membrane of the precellular embryo. The Spätzle signaling molecule is distributed in a gradient with peak levels in the ventralmost regions. As a result, more Toll receptors are activated in ventral regions than in lateral and dorsal regions. This gradient in Toll signaling creates a broad Dorsal nuclear gradient. (b) Side view of the embryo with anterior to the left and dorsal surface up; details of the Toll signaling cascade to the right. Activation of the Toll receptor leads to the activation of the Pelle kinase in the cytoplasm. Pelle either directly or indirectly phosphorylates the Cactus protein, which binds and inhibits the Dorsal protein. Phosphorylation of Cactus causes its degradation, so that Dorsal is released from the cytoplasm into nuclei.

## ► ADVANCED CONCEPTS

**Box 21-3 Overview of *Drosophila* Development**

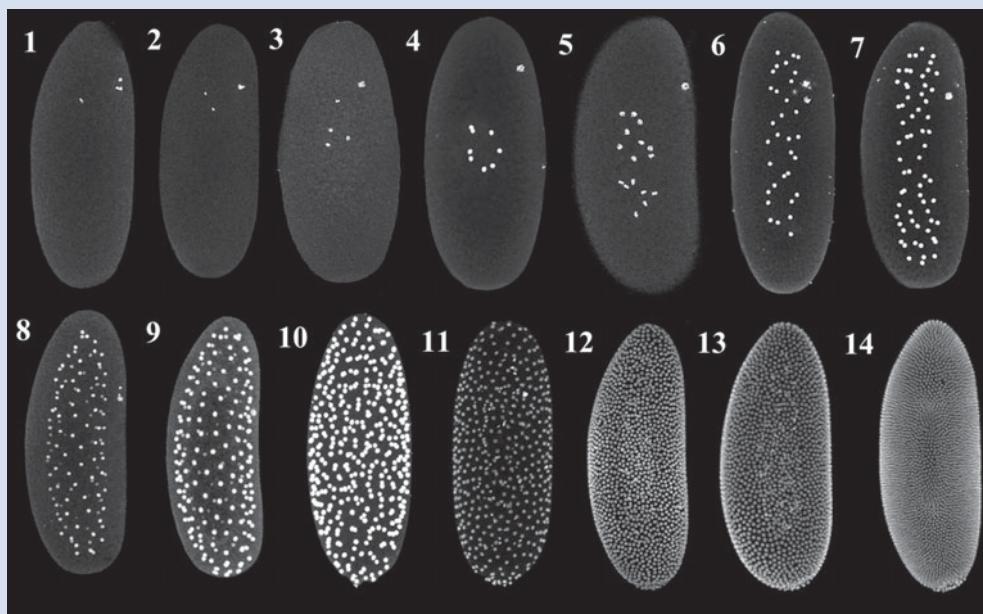
After the sperm and egg haploid nuclei fuse, the diploid, zygotic nucleus undergoes a series of 10 rapid and nearly synchronous cleavages within the central yolk regions of the egg. Large microtubule arrays emanating from the centrioles of the dividing nuclei help direct the nuclei from central regions toward the periphery of the egg (Box 21-3 Fig. 1). After eight cleavages, the 256 zygotic nuclei begin to migrate to the periphery. During this migration, they undergo two more cleavages (Box 21-3 Fig. 1, nuclear cleavage cycle 9). Most, but not all, of the resulting approximately 1000 nuclei enter the cortical regions of the egg (Box 21-3 Fig. 1, nuclear cleavage cycle 10). The others ("vitellophages") remain in central regions, where they have a somewhat obscure role in development.

Once the majority of the nuclei reach the cortex at  $\sim$ 90 min following fertilization, they first acquire competence to transcribe Pol II genes. Thus, as in many other organisms such as *Xenopus*, there seems to be a "midblastula transition," whereby early blastomeres (or nuclei) are transcriptionally silent during rapid periods of mitosis. Although causality is unclear, it does seem that DNA undergoing intense bursts of replication cannot simultaneously sustain transcription. These and other observations have led to the suggestion that there is competition between the large macromolecular complexes promoting replication and transcription.

After the nuclei reach the cortex, they undergo another three rounds of cleavage (for a total of 13 divisions after fertilization), leading to the dense packing of about 6000 columnar-shaped nuclei enclosing the central yolk (Box 21-3 Fig. 1, nuclear cleavage cycle 14). Technically, the embryo is still a syncytium, although histochemical staining of early embryos with antibodies against cytoskeletal proteins indicates a highly structured meshwork surrounding each nucleus. During a 1-h period, from 2 to 3 h after fertilization, the embryo undergoes a dramatic cellularization process, whereby cell membranes are formed between adjacent nuclei (Box 21-3 Fig. 1, nuclear cleavage cycle 14). By 3 h after fertilization, the embryo has been transformed into a cellular blastoderm, comparable to the "hollow ball of cells" that characterizes the blastulae of most other embryos.

The recently developed SPIM (single plane illumination microscopy) method has been used for the detailed imaging of *Drosophila* embryogenesis. High-resolution movies of early embryogenesis can be found on the following website: <http://www.nature.com/nmeth/journal/vaop/ncurrent/extref/nmeth.2062-sv1.mov>.

When the nuclei enter the cortex of the egg, they are totipotent and can form any adult cell type. The location of each nucleus, however, now determines its fate. The 30 or so nuclei



**BOX 21-3 FIGURE 1** *Drosophila* embryogenesis. *Drosophila* embryos are oriented with the future head pointed up. The numbers refer to the number of nuclear cleavages. Nuclei are stained white within the embryos. For example, stage 1 contains the single zygotic nucleus resulting from the fusion of the sperm and egg pronuclei. The stained material in the upper right areas of stages 1–7 are polar bodies. The zygotic nucleus of stage 1 and the nuclei of stages 2, 3, . . . , are in central regions of the embryo. Stage 2 contains two nuclei arising from the first division of the zygotic nucleus. At stage 10, there are approximately 500 nuclei, and most are arranged in a single layer at the cortex (periphery of the embryo). At nuclear cleavage cycle 14, there are more than 6000 nuclei densely packed in a monolayer in the cortex. Cellularization occurs during this stage. (Courtesy of W. Baker and G. Shubiger.)

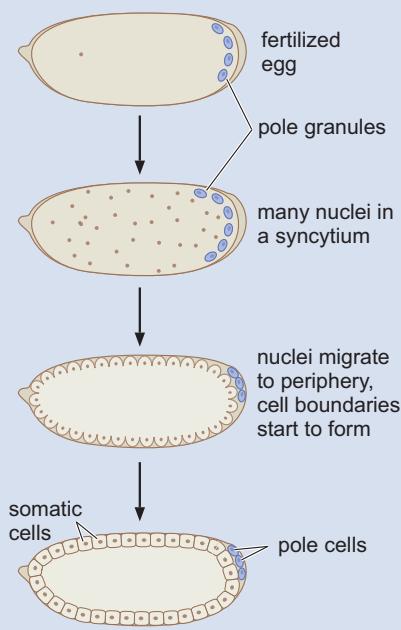
**Box 21-3** (Continued)

that migrate into posterior regions of the cortex encounter localized protein determinants, such as Oskar, which program these naive nuclei to form the germ cells (Box 21-3 Fig. 2). Among the putative determinants contained in the polar plasm are large nucleoprotein complexes, called polar granules. The posterior nuclei bud off from the main body of the embryo along with the polar granules, and the resulting pole cells differentiate into either sperm or eggs, depending on the sex of the embryo. The microinjection of polar plasm into abnormal locations, such as central and anterior regions, results in the differentiation of supernumerary pole cells.

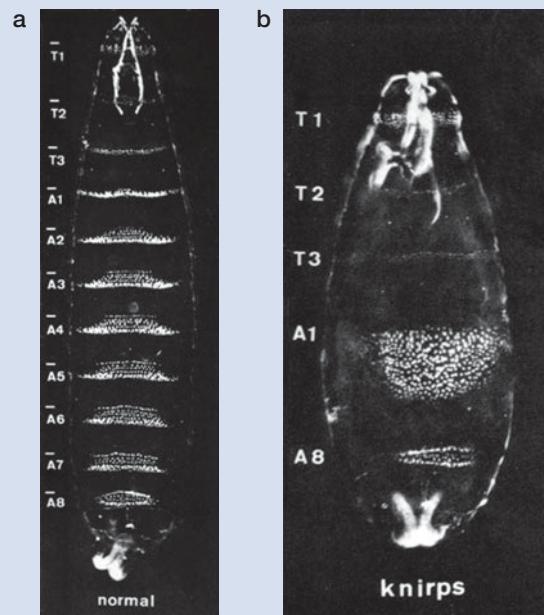
Cortical nuclei that do not enter the polar plasm are destined to form the somatic tissues. Again, these nuclei are totipotent and can form any adult cell type. However, within a very brief period (no more than an hour), each nucleus is rapidly programmed (or specified) to follow a particular pathway of differentiation. This specification process occurs during the period of cellularization, although there is no reason to believe that the deposition of cell membranes between neighboring nuclei is critical for determining cell fate. Different nuclei show distinct patterns of gene transcription before the

completion of cell formation. By 3 h after fertilization, each cell possesses a fixed positional identity, so that cells located in anterior regions of the embryo will form head structures in the adult fly, whereas cells located in posterior regions will form abdominal structures.

Systematic genetic screens by Eric Wieschaus and Christiane Nüsslein-Volhard identified approximately 30 "segmentation genes" that control the early patterning of the *Drosophila* embryo. This involved the examination of thousands of dead embryos. At the midpoint of embryogenesis, the ventral skin, or epidermis, secretes a cuticle that contains many fine hairs, or denticles. Each body segment of the embryo contains a characteristic pattern of denticles. Three different classes of segmentation genes were identified on the basis of causing specific disruptions in the denticle patterns of dead embryos. Mutations in the so-called "gap" genes cause the deletion of several adjacent segments (Box 21-3 Fig. 3). For example, mutations in the gap gene *knirps* cause the loss of the second through seventh abdominal segments (normal embryos possess eight such segments). Mutations in the "pair-rule" genes cause the loss of alternating segments.

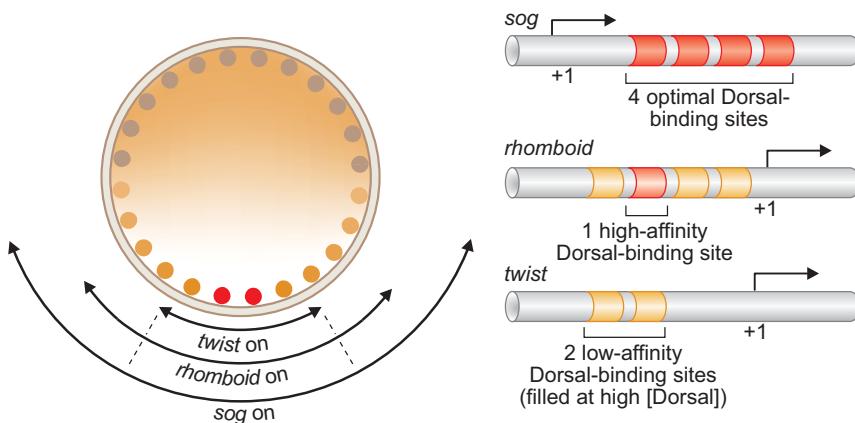


**BOX 21-3 FIGURE 2** Development of germ cells. Polar granules located in the posterior cytoplasm of the unfertilized egg contain germ cell determinants and the Nanos mRNA, which is important for the development of the abdominal segments. Nuclei (central dots) begin to migrate to the periphery. Those that enter posterior regions sequester the polar granules and form the pole cells, which form the germ cells. The remaining cells (somatic cells) form all of the other tissues in the adult fly. (Adapted, with permission, from Schneiderman H.A. 1976. *Symp. R. Entomol. Soc. Lond.* 8: 3–34. © Royal Entomological Society.)



**BOX 21-3 FIGURE 3** Dark-field images of normal and mutant cuticles. (a) The pattern of denticle hairs in this normal embryo is slightly different among the different body segments (labeled T1 through A8 in the image). (b) The Knirps mutant (having a mutation in the gap gene *knirps*), shown here, lacks the second through seventh abdominal segments. (Reprinted, with permission, from Nüsslein-Volhard C. and Wieschaus E. 1980. *Nature* 287: 795–801. © Macmillan. Images courtesy of Eric Wieschaus, Princeton University.)

**FIGURE 21-13** Three thresholds and three types of regulatory DNAs. The *twist* 5' regulatory DNA contains two low-affinity Dorsal-binding sites that are occupied only by peak levels of the Dorsal gradient. As a result, *twist* expression is restricted to ventral nuclei. The *rhombo* 5' enhancer contains a cluster of Dorsal-binding sites. Only one of these sites represents an optimal, high-affinity Dorsal recognition sequence. This mixture of high- and low-affinity sites allows both high and intermediate levels of the Dorsal gradient to activate *rhombo* expression in ventrolateral regions. Finally, the *sog* intronic enhancer contains four evenly spaced optimal Dorsal-binding sites. These allow high, intermediate, and low levels of the Dorsal gradient to activate *sog* expression throughout ventral and lateral regions.

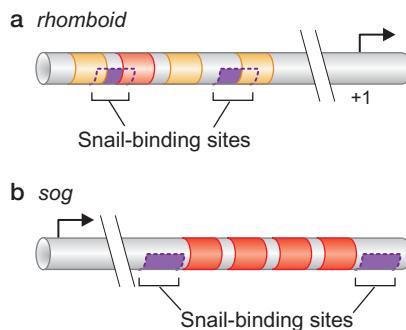
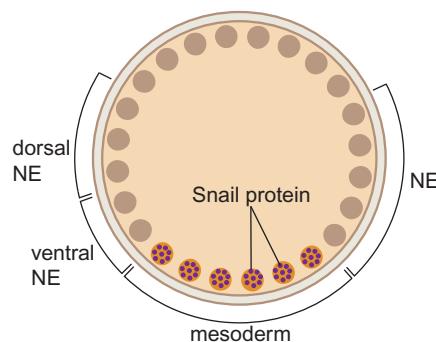


more lateral regions. Toll signaling causes the degradation of a cytoplasmic inhibitor, Cactus, and the release of Dorsal from the cytoplasm into nuclei. This leads to the formation of a corresponding Dorsal nuclear gradient in the ventral half of the early embryo. Nuclei located in the ventral regions of the embryo contain peak levels of the Dorsal protein, whereas those nuclei located in lateral regions contain lower levels of the protein.

The activation of some Dorsal target genes requires peak levels of the Dorsal protein, whereas others can be activated by intermediate and low levels, respectively. In this way, the Dorsal gradient specifies three major thresholds of gene expression across the dorsoventral axis of embryos undergoing cellularization ~2 h after fertilization. These thresholds initiate the differentiation of three distinct tissues: mesoderm, ventral neurogenic ectoderm, and dorsal neurogenic ectoderm (Fig. 21-13). Each of these tissues goes on to form distinctive cell types in the adult fly. The mesoderm forms flight muscles and internal organs, such as the fat body, which is analogous to our liver. The ventral and dorsal neurogenic ectoderm form distinct neurons in the ventral nerve cord.

We now consider the regulation of three different target genes that are activated by high, intermediate, and low levels of the Dorsal protein: *twist*, *rhombo*, and *sog*. The highest levels of the Dorsal gradient—that is, in nuclei with the highest levels of Dorsal protein—activate the expression of the *twist* gene in the ventralmost 18 cells that form the mesoderm (Fig. 21-13). The *twist* gene is not activated in lateral regions, the neurogenic ectoderm, where there are intermediate and low levels of the Dorsal protein. The reason for this is that the *twist* 5' regulatory DNA contains two low-affinity Dorsal-binding sites (Fig. 21-13). Therefore, peak levels of the Dorsal gradient are required for the efficient occupancy of these sites; the lower levels of Dorsal protein present in lateral regions are insufficient to bind and activate the transcription of the *twist* gene.

The *rhombo* gene is activated by intermediate levels of the Dorsal protein in the ventral neurogenic ectoderm. The *rhombo* 5'-flanking region contains a 300-bp enhancer located ~1.5 kb 5' of the transcription start site (Fig. 21-14a). This enhancer contains a cluster of Dorsal-binding sites, mostly low-affinity sites as seen in the *twist* 5' regulatory region. At least one of the sites, however, is an optimal, high-affinity site that permits the binding of intermediate levels of Dorsal protein—the amount present in lateral regions. In principle, the *rhombo* enhancer can be activated by both the high levels of Dorsal protein present in the mesoderm and the intermediate levels present in the ventral neurogenic ectoderm, but it is kept off in the mesoderm by a transcriptional repressor called **Snail**. The Snail repressor is only expressed in the mesoderm; it is not present in the neurogenic



**FIGURE 21-14** Regulatory DNAs. (a) The *rhomboid* enhancer contains binding sites for both Dorsal and the Snail repressor. Because the Snail protein is only present in ventral regions (the mesoderm), *rhomboid* is kept off in the mesoderm and restricted to ventral regions of the neurogenic ectoderm (ventral NE). (b) The intronic *sog* enhancer also contains Snail repressor sites. These keep *sog* expression off in the mesoderm and restricted to broad lateral stripes that encompass both ventral and dorsal regions of the neurogenic ectoderm (NE).

ectoderm. The 300-bp *rhomboid* enhancer contains binding sites for the Snail repressor, in addition to the binding sites for the Dorsal activator. This interplay between the broadly distributed Dorsal gradient and the localized Snail repressor leads to the restricted expression of the *rhomboid* gene in the ventral neurogenic ectoderm. We have already seen how the localized Ash1 repressor blocks the action of the SWI5 activator in the daughter cell of budding yeast, and further along in this chapter, we see the extensive use of this principle in other aspects of *Drosophila* development.

The lowest levels of the Dorsal protein, present in lateral regions of the early embryo, are sufficient to activate the *sog* gene in broad lateral stripes that encompass both the ventral and dorsal neurogenic ectoderm. Expression of *sog* is regulated by a 400-bp enhancer located within the first intron of the gene (Fig. 21-14b). This enhancer contains a series of four evenly spaced high-affinity Dorsal-binding sites that can therefore be occupied even by the lowest levels of the Dorsal protein. As seen for *rhomboid*, the presence of the Snail repressor precludes activation of *sog* expression in the mesoderm despite the high levels of Dorsal protein found there. Thus, the differential regulation of gene expression by different thresholds of the Dorsal gradient depends on the combination of the Snail repressor and the affinities of the Dorsal-binding sites.

The occupancy of Dorsal-binding sites is determined by the intrinsic affinities of the sites, as well as protein–protein interactions between Dorsal and other regulatory proteins bound to the target enhancers. For example, we have seen that the 300-bp *rhomboid* enhancer is activated by intermediate levels of the Dorsal gradient in the ventral neurogenic ectoderm. This enhancer contains mostly low-affinity Dorsal-binding sites. However, intermediate levels of Dorsal are sufficient to bind these sites because of protein–protein interactions with another activator protein called **Twist**. However, intermediate levels of Dorsal are sufficient to bind these sites because of protein–protein interactions with additional activators that bind to the *rhomboid* enhancer. Different mechanisms of cooperative interactions are discussed in Chapter 19 and in Box 21-4, Activator Synergy.

### Segmentation Is Initiated by Localized RNAs at the Anterior and Posterior Poles of the Unfertilized Egg

At the time of fertilization, the *Drosophila* egg contains two localized mRNAs. One, the *bicoid* mRNA, is located at the anterior pole, and the other, the *oskar* mRNA, is located at the posterior pole (Fig. 21-15a). The *oskar* mRNA encodes an RNA-binding protein that is responsible for the assembly of **polar granules**. These are large macromolecular complexes composed of a variety of different proteins and RNAs. The polar granules control the development of tissues that arise from posterior regions of the early embryo,

## ► KEY EXPERIMENTS

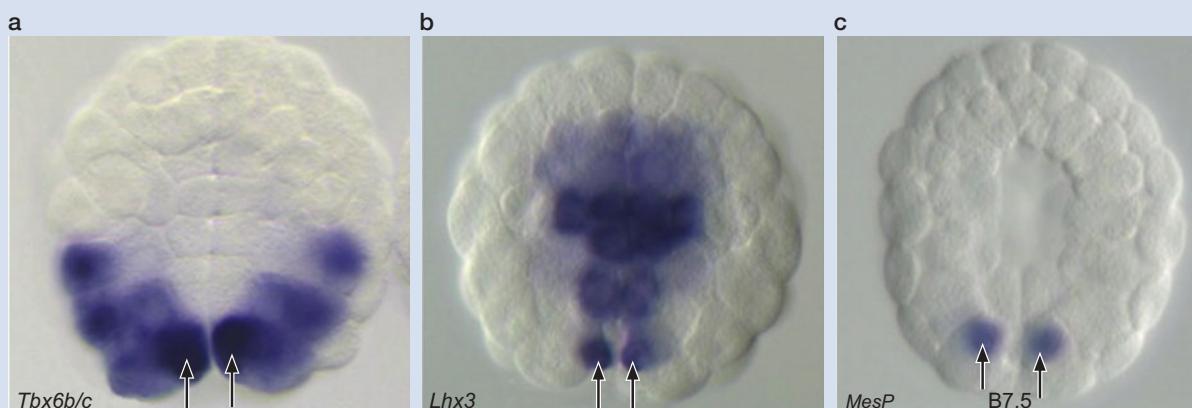
**Box 21-4 Activator Synergy**

Bacterial regulatory proteins such as the *lac* and  $\lambda$  repressors bind as dimers with high affinity. In yeast, the Gal4 activator binds as a dimer with high affinity to induce the expression of Gal1 and other genes required for galactose metabolism (see Chapter 19). In contrast, animal cells tend to lack such “dedicated” transcription factors. Many or most such factors bind to DNA as monomers with low affinities. Consequently, gene regulation is inherently more combinatorial in animal cells than in bacteria or yeast. Multiple proteins binding to multiple sites are required to achieve the activation or repression of gene expression.

This principle of combinatorial gene control is a pervasive feature of animal development. Quite often, activators A and B function in a synergistic manner to delineate a restricted pattern of gene expression. Neither A nor B alone is sufficient to do the job. There are many examples of activator synergy in animal development, but we illustrate the principle by con-

sidering the specification of the cardiac mesoderm (heart precursor cells) in the sea squirt embryo.

A regulatory gene called *MesP* is a critical determinant of cardiac mesoderm in both sea squirts and vertebrates. It is selectively activated in the B7.5 blastomeres of 110-cell embryos (arrows, Box 21-4 Fig. 1). These cells give rise to the beating heart of adult sea squirts. *MesP* is activated by two transcription factors, *Tbx6b/c* and *Lhx3*. *Tbx6b/c* is expressed throughout the developing tail muscles, as well as the B7.5 blastomeres (arrows, Box 21-4 Fig. 1a). *Lhx3* is expressed throughout the presumptive gut, along with the B7.5 blastomeres (arrows, Box 21-4 Fig. 1b). Only the B7.5 blastomeres contain both *Tbx6b/c* and *Lhx3*, and in these cells they work synergistically to activate *MesP* (Box 21-4 Fig. 1c). Because neither transcription factor alone is sufficient for activation, *MesP* expression is restricted to B7.5 and is inactive in the gut and tail muscles.



©2009 Elsevier

**BOX 21-4 FIGURE 1** *MesP* is synergistically activated by two transcription factors. Cells expressing each protein are stained blue. (a) Expression of *Tbx6b/c*. (b) Expression of *Lhx3*. (c) Expression of *MesP*. (Courtesy of Lionel Christiaen. Reproduced, with permission, from Christiaen L. et al. 2009. *Dev. Biol.* **328**: 552. Parts a, b, and c are from Fig. 3A, 3B, p. 556 and Fig. 6C, p. 558. © Elsevier.)

including the abdomen and the pole cells, which are the precursors of the germ cells (Fig. 21-15b).

The *oskar* mRNA is synthesized within the ovary of the mother fly. It is first deposited at the anterior end of the immature egg, or **oocyte**, by “helper” cells called **nurse cells**. Both the oocyte and associated nurse cells arise from specialized stem cells within the ovary (see Box 21-5, Stem Cell Niche). As the oocyte enlarges to form the mature egg, the *oskar* mRNA is transported from anterior to posterior regions. This localization process depends on specific sequences within the 3' UTR of the *oskar* mRNA (Fig. 21-16). We have already seen how the 3' UTR of the *ash1* mRNA mediates its localization to the daughter cell of budding yeast by interacting with the growing ends of microtubules. A remarkably similar process controls the localization of the *oskar* mRNA in the *Drosophila* oocyte.

The *Drosophila* oocyte is highly polarized. The nucleus is located in anterior regions; growing microtubules extend from the nucleus into the poste-

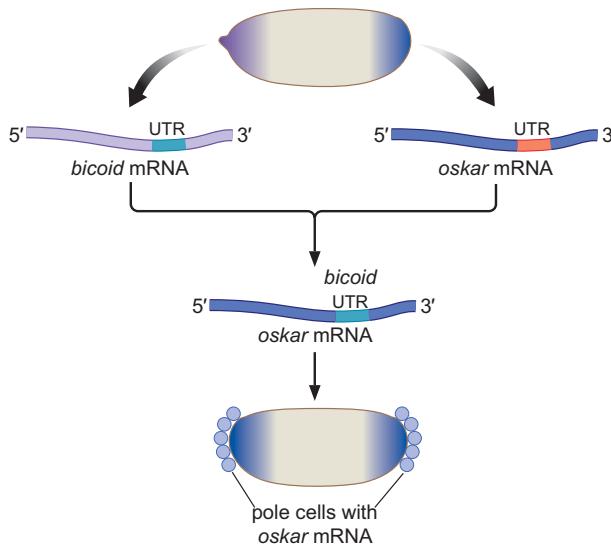
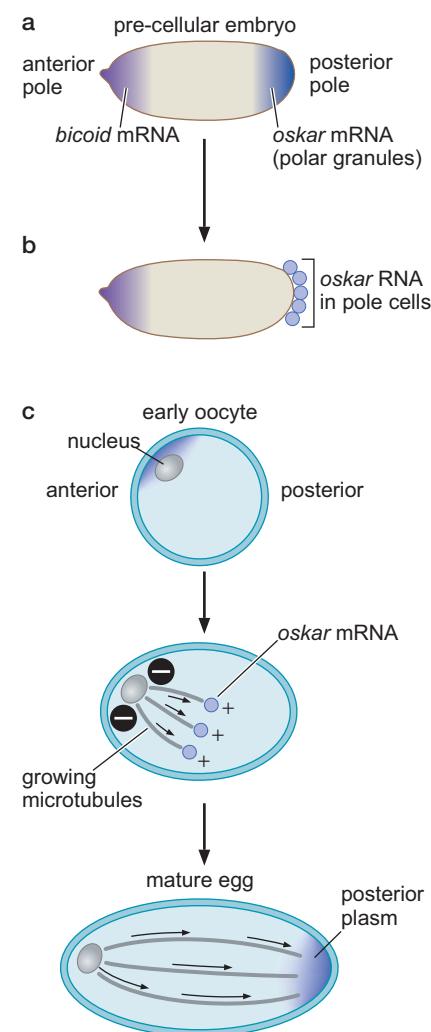
**FIGURE 21-15** Localization of maternal mRNAs in the *Drosophila* egg and embryo. (a) The unfertilized *Drosophila* egg contains two localized mRNAs: *bicoid* in anterior regions and *oskar* in posterior regions. (b) The *Oskar* protein helps coordinate the assembly of the polar granules in the posterior cytoplasm. Nuclei that enter this region bud off the posterior end of the embryo and form the pole cells. (c) During the formation of the *Drosophila* egg, polarized microtubules are formed that extend from the oocyte nucleus and grow toward the posterior plasm. The *oskar* mRNA binds adaptor proteins that interact with the microtubules and thereby transport the RNA to the posterior plasm. The “–” and “+” symbols indicate the direction of the growing strands of the microtubules.

rior cytoplasm. The *oskar* mRNA interacts with adaptor proteins that are associated with the growing “+” ends of the microtubules and are thereby transported away from anterior regions of the egg, where the nucleus resides, into the posterior plasm. After fertilization, the cells that inherit the localized *oskar* mRNA (and polar granules) form the pole cells.

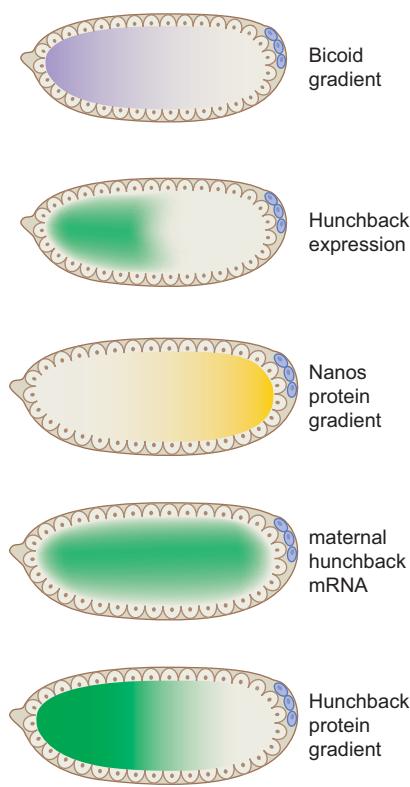
The localization of the *bicoid* mRNA in anterior regions of the unfertilized egg also depends on sequences contained within its 3' UTR. The nucleotide sequences of the *oskar* and *bicoid* mRNAs are distinct. As a result, they interact with different adaptor proteins and become localized to different regions of the egg. The importance of the 3' UTRs in determining where each mRNA becomes localized is revealed by the following experiment. If the 3' UTR from the *oskar* mRNA is replaced with that from *bicoid*, the hybrid *oskar* mRNA is located to anterior regions (just as *bicoid* normally is). This mislocalization is sufficient to induce the formation of pole cells at abnormal locations in the early embryo (see Fig. 21-16). In addition, the mislocalized polar granules suppress the expression of genes required for the differentiation of head tissues. As a result, embryonic cells that normally form head tissues are transformed into germ cells.

### Bicoid and Nanos Regulate *hunchback*

The Bicoid regulatory protein is synthesized before the completion of cellularization. As a result, it diffuses away from its source of synthesis at the anterior pole and becomes distributed in a broad concentration gradient along the length of the early embryo. Both high and intermediate concentrations of Bicoid are sufficient to activate *hunchback*, which is essential for the subdivision of the embryo into a series of segments (Fig. 21-17). The



**FIGURE 21-16** The *bicoid* and *oskar* mRNAs contain different UTR sequences. The *bicoid* UTR causes it to be localized to the anterior pole, and the distinct *oskar* UTR sequence causes localization in the posterior plasm. An engineered *oskar* mRNA that contains the *bicoid* UTR is localized to the anterior pole, just like the normal *bicoid* mRNA. This mislocalization of *oskar* causes the formation of pole cells in anterior regions. Pole cells also form from the posterior pole because of localization of the normal *oskar* mRNA in the posterior plasm.



**FIGURE 21-17** Hunchback protein gradient and translation inhibition by Nanos. The broad anteroposterior Bicoid protein gradient produces a sharp threshold of *hunchback* gene expression, as *hunchback* is activated by both high and intermediate levels of the Bicoid gradient. The Nanos mRNA is associated with polar granules; after its translation, the protein diffuses from posterior regions to form a gradient. The maternal *hunchback* mRNA is distributed throughout the early embryo, but its translation is arrested by the Nanos protein, which binds to specific sequences in the *hunchback* 3' UTR. The Nanos gradient thereby leads to the formation of a reciprocal Hunchback gradient in anterior regions.

*hunchback* gene is actually transcribed from two promoters: one is activated by the Bicoid gradient, and the other controls expression in the developing oocyte. The latter, “maternal” promoter leads to the synthesis of a *hunchback* mRNA that is evenly distributed throughout the cytoplasm of unfertilized eggs. The translation of this maternal transcript is blocked in posterior regions by an RNA-binding protein called **Nanos** (Fig. 21-17). Nanos is found only in posterior regions because its mRNA is, in turn, selectively localized there through interactions between its 3' UTR and the polar granules we encountered earlier.

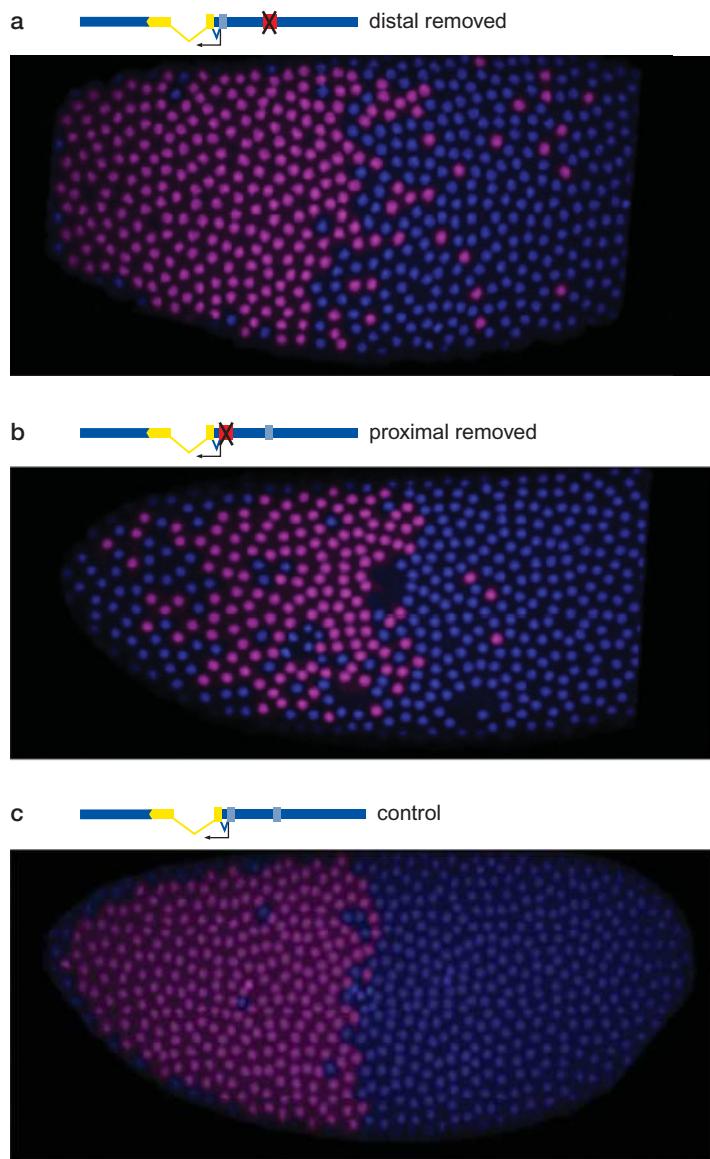
Nanos protein binds specific RNA sequences, NREs (Nanos response elements), located in the 3' UTR of the maternal *hunchback* mRNAs, and this binding causes a reduction in the *hunchback* poly-A tail, which, in turn, destabilizes the RNA and inhibits its translation (see Chapter 15). Thus, we see that the Bicoid gradient activates the zygotic *hunchback* promoter in the anterior half of the embryo, whereas Nanos inhibits the translation of the maternal *hunchback* mRNA in posterior regions (see Fig. 21-17). This dual regulation of *hunchback* expression produces a steep Hunchback protein gradient, with the highest concentrations located in the anterior half of the embryo and sharply diminishing levels in the posterior half. Further considerations of gradients and their implications in development are discussed in Box 21-6, Gradient Thresholds.

### Multiple Enhancers Ensure Precision of *hunchback* Regulation

Many patterning genes are regulated by “redundant” or multiple enhancers. As an example, consider the early activation of the *hunchback* gene by the Bicoid gradient. High and intermediate levels of the gradient activate *hunchback* expression in the anterior half of the embryo (see Fig. 21-17). Activation is mediated by two separate enhancers, which possess similar arrangements of Bicoid-binding sites and similar regulatory activities (Fig. 21-18). Why two enhancers rather than one? Two enhancers produce a sharper, more precise pattern of gene activation than either enhancer alone. Moreover, two enhancers help ensure reliable activation of the gene in large populations of embryos subjected to environmental variations such as changes in temperature. In some cases, multiple enhancers possessing overlapping activities prevent such variations to alter normal development. For example, a regulatory gene called *shavenbaby* is important for the development of tiny sensory hairs along the dorsal surface of advanced-stage embryos. *shavenbaby* is regulated by five separate enhancers distributed over an interval of 40 kb upstream of the transcription start site. Deletions of individual enhancers do not cause significant defects in the morphology of the hairs at optimal temperatures. But under adverse conditions, such as high (30°C) or low (15°C) temperatures, the removal of an enhancer results in fewer or misshapen sensory hairs.

### The Gradient of Hunchback Repressor Establishes Different Limits of Gap Gene Expression

Hunchback functions as a transcriptional repressor to establish different limits of expression of the so-called gap genes: *Krüppel*, *knirps*, and *giant* (discussed in Box 21-3). We will see that Hunchback also works in concert with the proteins encoded by these gap genes to produce segmentation stripes of gene expression, the first step in subdividing the embryo into a repeating series of body segments.



**FIGURE 21-18** Hunchback is regulated by two enhancers with similar activities. (a) Early activation of Hunchback transcription occurs from a transgene containing only the proximal enhancer intact. The distal shadow enhancer was inactivated by mutation (indicated by X in diagram). Note that expression is not restricted to the anterior half (left half) of the embryo. (b) This panel shows activation obtained when the proximal enhancer is inactivated, leaving only the distal shadow enhancer intact. Expression is sporadic in the anterior regions. (c) Uniform activation and a sharp border are observed when both enhancers are intact. (Courtesy of Mike Levine; described in Perry M.W. et al. 2011. *Proc. Natl. Acad. Sci.* **108**: 13570–13575, Fig. 2A–C, p. 13572.)

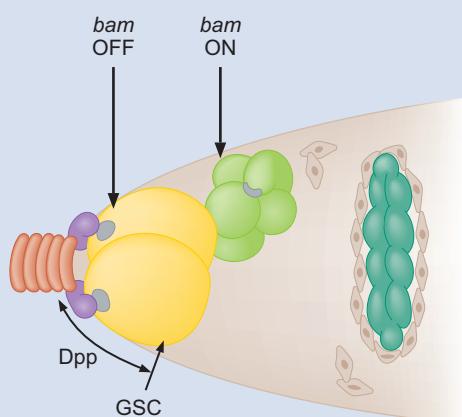
#### MEDICAL CONNECTIONS

##### Box 21-5 Stem Cell Niche

In *Drosophila*, the egg or oocyte arises from a stem cell precursor called the **germline stem cell** (GSC). Quite a lot is known regarding the transition of GSCs into oocytes within the *Drosophila* ovary, and it is likely that many aspects of this mechanism will apply to the development of other classes of stem cells in both flies and humans. Stem cells proliferate only when in direct physical contact with specialized cells, collectively known as the “niche,” which produce a signal that triggers proliferation. When stem cells become detached from the niche, proliferation stops and the cells undergo differentiation into specialized cell types. In the *Drosophila* example, detachment of GSCs from the ovary niche causes them to develop into nondividing oocytes, in a process mediated by signal-induced repression. This process is now well understood at the molecular level and works as follows.

Niche cells within the *Drosophila* ovary, called Cap cells, secrete a diffusible signaling molecule called Dpp. Activation of the Dpp receptor within the associated GSCs results in silencing of a critical regulatory gene called *bam*: when transcription of *bam* is blocked, GSCs proliferate. This silencing of *bam* expression depends on direct physical contact between Cap cells and GSCs, similar to the process that results in activation of Notch signaling during formation of the insect nervous system. As GSCs proliferate, some of the daughter cells become detached from the Cap cells and thus are no longer targets of Dpp signaling. In the absence of signaling, *bam* transcription is activated, and the cell stops proliferating; instead, it differentiates into an oocyte (Box 21-5 Fig. 1).

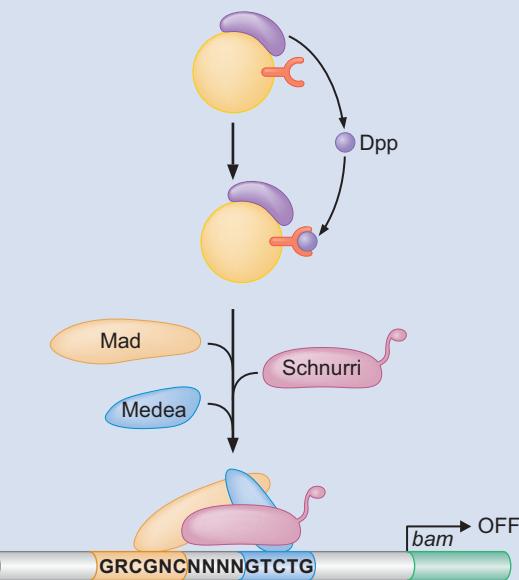
The basic choice between stem cell proliferation and oocyte differentiation therefore depends on the on/off regulation of

**Box 21-5 (Continued)**

**BOX 21-5 FIGURE 1** *bam* expression in developing oocytes. The scheme represents the patterns of expression and distribution of *bam* mRNA and protein. Cap cells (purple) secrete Dpp, which activates its receptor on germline stem cells (GSCs, yellow), resulting in a signaling process that ultimately represses *bam* expression. As GSCs detach from the Cap cells, Dpp signaling is lost, and *bam* mRNA is expressed, leading to production of high levels of its protein in the cytoplasm. In the presence of Bam protein, the detached daughter cells develop into oocyte progenitor cells (green) and further into eight-cell cysts (dark green). (Adapted, with permission, from Chen D. and McKearin D.M. 2003. *Development* 130: 1159–1170, Fig. 1. © Company of Biologists.)

*bam* expression. And this regulation is now known to be mediated by a silencer element in the 5' regulatory region of *bam*, having the sequence GRCGN(C<sub>N</sub>)<sub>5</sub>GTCTG (Box 21-5 Fig. 2). Dpp signaling triggers nuclear transport of two Smad regulatory proteins, called Mad and Medea. These proteins bind the two half-sites in the silencer element and, in turn, recruit a transcriptional repressor, called ZF6-6 or Schnurri, that prevents transcription of *bam*. This recruitment of Schnurri and consequent repression of *bam* occurs only in GSCs that remain in contact with the Cap cells. As a result, these cells divide to produce more stem cells. In contrast, in GSC daughter cells that detach from the Cap cells, *bam* is actively transcribed because the

signaling pathway leading to gene silencing is disrupted. In these cells, the Dpp receptor is not activated (because signaling is disrupted), and Mad and Medea are not transported to the nucleus and thus do not bind the 5' silencer element or recruit the Schnurri repressor. Under these conditions, *bam* is expressed, and the daughter cells no longer proliferate but rather differentiate into oocytes. This requirement for direct physical contact between the niche and stem cell and resulting signal-induced repression may be a general mechanism for continuing stem cell proliferation.



**BOX 21-5 FIGURE 2** The Dpp pathway actively represses key developmental genes. The binding of Dpp, secreted by Cap cells (purple), to the Dpp receptor on germline stem cells (GSC, yellow) initiates a signal that prompts the transport of Mad (orange) and Medea (blue) into the nucleus. Repressed target genes (here, *bam* is shown as an example) contain a *cis*-acting silencing element that binds Mad and Medea, which together recruit Schnurri (pink) to effectively block transcription. (Adapted, with permission, from Pyrowolakis G. et al. 2004. *Dev. Cell* 7: 229–240, Fig. 7. © Elsevier.)

The Hunchback protein is distributed in a steep gradient that extends through the presumptive thorax and into the abdomen. High levels of the Hunchback protein repress the transcription of *Krüppel*, whereas intermediate and low levels of the protein repress the expression of *knirps* and *giant*, respectively (Fig. 21-19a). We have seen that the binding affinities of the Dorsal activator are responsible for producing different thresholds of gene expression. The Hunchback repressor gradient might not work in the same way. Instead, the number of Hunchback repressor sites may be a more critical determinant for distinct patterns of *Krüppel*, *knirps*, and *giant* expression (Fig. 21-19b). The *Krüppel* enhancer contains only three Hunchback-binding sites and is repressed by high levels of the Hunchback gradient. In contrast, the *giant* enhancer contains seven Hunchback sites and is repressed

## ► ADVANCED CONCEPTS

**Box 21-6 Gradient Thresholds**

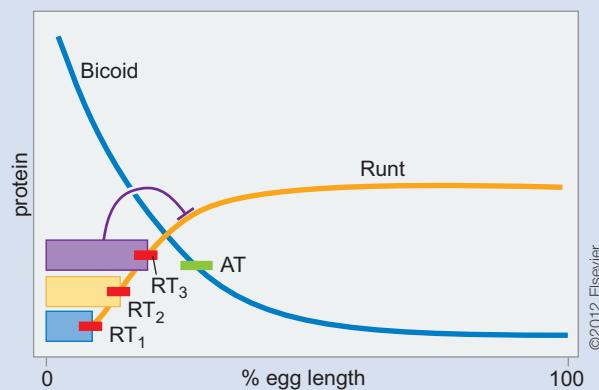
We have encountered several examples of regulatory gradients producing different patterns of gene expression. Sonic hedgehog and its transcriptional effector Gli establish differential patterns of Nkx2.2, Olig2, and Pax6 expression in the developing neural tube of vertebrate embryos (Fig. 21-11). The Dorsal gradient generates different patterns of gene expression in the ventral mesoderm, lateral (neurogenic) ectoderm, and dorsal ectoderm of precellular *Drosophila* embryos (Fig. 21-12). The famous Bicoid gradient establishes sequential patterns of "gap" gene expression across the anterior–posterior axis of the precellular embryo (Box 21-3 Fig. 3).

Until recently, it was generally assumed that the affinity of Gli-, Dorsal-, and Bicoid-binding sites determined the spatial limits of gene expression. Indeed, we have discussed the evidence that such a mechanism is used for the Dorsal gradient. But there is emerging evidence that different binding affinities might not be sufficient to account for the diverse patterns of gene expression produced by the Gli and Bicoid gradients. For example, it was recently shown that Bicoid target genes activated by high levels of the Bicoid gradient contain similar binding affinities as those regulated by low levels of the gradient. In contrast, a simple binding affinity model would predict that genes activated by high levels of the gradient contain low-affinity sites, whereas genes activated by low levels contain optimal, high-affinity sites.

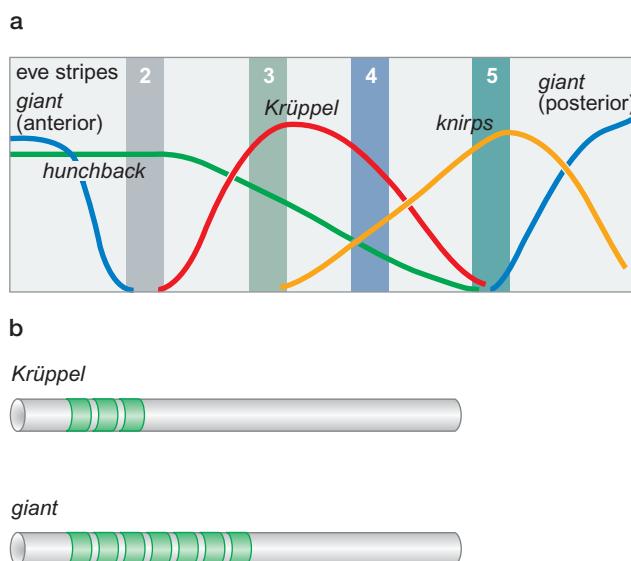
In fact, it appears that different threshold readouts of the Bicoid gradient depend on opposing repressor gradients, including the Runt repressor (Box 21-6 Fig. 1). Target genes RT<sub>1</sub>, RT<sub>2</sub>, and RT<sub>3</sub> are activated by progressively lower levels of the Bicoid gradient. But RT<sub>1</sub> and RT<sub>2</sub> contain Bicoid-binding sites with similar affinities. Their distinctive limits of expression

appear to depend on differential repression by Runt. RT<sub>3</sub> is repressed by high levels of the Runt gradient, whereas RT<sub>2</sub> and RT<sub>1</sub> are repressed by progressively lower levels. It is currently uncertain whether these target genes contain similar Runt-binding sites. Perhaps their differential responses depend on different numbers of sites, with RT<sub>1</sub> containing more Runt repressor sites than RT<sub>2</sub> or RT<sub>3</sub>.

As discussed above, the Gli activator gradient in the vertebrate neural tube might also rely on the use of transcriptional repressors to produce different readouts of the Sonic hedgehog gradient (see Fig. 21-11).

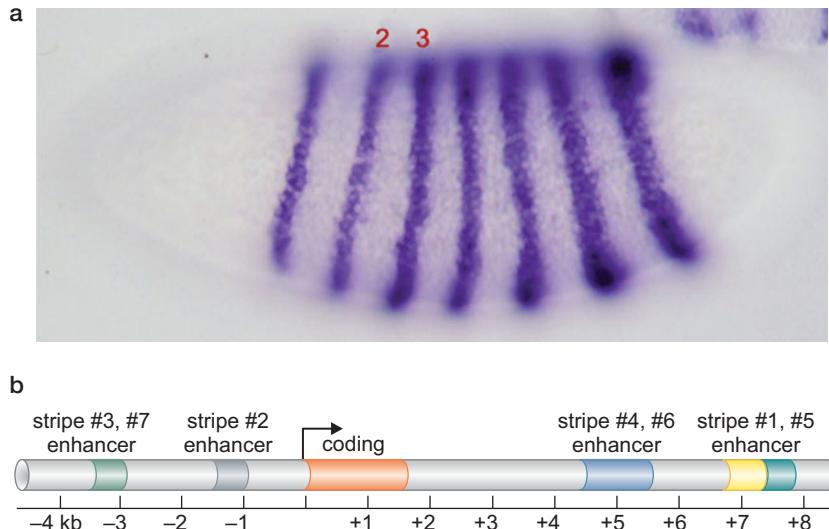


**BOX 21-6 FIGURE 1** Cooperation of activator and repressor gradients. See text for details. (Adapted, with permission, from Roth S. and Lynch J. 2012. *Cell* 149: 511, Fig. 1, p. 512. © Elsevier.)



**FIGURE 21-19** Expression of *hunchback* forms sequential gap expression patterns. (a) The anteroposterior Hunchback repressor gradient establishes different limits of Krüppel, knirps, and giant expression. High levels of Hunchback are required for the repression of Krüppel, but low levels are sufficient to repress giant. (b) The Krüppel and giant 5' regulatory DNAs contain different numbers of Hunchback repressor sites. There are three sites in Krüppel, but seven sites in giant. The increased number of Hunchback sites in the giant enhancer may be responsible for its repression by low levels of the Hunchback gradient. (a, Redrawn, with permission, from Gilbert S.E. 1997. *Developmental biology*, 5th ed., p. 565, Fig. 14-23. © Sinauer.)

**FIGURE 21-20** Expression of the *eve* gene in the developing embryo. (a) *eve* expression pattern in the early embryo. (b) The *eve* locus contains more than 12 kb of regulatory DNA. The 5' regulatory region contains two enhancers, which control the expression of stripes 2, 3, and 7. Each enhancer is 500 bp in length. The 3' regulatory region contains three enhancers that control the expression of stripes 4 and 6, stripe 1, and stripe 5, respectively. The five enhancers produce seven stripes of *eve* expression in the early embryo. (a, Image courtesy of Michael Levine.)



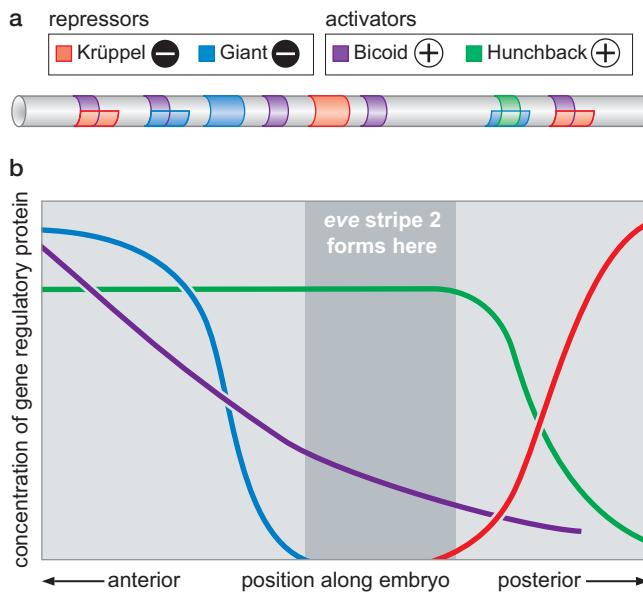
by low levels of the Hunchback gradient. The underlying mechanism here is unknown. Perhaps different thresholds of repression are produced by the additive effects of the individual Hunchback repression domains.

### Hunchback and Gap Proteins Produce Segmentation Stripes of Gene Expression

A culminating event in the regulatory cascade that begins with the localized *bicoid* and *oskar* mRNAs is the expression of a “pair-rule” gene called *even-skipped*, or simply *eve*. The *eve* gene is expressed in a series of seven alternating, or pair-rule, stripes that extend along the length of the embryo (Fig. 21-20). Each *eve* stripe encompasses four cells, and neighboring stripes are separated by interstripe regions—also four cells wide—that express little or no *eve*. These stripes foreshadow the subdivision of the embryo into a repeating series of body segments.

The *eve* protein-coding sequence is rather small, <2 kb in length. In contrast, the flanking regulatory DNAs that control *eve* expression encompass more than 12 kb of genomic DNA: ~4 kb located 5' of the *eve* transcription start site, and ~8 kb in the 3'-flanking region (see Fig. 21-20). The 5' regulatory region is responsible for initiating stripes 2, 3, and 7, and the 3' region regulates stripes 1, 4, 5, and 6. The 12 kb of regulatory DNA contains five separate enhancers that together produce the seven different stripes of *eve* expression seen in the early embryo. Each enhancer initiates the expression of just one or two stripes. (In Box 21-7, *cis*-Regulatory Sequences in Animal Development and Evolution, we discuss further aspects and examples of the modular organization of regulatory elements within animal genomes.) We now consider the regulation of the enhancer that controls the expression of *eve* stripe 2.

The stripe 2 enhancer is 500 bp in length and located 1 kb upstream of the *eve* transcription start site. It contains binding sites for four different regulatory proteins: Bicoid, Hunchback, Giant, and Krüppel (Fig. 21-21). We have seen how Hunchback functions as a repressor when controlling the expression of the gap genes; in the context of the *eve* stripe 2 enhancer, it works as an activator. In principle, Bicoid and Hunchback can activate the stripe 2 enhancer in the entire anterior half of the embryo because both proteins are present there, but Giant and Krüppel function as repressors that establish the edges of the stripe 2 pattern—the anterior and posterior borders, respectively (see Fig. 21-21).



**FIGURE 21-21** Regulation of *eve* stripe

2. (a) The 500-bp enhancer contains a total of 12 binding sites for the Bicoid, Hunchback, Krüppel, and Giant proteins. The distribution of these regulatory proteins in the early *Drosophila* embryo is summarized in the diagram shown in b. There are high levels of the Bicoid and Hunchback proteins in the cells that express *eve* stripe 2. The borders of the stripes are formed by the Giant and Krüppel repressors. (Giant is expressed in anterior and posterior regions. Only the anterior pattern is shown; the posterior pattern, which is regulated by Hunchback, is not shown.) (Adapted, with permission, from Alberts B. et al. 2002. *Molecular biology of the cell*, 4th ed.: a, p. 409, Fig. 7-55; b, p. 410, Fig. 7-56. © Garland Science/Taylor & Francis LLC.)

## ► KEY EXPERIMENTS

### Box 21-7 *cis*-Regulatory Sequences in Animal Development and Evolution

*cis*-regulatory sequences are organized in a modular fashion within animal genomes. In general, there are separate enhancers for the individual components of a complex expression pattern. Consider a gene that is expressed in multiple tissues and organs within a developing mouse embryo, such as the liver, pancreas, and pituitary gland. Odds are that the gene contains separate enhancers for each of these sites of expression. We have seen that the *eve* locus contains five separate enhancers located in the 5'- and 3'-flanking regions (see Fig. 21-20). Each enhancer directs the expression of just one or two of the seven *eve* stripes in the early *Drosophila* embryo. This type of modular organization facilitates morphological diversity via evolution of *cis*-regulatory sequences, as we discuss later.

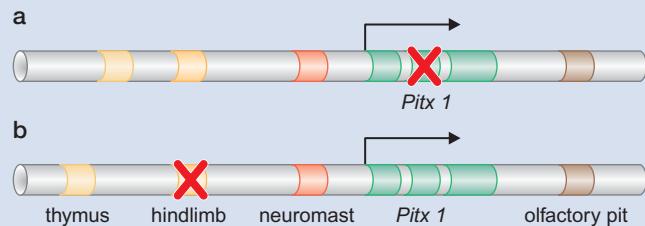
#### Modular Organization Circumvents Pleiotropy

How do patterns of gene expression change during evolution? There is emerging evidence that nucleotide changes within critical activator binding sites eliminate gene expression within a specific tissue or cell type during evolution. Consider the example of pelvic fins in stickleback fish. There are natural variants of sticklebacks that lack pelvic fins. When mated with individuals containing fins, it was possible to identify a major genetic locus responsible for reduced fins. It maps within the 5'-flanking region of the *Pitx1* gene. *Pitx1* is a developmental control gene that is essential for the development of several different tissues in mice, including the thymus, olfactory pit, and hindlimbs. In sticklebacks, it would appear that reduced fins result from point mutations in critical activator sites within the pelvic fin ("hindlimb") enhancer (Box 21-7 Fig. 1). These mutations disrupt expression in the developing pelvic fins, but they do not interfere with the activities of the other enhancers required for regulating *Pitx1* in the thymus, olfactory pit, and other tissues where the *Pitx1* gene is active.

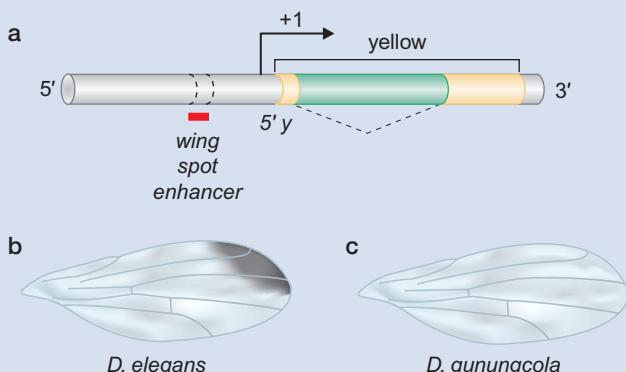
Specific alterations within a modular, *cis*-regulatory region are also responsible for the evolution of distinct pigmentation patterns in different species of *Drosophila*. The classical *yellow* (*y*) locus is critical for pigmentation, and simple mutations in the gene result in flies with a yellow body color that lack localized foci of melanin. The *y* gene is regulated by separate enhancers for expression in the bristles, wings, and abdomen, as we now describe.

*D. melanogaster* adults (particularly males) contain intense pigmentation in the posterior abdominal segments. This pigmentation is due to the direct activation of the *y* abdominal enhancer by the Hox protein Abd-B. *Drosophilids* lacking abdominal segmentation, such as *Drosophila kikkawai*, contain point mutations in a critical Abd-B activator site. This causes a loss of *y* expression in the abdomen and the observed loss of pigmentation.

A separate enhancer controls *y* expression in the wings. In some *Drosophila* species, this enhancer directs a spot of pigmentation in a specific quadrant of the adult male wing (Box 21-7



**BOX 21-7 FIGURE 1** The developmental control gene *Pitx1*. (a) The structure of the *Pitx1* gene with 5' upstream sequences. Shown here is a lethal null mutation (of a laboratory mouse) within the coding region (second exon) of the gene. (b) In the wild stickleback, a viable regulatory mutation within the 5' upstream sequence results in reduced pelvic fin size.

**Box 21-7 (Continued)**

**BOX 21-7 FIGURE 2** The yellow (*y*) locus of *Drosophila*. (a) The panel shows the structure and upstream regulatory sequences (enhancer sequences) of the *yellow* gene. (b) The normal pigmentation (the “mating spot”) of the adult male wing in one species of *Drosophila*. (c) The wing of a species lacking pigmentation; this species carries a mutation in the 5' spot enhancer.

Fig. 2). This spot is a critical component of the courtship ritual. Species lacking the mating spot contain point mutations in the wing enhancer, causing the restricted loss of *y* gene activity without compromising its function in other tissues such as the bristles and abdominal cuticle.

#### Changes in Repressor Sites Can Produce Big Changes in Gene Expression

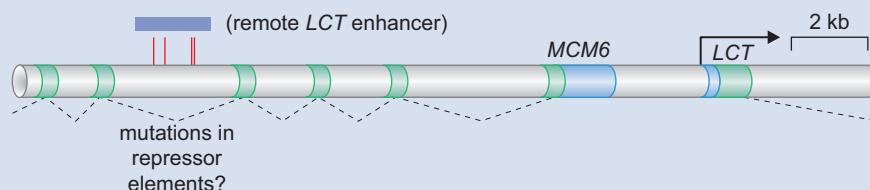
The simple loss of critical activator sites within discrete enhancer modules can explain the localized loss of *Pitx1* and *y* gene ac-

tivities. New patterns of gene expression might arise through the loss of repressor elements.

Most or all of the enhancers active in the early *Drosophila* embryo have repressor binding sites that are responsible for creating sharp boundaries of gene expression. For example, the *eve* stripe 2 enhancer contains binding sites for the Giant and Krüppel repressors, which produce sharp anterior and posterior borders of gene expression (see Fig. 21-21). Mutations in these sites cause a dramatic expansion in the normal expression pattern: a broad band of expression rather than a tight stripe.

A possible example of evolution via repressor elements is seen for the lactase (*LCT*) gene in human populations. In most primates, the *LCT* gene is expressed at high levels in the small intestines of infants, during the time they obtain milk from their mothers. However, the *LCT* gene is shut off after adolescence. Certain populations of humans are unusual in retaining *LCT* gene expression as adults. This persistence correlates with pastoral societies that use dietary milk long after weaning. Individual populations with persistent *LCT* expression contain nucleotide substitutions in an intronic sequence within the *MCM6* gene, located immediately 5' of *LCT* (Box 21-7 Fig. 3).

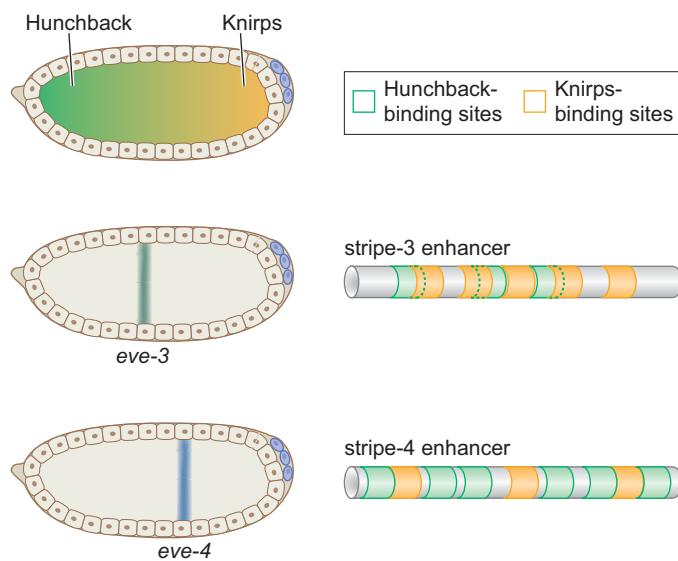
These nucleotide changes might damage repressor elements that normally bind a silencer protein responsible for repressing *LCT* expression in the small intestines of adolescents and adults. Such a loss of critical *cis*-regulatory elements would be comparable to the inactivation of the hindlimb/pelvic fin enhancer in the *Pitx1* gene in sticklebacks or the inactivation of the abdominal and wing enhancers in the *y* gene of *Drosophila*. But in the case of the lactase gene, a novel pattern of gene expression is evolved, temporal persistence of *LCT* activity, because of the loss of repression elements.



**BOX 21-7 FIGURE 3** Structure of the *LCT* gene and its 5' upstream regulatory region.

#### Gap Repressor Gradients Produce Many Stripes of Gene Expression

*eve* stripe 2 is formed by the interplay of broadly distributed activators (Bicoid and Hunchback) and localized repressors (Giant and Krüppel). The same basic mechanism applies to the regulation of the other *eve* enhancers as well. For example, the enhancer that directs the expression of *eve* stripe 3 can be activated throughout the early embryo by ubiquitous transcriptional activators. The stripe borders are defined by localized gap repressors: Hunchback establishes the anterior border, whereas Krüppel specifies the posterior border (Fig. 21-22).



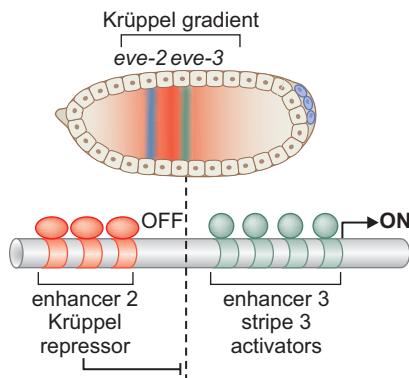
**FIGURE 21-22** Differential regulation of the stripe 3 and stripe 4 enhancers by opposing gradients of the Hunchback and Knirps repressors. The two stripes are positioned in different regions of the embryo. The *eve* stripe 3 enhancer is repressed by high levels of the Hunchback gradient but low levels of the Knirps gradient. Conversely, the stripe 4 enhancer is repressed by low levels of the Hunchback gradient but high levels of Knirps. The stripe 3 enhancer contains just a few Hunchback-binding sites, and as a result, high levels of the Hunchback gradient are required for its repression. The stripe 3 enhancer contains many Knirps-binding sites, and consequently, low levels of Knirps are sufficient for repression. The stripe 4 enhancer has the opposite organization of repressor-binding sites. There are many Hunchback sites, and these allow low levels of the Hunchback gradient to repress stripe 4 expression. The stripe 4 enhancer contains just a few Knirps sites, so that high levels of the Knirps gradient are required for repression. Note that the stripe 3 enhancer actually directs the expression of two stripes, 3 and 7. The stripe 4 enhancer directs the expression of stripes 4 and 6. For simplicity, we consider only one of the stripes from each enhancer.

The enhancer that controls the expression of *eve* stripe 4 is also repressed by Hunchback and Knirps. However, different concentrations of these repressors are required in each case. Low levels of the Hunchback gradient that are insufficient to repress the *eve* stripe 3 enhancer are sufficient to repress the *eve* stripe 4 enhancer (Fig. 21-22). This differential regulation of the two enhancers by the Hunchback repressor gradient produces distinct anterior borders for the stripe 3 and stripe 4 expression patterns. The Knirps protein is also distributed in a gradient in the precellular embryo. Higher levels of this gradient are required to repress the stripe 4 enhancer than are needed to repress the stripe 3 enhancer. This distinction produces discrete posterior borders of the stripe 3 and stripe 4 expression patterns.

We have seen that the Hunchback repressor gradient produces different patterns of Krüppel, Knirps, and Giant expression. This differential regulation might be due to the increasing number of Hunchback-binding sites in the Krüppel, Knirps, and Giant enhancers. A similar principle applies to the differential regulation of the stripe 3 and stripe 4 enhancers by the Hunchback and Knirps gradients. The *eve* stripe 3 enhancer contains relatively few Hunchback binding sites but many Knirps sites, whereas the *eve* stripe 4 enhancer contains many Hunchback sites but relatively few Knirps sites (see Fig. 21-22). Similar principles are likely to govern the regulation of the remaining stripe enhancers that control the *eve* expression pattern (as well as the expression of other pair-rule genes).

### Short-Range Transcriptional Repressors Permit Different Enhancers to Work Independently of One Another within the Complex *eve* Regulatory Region

We have seen that *eve* expression is regulated in the early embryo by five separate enhancers. In fact, there are additional enhancers that control *eve* expression in the heart and central nervous system (CNS) of older embryos. This type of complex regulation is not a peculiarity of *eve*. There are genetic loci that contain even more enhancers distributed over even larger distances. For example, several genes are known to be regulated by as many as 10 different enhancers, perhaps more, that are scattered over distances approaching 100 kb (as we discuss later). Thus, genes engaged in important developmental processes are often regulated by multiple enhancers. How do



**FIGURE 21-23** Short-range repression and enhancer autonomy. Different enhancers work independently of one another in the *eve* regulatory region because of short-range transcriptional repression. Repressors bound to one enhancer do not interfere with activators in the neighboring enhancers. For example, the Krüppel repressor binds to the stripe 2 enhancer and keeps stripe 2 expression off in central regions of the embryo. The *eve* stripe 3 enhancer is expressed in these regions. It is not repressed by Krüppel because it lacks the specific DNA sequences that are recognized by the Krüppel protein. In addition, Krüppel repressors bound to the stripe 2 enhancer do not interfere with the stripe 3 activators because they map too far away. Krüppel must bind no more than 100 bp from upstream activators to block their ability to stimulate transcription. The stripe 2 and stripe 3 enhancers are separated by a 1.5-kb spacer sequence.



**FIGURE 21-24** A dominant mutation in the *Antp* gene results in the homeotic transformation of antennae into legs. The fly on the right is normal. Note the rudimentary set of antennae at the front end of the head. The fly on the left is heterozygous for a dominant *Antp* mutation (*AntpD/+*). It is fully viable and mainly normal in appearance except for the remarkable set of legs emanating from the head in place of antennae. (Courtesy of Matthew Scott.)

these enhancers work independently of one another to produce additive patterns of gene expression? In the case of *eve*, five separate enhancers produce seven different stripes.

Short-range transcriptional repression is one mechanism for ensuring enhancer autonomy—the independent action of multiple enhancers to generate additive patterns of gene expression. This means that repressors bound to one enhancer do not interfere with the activators bound to another enhancer within the regulatory region of the same gene. For example, we have seen that the Krüppel repressor binds to the *eve* stripe 2 enhancer and establishes the posterior border of the stripe 2 pattern. The Krüppel repressor works only within the limits of the 500-bp stripe 2 enhancer. It does not repress the core promoter or the activators contained within the stripe 3 enhancer, both of which map more than 1 kb away from the Krüppel repressor sites within the stripe 2 enhancer (Fig. 21-23). If Krüppel could function over long distances, or if it mapped near the promoter (like bacterial repressors), then it would interfere with the expression of *eve* stripe 3, because high levels of the Krüppel repressor are present in that region of the embryo where the *eve* stripe 3 enhancer is active.

## HOMEOTIC GENES: AN IMPORTANT CLASS OF DEVELOPMENTAL REGULATORS

The genetic analysis of *Drosophila* development led to the discovery of an important class of regulatory genes, the homeotic genes, which cause the morphological diversification of the different body segments. Some homeotic genes control the development of mouth parts and antennae from head segments, whereas others control the formation of wings and halteres from thoracic segments. The two best-studied homeotic genes are *Antp* and *Ubx*, responsible for suppressing the development of antennae and wings, respectively.

*Antp* (*Antennapedia*) controls the development of the middle segment of the thorax, the mesothorax. The mesothorax produces a pair of legs that are morphologically distinct from the forelegs and hindlegs. *Antp* encodes a homeodomain regulatory protein that is normally expressed in the mesothorax of the developing embryo. The gene is not expressed, for example, in the developing head tissues. But a dominant *Antp* mutation, caused by a chromosome inversion, brings the *Antp* protein-coding sequence under the control of a “foreign” regulatory DNA that mediates gene expression in head tissues, including the antennae (see Fig. 21-24). When misexpressed in the head, *Antp* causes a striking change in morphology: legs develop instead of antennae.

*Ubx* (*Ultrabithorax*) encodes a homeodomain regulatory protein that controls the development of the third thoracic segment, the metathorax. *Ubx* specifically represses the expression of genes that are required for the development of the second thoracic segment, or mesothorax. Indeed, *Antp* is one of the genes that it regulates: *Ubx* represses *Antp* expression in the metathorax and restricts its expression to the mesothorax of developing embryos. Mutants that lack the *Ubx* repressor show an abnormal pattern of *Antp* expression. The gene is not only expressed within its normal site of action in the developing mesothorax but also misexpressed in the developing metathorax. This misexpression of *Antp* causes a transformation of the metathorax into a duplicated mesothorax.

In adult flies, the mesothorax contains a pair of legs and wings, whereas the metathorax contains a pair of legs and halteres (see Fig. 21-25). The halteres are

considerably smaller than the wings and function as balancing structures during flight. *Ubx* mutants show a spectacular phenotype: they have four fully developed wings, because of the transformation of the halteres into wings.

The expression of *Ubx* in the different tissues of the metathorax depends on the regulatory sequences that encompass more than 80 kb of genomic DNA. A mutation called *Cbx* (*Contrabithorax*) disrupts this *Ubx* regulatory DNA without changing the *Ubx* protein-coding region. The *Cbx* mutation causes *Ubx* to be misexpressed in the mesothorax, in addition to its normal site of expression in the metathorax (Fig. 21-26). *Ubx* now represses the expression of *Antp*, as well as the other genes needed for the normal development of the mesothorax. As a result, the mesothorax is transformed into a duplicated copy of the normal metathorax. This is a striking phenotype: the wings are transformed into halteres, and the resulting *Cbx* mutant flies look like wingless ants.

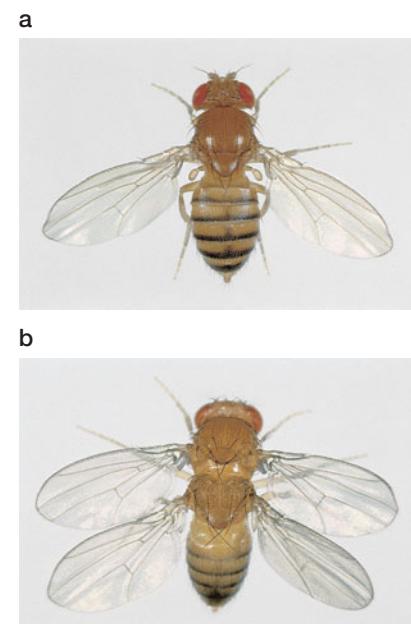
### Changes in Homeotic Gene Expression Are Responsible for Arthropod Diversity

The interdisciplinary field known as “evo–devo” lies at the cusp of two traditionally isolated areas of research: evolutionary biology and developmental biology. The impetus for evo–devo research is that genetic analysis of development in flies, nematode worms, and other model organisms has identified the key genes responsible for evolutionary diversity. The homeotic genes represent premiere examples of such genes.

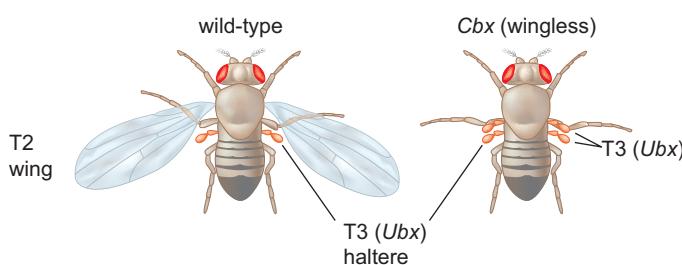
The *Drosophila* genome contains a total of eight homeotic genes organized in two gene clusters or complexes: the Antennapedia complex and the Bithorax complex (see Box 21-8, Homeotic Genes of *Drosophila* Are Organized in Special Chromosome Clusters). A typical invertebrate genome contains eight to 10 homeotic genes, usually located within just one complex. Vertebrates have duplicated the ancestral Hox complex and contain four clusters. Changes in the expression and function of individual homeotic genes are responsible for altering limb morphology in arthropods and the axial skeletons of vertebrates. We describe later how changes in *Ubx* activity have produced evolutionary modifications in insects and other arthropods.

### Changes in *Ubx* Expression Explain Modifications in Limbs among the Crustaceans

Crustaceans include most, but not all, of the arthropods that swim. Some live in the ocean, whereas others prefer fresh water. They include some of our favorite culinary dishes, such as shrimp, crab, and lobster. One of the



**FIGURE 21-25** *Ubx* mutants cause the transformation of the metathorax into a duplicated mesothorax. (a) A normal fly is shown that contains a pair of prominent wings and a smaller set of halteres just behind the wings. (b) A mutant that is homozygous for a weak mutation in the *Ubx* gene is shown. The metathorax is transformed into a duplicated mesothorax. As a result, the fly has two pairs of wings rather than one set of wings and one set of halteres. (Courtesy of E.B. Lewis.)



**FIGURE 21-26** Misexpression of *Ubx* in the mesothorax results in the loss of wings. The *Cbx* mutation disrupts the regulatory region of *Ubx*, causing its misexpression in the mesothorax, and results in its transformation into the metathorax.

## ► ADVANCED CONCEPTS

**Box 21-8** Homeotic Genes of *Drosophila* Are Organized in Special Chromosome Clusters

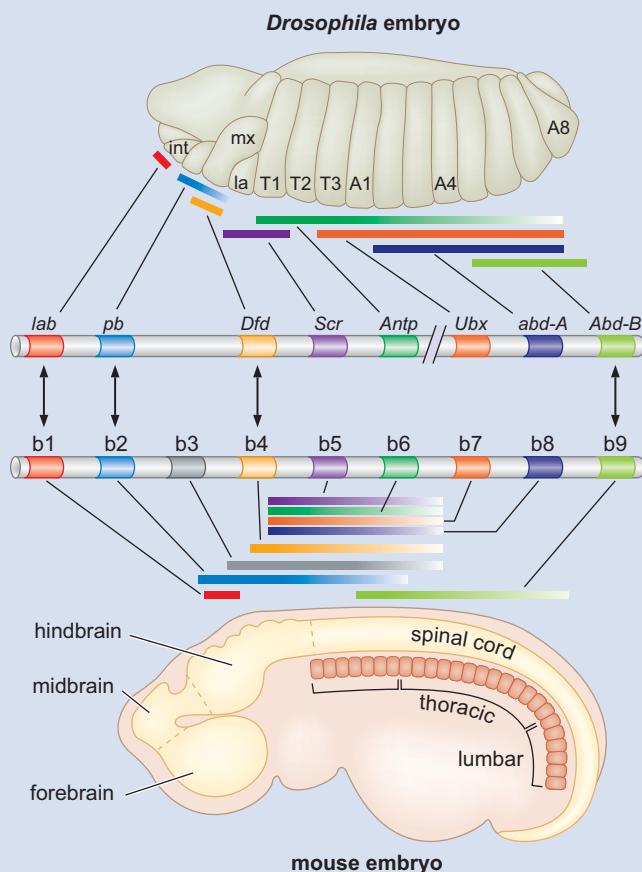
*Antp* and *Ubx* represent only two of the eight homeotic genes in the *Drosophila* genome. The eight homeotic genes of *Drosophila* are located in two clusters, or gene complexes. Five of the eight genes are located within the Antennapedia complex, and the remaining three genes are located within the Bithorax complex (see Box 21-8 Fig. 1). Do not confuse the names of the complex with the individual genes within the complex. For example, the Antennapedia complex is named in honor of the *Antennapedia* (*Antp*) gene, which was the first homeotic gene identified within the complex. There are four other homeotic genes in the Antennapedia complex: *labial* (*lab*), *proboscipedia* (*pb*), *Deformed* (*Dfd*), and *Sex combs reduced* (*Scr*). Similarly, the Bithorax complex is named in honor of the *Ultrabithorax* (*Ubx*) gene, but there are two others in this complex: *abdominal-A* (*abd-A*) and *Abdominal-B* (*Abd-B*). Another insect, the flour beetle, contains a single complex of homeotic genes that includes homologs of all eight homeotic genes contained in the *Drosophila* Antennapedia and Bithorax complexes. The two complexes probably arose from a chromosomal rearrangement within a single ancestral complex.

There is a collinear correspondence between the order of the homeotic genes along the chromosome and their patterns of expression across the anteroposterior axis in developing embryos (see Box 21-8 Fig. 1). For example, the *lab* gene, located in the 3'-most position of the Antennapedia complex, is expressed in the anteriormost head regions of the developing *Drosophila* embryo. In contrast, the *Abd-B* gene, which is located in the 5'-most position of the Bithorax complex, is expressed in the posteriormost regions (see Box 21-8 Fig. 1). The significance of this colinearity has not been established, but it must be important because it is preserved in each of the major groups of arthropods (including flour beetles), as well as all vertebrates that have been studied, including mice and humans.

#### Mammalian Hox Gene Complexes Control Anteroposterior Patterning

Mice contain 38 *Hox* genes arranged within four clusters (*Hoxa*, *Hoxb*, *Hoxc*, *Hoxd*). Each cluster or complex contains nine or 10 *Hox* genes and corresponds to the single homeotic gene cluster in insects that formed the Antennapedia and Bithorax complexes in *Drosophila* (Box 21-8 Fig. 2). For example, the *Hoxa-1* and *Hoxb-1* genes are most closely related to the *lab* gene in *Drosophila*, whereas *Hoxa-9* and *Hoxb-9*—located at the other end of their respective complexes—are similar to the *Abd-B* gene.

In addition to this “serial” homology between mouse and fly *Hox* genes, each mouse *Hox* complex shows the same type of colinearity as that seen in *Drosophila*. For example, *Hox* genes located at the 3' end of each complex, such as the *Hoxa-1* and *Hoxb-1*, are expressed in the anteriormost regions of developing mouse embryos (future hindbrain). In contrast, *Hox* genes located near the 5' end of each complex, such as *Hoxa-9* and *Hoxb-9*, are expressed in posterior regions of the embryo

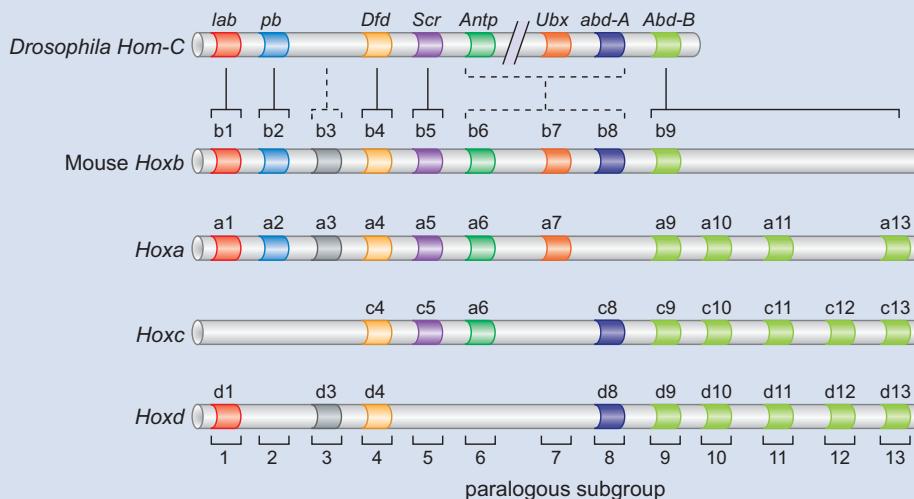


**BOX 21-8 FIGURE 1** Organization and expression of *Hox* genes in *Drosophila* and in the mouse. The figure compares the collinear sequences and transcription patterns of the *Hox* genes in *Drosophila* and in the mouse. (Adapted, with permission, from McGinnis W. and Krumlauf R. 1992. *Cell* 68: 283–302, Fig. 2. © Elsevier.)

(thoracic and lumbar regions of the developing spinal cord). The *Hoxd* complex shows sequential expression across the anteroposterior axis of the developing limbs. A comparable pattern is not observed in insect limbs, suggesting that the *Hoxd* genes have acquired “novel” regulatory DNAs during vertebrate evolution. Indeed, we have already seen in Chapter 19 that a specialized global control region (GCR) coordinates the expression of the individual *Hoxd* genes in developing limbs.

#### Altered Patterns of Hox Expression Create Morphological Diversity in Vertebrates

Mutations in mammalian *Hox* genes cause disruptions in the axial skeleton, which consists of the spinal cord and the different vertebrae of the backbone. These alterations are evocative of some of the changes in morphology we have seen for the *Antp* and *Ubx* mutants in *Drosophila*.

**Box 21-8** (Continued)

**BOX 21-8 FIGURE 2** Conservation of organization and expression of the homeotic gene complexes in *Drosophila* and in the mouse. (Adapted, with permission, from Gilbert S.E. 2000. *Developmental biology*, 6th ed., Fig. 11.36a. © Sinauer.)

Consider the *Hoxc-8* gene in mice, which is most closely related to the *abd-A* gene of the *Drosophila* Bithorax complex. It is normally expressed near the boundary between the developing rib cage and lumbar region of the backbone, the anterior “tail.” (The *abd-A* gene is expressed in the anterior abdomen of the *Drosophila* embryo.) The first lumbar vertebra normally lacks ribs. However, mutant embryos that are homozygous for a knockout mutation in the *Hoxc-8* gene show a dramatic mutant phenotype. The first lumbar vertebra develops an extra pair of vestigial ribs. This type of developmental abnormality is sometimes called a “homeotic” transformation, one in which the proper structure develops in the wrong place. In this case, a vertebra that is typical of the posterior thoracic region develops within the anterior lumbar region.

#### Maintenance of Hox Gene Expression Patterns

Localized patterns of *Hox* gene expression are established in early fly and mouse embryos by combinations of sequence-specific transcriptional activators and repressors. Some of these regulatory proteins are modulated by cell signaling pathways, such as the FGF and Wnt pathways. In *Drosophila*, many of the same gap repressors that establish localized stripes of *eve* expression also control the initial patterns of *Hox* gene expression. These patterns are maintained throughout the life cycle long after the gap repressors are lost.

Consider, as an example, the *Abd-B* *Hox* gene in *Drosophila*. It is specifically expressed in the posterior abdomen, including the primordia of the fifth through eighth abdominal segments. *Abd-B* expression is initially repressed by the *Hb*, *Kr*, and *Kni* gap repressors in the head, thorax, and anterior abdomen of the early *Drosophila* embryo. These are the same repressors that establish localized stripes of *eve* expression (see Figs. 21-19 and 21-20). These repressors restrict *Abd-B* expression to the posterior abdominal segments.

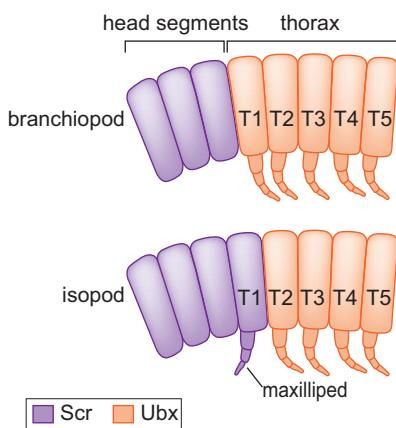
The maintenance of *Abd-B* expression, as well as the expression of most other *Hox* genes in flies and mammals, depends

on a large protein complex, called the Polycomb repression complex (PRC). The PRC binds to *Abd-B* regulatory sequences in cells that fail to activate the gene in the early embryo: the progenitors of the head, thorax, and anterior abdomen. In all of these cells, the PRC causes methylation of lysine 27 on histone H3, and this methylation correlates with the repression of the associated *Abd-B* transcription unit. Conversely, a ubiquitous activator complex, the Trithorax complex (TRC), binds to *Abd-B* regulatory sequences in cells that express the gene in the early embryo (i.e., the posterior abdominal segments). The binding of the TRC leads to the methylation of lysine 4 on histone H3, and this correlates with active transcription of *Abd-B*.

Thus, the PRC and TRC maintain on/off states of *Hox* gene expression depending on the initial expression patterns of these genes in the early embryo. If a given *Hox* gene is repressed in a particular cell, PRC binds and keeps the gene off in all the descendants of that cell. Conversely, if a given *Hox* gene is activated in a particular cell, then TRC will bind and ensure stable expression of the gene in all of its descendants. TRC and PRC serve to maintain a regulatory “memory” of *Hox* gene expression patterns.

#### MicroRNAs Modulate Hox Activity

Many *Hox* gene complexes contain microRNA (miRNA or miR) genes. For example, the fly *ANT-C* contains *miR-10*, and *BX-C* contains *miR-iab4*. The encoded miRNAs are thought to inhibit or attenuate the synthesis of different *Hox* proteins. The *iab4* miRNA inhibits *Ubx* protein synthesis in abdominal tissues. Vertebrate *Hox* complexes also contain miR genes, including *miR-10*. The *miR-196* gene is located in 5' regions of several vertebrate *Hox* complexes. The encoded miRNA is thought to inhibit the synthesis of the *Hoxb8* protein in posterior regions of mouse embryos. See Chapter 20 for more details regarding how miRNAs block or attenuate protein synthesis.



**FIGURE 21-27** Changing morphologies in two different groups of crustaceans. In branchiopods, *Scr* expression is restricted to head regions, where it helps promote the development of feeding appendages, whereas *Ubx* is expressed in the thorax, where it controls the development of swimming limbs. In isopods, *Scr* expression is detected in both the head and the first thoracic segment (T1), and as a result, the swimming limb in T1 is transformed into a feeding appendage (the maxilliped). This posterior expansion of *Scr* was made possible by the loss of *Ubx* expression in T1 because *Ubx* normally represses *Scr* expression. (Adapted from Levine M. 2002. *Nature* 415: 848–849, Fig. 2. © Macmillan.)

most popular groups of crustaceans for study is *Artemia*, also known as “sea monkeys.” Their embryos arrest as tough spores that can be purchased at toy stores. The spores quickly resume development upon addition of salt water.

The heads of these shrimp contain feeding appendages. The thoracic segment nearest the head, T1, contains swimming appendages that look like those further back on the thorax (the second through 11th thoracic segments, T2–T11). *Artemia* belongs to an order of crustaceans known as **branchiopods**. Consider a different order of crustaceans, called **isopods**. Isopods contain swimming limbs on the second through eighth thoracic segments, just like the branchiopods. But the limbs on the first thoracic segment of isopods have been modified. They are smaller than the others and function as feeding limbs (Fig. 21-27). These modified limbs are called maxillipeds (otherwise known as jaw feet), and look like appendages found on the head (although these are not shown in the figure).

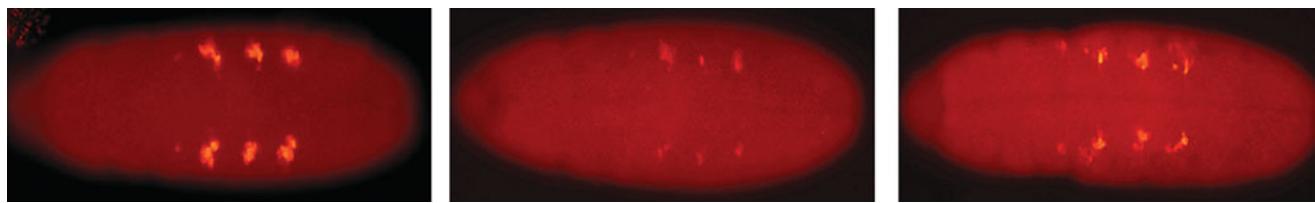
Slightly different patterns of *Ubx* expression are observed in branchiopods and isopods. These different expression patterns are correlated with the modification of the swimming limbs on the first thoracic segment of isopods. Perhaps the last shared ancestor of the present branchiopods and isopods contained the arrangement of thoracic limbs seen in *Artemia* (which is itself a branchiopod): all thoracic segments contain swimming limbs. During the divergence of branchiopods and isopods, the *Ubx* regulatory sequences changed in isopods. As a result of this change, *Ubx* expression was eliminated in the first thoracic segment and restricted to segments T2–T8. This shift in *Ubx* expression permitted the formation of a maxilliped in place of the T1 swimming limb. There is a tight correlation between the absence of *Ubx* expression in the thorax and the development of feeding appendages in different crustaceans. For example, lobster embryos lack *Ubx* expression in the first two thoracic segments and contain two pairs of maxillipeds. Cleaner shrimp lack *Ubx* expression in the first three thoracic segments and contain three pairs of maxillipeds.

### How Insects Lost Their Abdominal Limbs

All insects have six legs, two on each of the three thoracic segments; this applies to every one of the more than 1 million species of insects. In contrast, other arthropods, such as crustaceans, have a variable number of limbs. Some crustaceans have limbs on every segment in both the thorax and abdomen. This evolutionary change in morphology, the loss of limbs on the abdomen of insects, is not due to altered expression of pattern-determining genes, as seen in the case of maxilliped formation in isopods. Rather, the loss of abdominal limbs in insects is due to functional changes in the *Ubx* regulatory protein.

In insects, *Ubx* and *abd-A* repress the expression of a critical gene that is required for the development of limbs, called *Distal-less* (*Dll*). In developing *Drosophila* embryos, *Ubx* is expressed at high levels in the metathorax and anterior abdominal segments; *abd-A* expression extends into more posterior abdominal segments. Together, *Ubx* and *abd-A* keep *Dll* off in the first seven abdominal segments. Although *Ubx* is expressed in the metathorax, it does not interfere with the expression of *Dll* in that segment, because *Ubx* is not expressed in the developing T3 legs until after the time when *Dll* is activated. As a result, *Ubx* does not interfere with limb development in T3.

In crustaceans, such as the branchiopod *Artemia* already mentioned, there are high levels of both *Ubx* and *Dll* in all 11 thoracic segments. The expression of *Dll* promotes the development of swimming limbs. Why does *Ubx* repress *Dll* expression in the abdominal segments of insects but



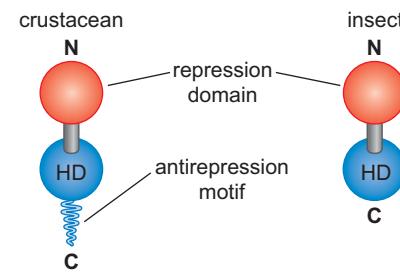
**FIGURE 21-28** Evolutionary changes in Ubx protein function. (a) The *Dll* enhancer (*Dll*304) is normally activated in three pairs of “spots” in *Drosophila* embryos. These spots go on to form the three pairs of legs in the adult fly. (b) The misexpression of the *Drosophila* Ubx protein (DmUbxHA) strongly suppresses expression from the *Dll* enhancer. (c) In contrast, the misexpression of the Ubx protein from the brine shrimp *Artemia* (AfUbxHA) causes only a slight suppression of the *Dll* enhancer. (Adapted, with permission, from Ronshaugen M. et al. 2002. *Nature* 415: 914–917, Fig. 2c. © Macmillan. Images courtesy of William McGinnis and Matt Ronshaugen.)

not crustaceans? The answer is that the Ubx protein has diverged between insects and crustaceans. This was shown in the following experiment.

The misexpression of *Ubx* throughout all of the tissues of the presumptive thorax in transgenic *Drosophila* embryos suppresses limb development because of the repression of *Dll* (Fig. 21-28). In contrast, the misexpression of the crustacean Ubx protein in transgenic flies does not interfere with *Dll* gene expression and the formation of thoracic limbs. These observations indicate that the *Drosophila* Ubx protein is functionally distinct from Ubx in crustaceans. The fly protein represses *Dll* gene expression, whereas the crustacean Ubx protein does not.

What is the basis for this functional difference between the two Ubx proteins? (They share only 32% overall amino acid identity, but their homeodomains are virtually identical—59/60 matches.) It turns out that the crustacean protein has a short motif containing 29 amino acid residues that blocks repression activity. When this sequence is deleted, the crustacean Ubx protein is just as effective as the fly protein at repressing *Dll* gene expression (Fig. 21-29).

Both the crustacean and fly Ubx proteins contain multiple repression domains. As discussed in Chapter 19, it is likely that these domains interact with one or more transcriptional repression complexes. The “antirepression” peptide present in the crustacean Ubx protein might interfere with the ability of the repression domains to recruit these complexes. When this peptide is attached to the fly protein, the hybrid protein behaves like the crustacean Ubx protein and no longer represses *Dll*.



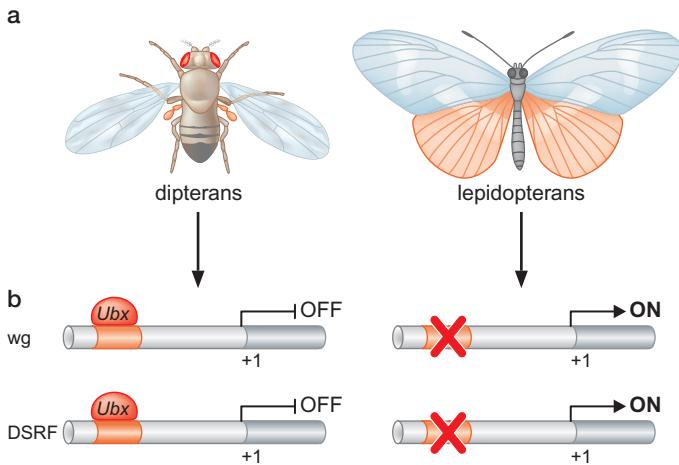
**FIGURE 21-29** Comparison of Ubx in crustaceans and in insects. (Left) Ubx in crustaceans. The carboxy-terminal antirepression peptide blocks the activity of the amino-terminal repression domain. (Right) Ubx in insects. The carboxy-terminal antirepression peptide was lost through mutation. (Adapted, with permission, from Ronshaugen M. et al. 2002. *Nature* 415: 914–917, Fig. 4b. © Macmillan.)

### Modification of Flight Limbs Might Arise from the Evolution of Regulatory DNA Sequences

Ubx has dominated our discussion of morphological change in arthropods. Changes in the Ubx expression pattern appear to be responsible for the transformation of swimming limbs into maxillipeds in crustaceans. Moreover, the loss of the antirepression motif in the Ubx protein likely accounts for the suppression of abdominal limbs in insects. In this final section on that theme, we review evidence that changes in the regulatory sequences in *Ubx* target genes might explain the different wing morphologies found in fruit flies and butterflies.

In *Drosophila*, Ubx is expressed in the developing halteres, where it functions as a repressor of wing development. Approximately five to 10 target genes are repressed by Ubx. These genes encode proteins that are crucial for the growth and patterning of the wings (Fig. 21-30), and all are expressed

**FIGURE 21-30** Changes in the regulatory DNA of *Ubx* target genes. (a) The *Ubx* repressor is expressed in the halteres of dipterans and hindwings of lepidopterans (orange). (b) Different target genes contain *Ubx* repressor sites in dipterans. These have been lost in lepidopterans.



in the developing wing. In *Ubx* mutants, these genes are no longer repressed in the halteres, and as a result, the halteres develop into a second set of wings.

Fruit flies are dipterans, and all of the members of this order contain a single pair of wings and a set of halteres. It is likely that *Ubx* functions as a repressor of wing development in all dipterans. Butterflies belong to a different order of insects, the lepidopterans. All of the members of this order (which also includes moths) contain two pairs of wings rather than a single pair of wings and a set of halteres. What is the basis for these different wing morphologies in dipterans and lepidopterans?

The two orders diverged from a common ancestor more than 250 million years ago. This is about the time of divergence that separates humans and nonmammalian vertebrates such as frogs. It would seem to be a sufficient period of time to alter *Ubx* gene function through any or all of the three strategies that we have discussed. The simplest mechanism would be to change the *Ubx* expression pattern so that it is lost in the progenitors of the hindwings in Lepidoptera. Such a loss would permit the developing hindwings to express all of the genes that are normally repressed by *Ubx*. The transformation of swimming limbs into maxillipedes in isopods provides a clear precedent for such a mechanism. However, there is no obvious change in the *Ubx* expression pattern in flies and butterflies; *Ubx* is expressed at high levels throughout the developing hindwings of butterflies.

That leaves us with two possibilities. First, the *Ubx* protein is functionally distinct in flies and butterflies. The second is that each of the approximately five to 10 target genes that are repressed by *Ubx* in *Drosophila* has evolved changes in its regulatory DNAs so that they are no longer repressed by *Ubx* in butterflies (see Fig. 21-30). It seems easier to modify repression activity than to change the regulatory sequences of five to 10 different *Ubx* target genes.

Surprisingly, it appears that the less likely explanation—changes in the regulatory sequences of several *Ubx* target genes—accounts for the different wing morphologies. The *Ubx* protein appears to function in the same way in fruit flies and butterflies. For example, in butterflies, the loss of *Ubx* in patches of cells in the hindwing causes them to be transformed into forewing structures (see Fig. 21-30a for the difference between forewings and hindwings). This observation suggests that the butterfly *Ubx* protein functions as a repressor that suppresses the development of forewings. Although not proven, it is possible that the regulatory DNAs of the

wing-patterning genes have lost the Ubx-binding sites (Fig. 21-30b) and they are no longer repressed by Ubx in the developing hindwing.

## GENOME EVOLUTION AND HUMAN ORIGINS

---

We now consider specific examples of comparative genome analysis, with a particular focus on the comparison of animal genomes. Our final discussion of the comparison of the Neanderthal genome with those of chimpanzee and human provides a few startling insights into human origins.

### Diverse Animals Contain Remarkably Similar Sets of Genes

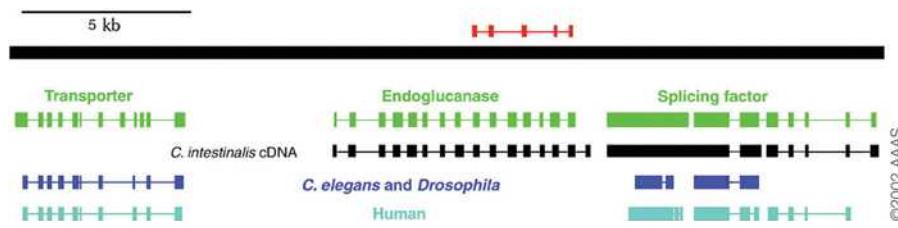
About 100 different animal genomes have been fully sequenced and assembled, but the majority of these sequences correspond to just a few animal groups, centered around the human genome, as well as those of key model organisms such as the fruit fly, *Drosophila melanogaster*, and the nematode worm, *Caenorhabditis elegans*. Thus, several primate genomes (chimpanzees, rhesus monkey, etc.) have been determined to help identify the distinctive features of the human genome (see later discussion). Twelve different species of *Drosophila* have been sequenced to help understand the diversification of distinct species of fruit flies. Currently, just one-third of all animal phyla are represented by a member species with a complete genome sequence assembly.

By far, the most startling discovery arising from comparative genome sequence analysis is the fact that wildly divergent animals, from the sea anemone to humans, possess a highly conserved set of genes. A typical invertebrate genome (e.g., sea anemone, worm, insect) contains approximately 15,000 protein-coding genes. Vertebrates contain a larger number, with an average of about 25,000 genes. However, this larger gene number is not generally due to the invention of “new” genes unique to vertebrates; rather, it is due to the duplication of “old” genes already present in invertebrate genomes. For example, invertebrates contain just a few copies of genes encoding a growth factor called fibroblast growth factor (FGF), whereas a typical vertebrate genome contains more than 20 different FGF genes.

A glimpse into the set of genes required for the distinctive attributes of all animals is provided by the genome sequence assembly of a single-cell eukaryote, a protozoan, called *Monosiga*. This organism is the closest living relative of modern animals. Yet, it lacks many of the genes required for animal development, including those encoding signaling molecules, such as Wingless, the transforming growth factor- $\beta$  (TGF- $\beta$ ), Hedgehog, and Notch. It also lacks critical regulatory genes responsible for differential gene activity in developing animal embryos, including *Hox* genes and *Hox* clusters. Thus, the evolutionary transition of simple eukaryotes into modern animals required the creation of a large number of novel genes not seen among the simple organisms that lived in the ancient oceans more than 1 billion years ago.

### Many Animals Contain Anomalous Genes

Despite a constant set, or “toolkit,” of basic genes required for the development of all animals, every genome contains its own distinctive—and sometimes surprising—attribute. Consider the case of the sea squirt. It contains a gene encoding cellulose synthase (Fig. 21-31). This enzyme is used by plants to produce cellulose, the major biopolymer of wood. It is absent in virtually all animals, so what is it doing in the sea squirt? The adult is immobile



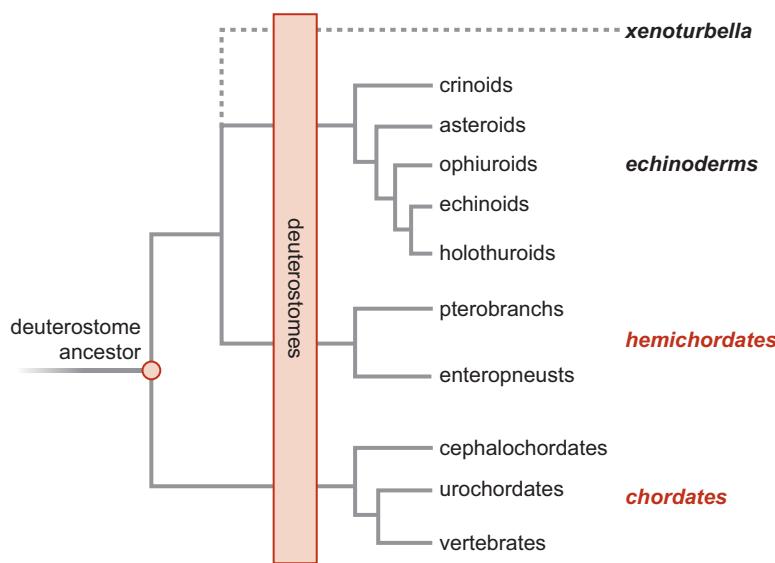
**FIGURE 21-31** A plant gene in the *Ciona* genome compared with sequences from other animals. A 20-kb region of one of the *Ciona* scaffolds is shown. This sequence contains an endoglucanase gene, which encodes an enzyme that is required for the degradation and synthesis of cellulose, a major component of plant cell walls. The red rectangles on top represent the *Kerrigan-1* gene of *Arabidopsis*. The gene finder program identified 15 putative exons in the *Ciona* gene, indicated as green rectangles. In reality, there is a 5' exon present in the cDNA (black rectangles below) that was missed by the computer program. Similarly, a flanking gene, which encodes an RNA splicing factor, is predicted to contain a small intron in a large coding region, whereas the cDNA sequence suggests that there is no intron. There is also a discrepancy in the size of the 5'-most exon. The flanking genes are conserved in worms, flies, and humans, whereas the endoglucanase gene is unique to *Ciona*, which contains a cellulose sheath. Note differences in the detailed intron–exon structures of the flanking genes among the different animal genomes. (Reprinted, with permission, from Dehal et al. 2002. *Science* 298: 2157–2167, Fig. 8. © AAAS.)

and sits in tide pools where it filters seawater. It contains a rubbery protective sheath composed of tunicin, a biopolymer related to plant cellulose. However, prior to the genome assembly, it was unclear whether the sea squirt contained its own endogenous cellulose synthase gene or employed a symbiotic organism for producing the tunicin sheath. Indeed, there are numerous examples of animals using simple symbionts for unusual genetic functions. For example, termites and wood-eating cockroaches contain symbiotic bacteria in their hindguts that contain the necessary genes required for digesting wood.

Another surprise came from the analysis of the sea urchin genome; it contains two genes, *RAG1* and *RAG2*, required for the rearrangement of immunoglobulin genes in humans and other vertebrates (see Chapter 12). One of the distinctive attributes of vertebrates is the ability to mount an adaptive immune response upon infection or injury. This includes the production of specific antibodies that recognize foreign antigens with great specificity and precision. Invertebrates possess a general innate immunity, but they lack the capacity to produce an adaptive immune response. Prior to the sea urchin genome assembly, it was thought that an ancestor of the modern vertebrates acquired a virus or transposon containing the *RAG1* and *RAG2* genes. However, the identification of these genes in sea urchins suggests that this is not true. Instead, the *RAG* genes were acquired by a much more distant ancestor, a progenitor of the so-called Deuterostomes, which diverged into modern echinoderms (e.g., sea urchins) and chordates (e.g., vertebrates) (see Fig. 21-32). It would appear that several descendants of this hypothetical ancestor, such as sea squirts, lost the *RAG* genes.

### Synteny Is Evolutionarily Ancient

One of the striking findings of comparative genome analysis is the high degree of **synteny**, conservation in genetic linkage, between distantly related animals. There is extensive synteny between mice and humans. In many cases, this linkage even extends to the pufferfish, which last shared a common ancestor with mammals more than 400 million years ago. What is even more remarkable is that some of the linkage relationships are conserved between

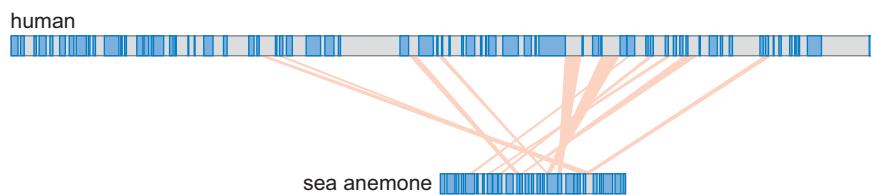


**FIGURE 21-32 Deuterostome phylogeny.** The deuterostomes include four animal phyla: Xenoturbellida, Echinodermata, Hemichordata, and Chordata. There are five classes of organisms within the echinoderms, two classes of hemichordates, and three classes of chordates. Note that the closest living relatives of the vertebrates are the urochordates, which include the sea squirts (see Box 19-3). (Adapted, with permission, from Gerhart J. 2006. *J. Cell Physiol.* **209**: 677–685. © Wiley-Liss, Inc.)

humans and simple invertebrates, such as sea anemones, which last shared a common ancestor more than 700 million years ago, well before the Cambrian radiation that produced most of the modern animal phyla (Fig. 21-33).

Genetic linkage is essential in prokaryotes, where linked genes are coregulated within a common operon (Chapter 18). Such linkage is generally absent in metazoan genomes, although the nematode worm *C. elegans* has been shown to contain a few operons. In other words, neighboring genes are no more likely to be coexpressed (e.g., in blood cells) than unlinked genes. Early comparative genome analyses appeared to confirm that genetic linkage bore no impact on gene regulation. For example, there is no obvious synteny in the arrangement of related genes in mammalian genomes (e.g., mouse and human) and invertebrate genomes such as *C. elegans* and *Drosophila*. However, there is emerging evidence that the genomes of nematode worms and fruit flies are highly “derived.” That is, they have undergone distinctive rearrangements and changes not seen in other genomes. Evidence for this view stems from the analysis of the genome of *Nematostella*, a simple sea anemone.

Sea anemones are ancient creatures. They appear in pre-Cambrian fossils, before the first appearance of Arthropods (e.g., trilobites) and annelids. Despite their simplicity and ancient history, they contain several genes that have been lost in flies and worms. What is even more remarkable is that about half of the genetic linkages seen in the human genome are re-



**FIGURE 21-33 Conservation of genetic linkage between sea anemones and humans.** The top diagram shows a 4-Mb region of human chromosome 10 (the q24 region). The lines show alignments between 11 different genes in this interval and corresponding sequences within a 1-Mb region of a sea anemone chromosome. All 11 genes are located together in both chromosomes, but the exact order of the genes has changed during the course of the ~700 million years since humans and sea anemones last shared a common ancestor.

tained, albeit in a somewhat scrambled order, in the *Nematostella* genome (Fig. 21-33). Consider the q24 region of human chromosome 10. This region contains 11 genes within a 4-Mb interval, including the gene for actin and *SLK*, which encodes a kinase required for cell division. In the smaller *Nematostella* genome, these 11 genes are not only present but also linked within a 1-Mb interval. The conservation of this local synteny raises the possibility that linkage might influence gene function in some subtle manner, which we are currently unable to explain. By sequencing additional animal genomes, particularly those representing ancient creatures such as sponges and flatworms, it might be possible to reconstruct the ancestral karyotype—the exact chromosome complement and genetic linkages of the metazoan ancestor that generated all the modern animal phyla seen today.

### Deep Sequencing Is Being Used to Explore Human Origins

The ability to sequence large quantities of DNA quickly and inexpensively has created an opportunity to perform experiments that were impossible to imagine even a year ago. One recent example concerns the analysis of the Neanderthal genome.

Modern humans appeared approximately 100,000 years ago and last shared a common ancestor with Neanderthals about 500,000 years ago. There is evidence that modern humans and Neanderthals coexisted in certain locations prior to the disappearance of the Neanderthals about 30,000 years ago. It has been suggested that the two groups mated, resulting in the occurrence of at least some “Neanderthal genes” in the modern human genome. To test this possibility, scientists have recently determined the complete sequence of the Neanderthal genome.

Neanderthal DNA samples have been obtained from well-preserved fossils. However, the DNA is heavily contaminated with bacteria and fungi. Nonetheless, the ability to generate hundreds of thousands of short DNA sequence “reads” (see Chapter 7) permits the identification of authentic Neanderthal DNA among the mixture of contaminating DNAs. In fact, just 2%–3% of the total DNA obtained from a well-preserved Neanderthal fossil corresponds to authentic Neanderthal DNA that matches chimpanzee and human reference genome sequences. The detailed comparison of these Neanderthal sequences with the chimpanzee and human genomes suggests that there was indeed comingling of Neanderthals and modern humans. It is amazing to think that the genomes of extinct organisms can be “resurrected.”

## SUMMARY

---

The cells of a developing embryo follow divergent pathways of development by expressing different sets of genes. Most differential gene expression is regulated at the level of transcription initiation. There are three major strategies: mRNA localization, cell-to-cell contact, and the diffusion of secreted signaling molecules.

mRNA localization is achieved by the attachment of specific 3'-UTR sequences to the growing ends of microtubules. This mechanism is used to localize the *ash1* mRNA to the daughter cells of budding yeast. It is also used to localize the *oskar* mRNA to the posterior plasm of the unfertilized egg in *Drosophila*.

In cell-to-cell contact, a membrane-bound signaling molecule alters gene expression in neighboring cells by activating

a signaling pathway. In some cases, a dormant transcriptional activator, or coactivator protein, is released from the cell surface into the nucleus. In other cases, a quiescent transcription factor (or transcriptional repressor) already present in the nucleus is modified so that it can activate gene expression. Cell-to-cell contact is used by *B. subtilis* to establish different programs of gene expression in the mother cell and forespore. A remarkably similar mechanism is used to prevent skin cells from becoming neurons during the development of the insect central nervous system.

Extracellular gradients of secreted cell-signaling molecules can establish multiple cell types during the development of a complex tissue or organ. These gradients produce intracellular gradients of activated transcription factors,

which, in turn, control gene expression in a concentration-dependent fashion. An extracellular Sonic hedgehog gradient leads to a Gli activator gradient in the ventral half of the vertebrate neural tube. Different levels of Gli regulate distinct sets of target genes and thereby produce different neuronal cell types. Similarly, the Dorsal gradient in the early *Drosophila* embryo elicits different patterns of gene expression across the dorsoventral axis. This differential regulation depends on the binding affinities of Dorsal-binding sites in the target enhancers.

The segmentation of the *Drosophila* embryo depends on a combination of localized mRNAs and gradients of regulatory factors. Localized *bicoid* and *oskar* mRNAs, at the anterior and posterior poles, respectively, lead to the formation of a steep Hunchback repressor gradient across the anteroposterior axis. This gradient establishes sequential patterns of Krüppel, Knirps, and Giant in the presumptive thorax and abdomen. These four proteins are collectively called gap proteins; they function as transcriptional repressors that establish localized stripes of pair-rule gene expression. Individual stripes are regulated by separate enhancers located in the regulatory regions of pair-rule genes such as *eve*. Each enhancer contains multiple binding sites for both activators and gap repressors. It is the interplay of broadly distributed activators, such as Bicoid, and localized gap repressors that establish the anterior and posterior borders of individual pair-rule stripes. Separate stripe enhancers work independently of one another to produce composite, seven-stripe patterns of pair-rule expression. This enhancer autonomy is due, in part, to short-range transcriptional repression. A gap repressor bound to one enhancer does not interfere with the activities of a neighboring stripe enhancer located in the same gene.

Homeotic genes encode regulatory proteins responsible for making the individual body segments distinct from one another. The two best-studied homeotic genes, *Antp* and *Ubx*, control the development of the second and third thoracic segments, respectively, of the fruit fly. The misexpression of *Ubx* in the developing wings causes the development of wingless flies, whereas the misexpression of *Antp* in the head causes a transformation of antennae into legs.

In terms of sheer numbers and diversity, the arthropods can be considered the most successful of all animal phyla. More is known regarding the molecular basis of arthropod diversity than any other group of animals. For example, changes in the expression profile of the *Ubx* gene are correlated with the conversion of swimming limbs into maxillipedes in different groups of crustaceans. Functional changes in the *Ubx* protein might account for the repression of abdominal limbs in insects. Finally, changes in *Ubx* target enhancers might explain the different morphologies of the halteres in dipterans and the hindwings of butterflies.

Whole-genome assemblies of diverse animal groups reveal remarkable conservation of the core “genetic toolkit.” Most animal genomes contain a similar set of genes, and most differences result from duplication and divergence of “old” genes rather than the invention of new genes. Not only are most genes conserved in different animal groups, but there is also conservation of genetic linkage, or synteny. As many as one-half of all the genes in the human genome are located near the same neighbors in highly divergent animal groups such as sea anemones. Whole-genome assemblies are being used to obtain insights into our own human origins. A comparison of the chimpanzee and Neanderthal genomes suggest that modern humans contain significant contributions from “extinct” Neanderthals.

## BIBLIOGRAPHY

### Books

- Carroll S.B., Grenier J.K., and Weatherbee S.D. 2005. *From DNA to diversity: Molecular genetics and the evolution of animal design*, 2nd ed. Blackwell, Malden, Massachusetts.
- Gilbert S.F. 2010. *Developmental biology*, 9th ed. Sinauer Associates, Sunderland, Massachusetts.
- Wolpert L. and Tickle C. 2010. *Principles of development*, 4th ed. Oxford University Press, New York.

### mRNA Localization

- Macdonald P.M. 2011. mRNA localization: Assembly of transport complexes and their incorporation into particles. *Curr. Opin. Genet. Dev.* **21**: 407–413.
- Martin K.C. and Ephrussi A. 2009. mRNA localization: Gene expression in the spatial dimension. *Cell* **136**: 719–730.
- Medioni C., Mowry K., and Besse F. 2012. Principles and roles of mRNA localization in animal development. *Development* **139**: 3263–3276.

### Cell-to-Cell Contact

- Barad O., Hornstein E., and Barkai N. 2011. Robust selection of sensory organ precursors by the Notch-Delta pathway. *Curr. Opin. Cell Biol.* **23**: 663–667.

Schweisguth F. 2004. Notch signaling activity. *Curr. Biol.* **14**: R129–R138.

Shapiro L., McAdams H.H., and Losick R. 2002. Generating and exploiting polarity in bacteria. *Science* **298**: 1942–1946.

### Morphogens

- Ashe H.L. and Briscoe J. 2006. The interpretation of morphogen gradients. *Development* **133**: 385–394.
- Rogers K.W. and Schier A.F. 2011. Morphogen gradients: From generation to interpretation. *Annu. Rev. Cell Dev. Biol.* **27**: 377–407.
- Roth S. and Lynch J. 2012. Does the Bicoid gradient matter? *Cell* **149**: 511–512.

### Developmental Enhancers

- Arnosti D.N. and Kulkarni M.M. 2005. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J. Cell Biochem.* **94**: 890–898.
- Jaeger J. and Reinitz J. 2006. On the dynamic nature of positional information. *BioEssays* **28**: 1102–1111.
- Levine M. 2010. Transcriptional enhancers in animal development and evolution. *Curr. Biol.* **20**: R754–R763.

## Segmentation

Lemons D. and McGinnis W. 2006. Genomic evolution of Hox gene clusters. *Science* **313**: 1918–1922.

Lewis E.B. 1978. A gene complex controlling segmentation in *Drosophila*. *Nature* **276**: 565–570.

Tschopp P. and Duboule D. 2011. A genetic approach to the transcriptional regulation of Hox gene clusters. *Ann. Rev. Genet.* **45**: 145–166.

## QUESTIONS

## MasteringBiology®

For instructor-assigned tutorials and problems, go to *MasteringBiology*.

For answers to even-numbered questions, see Appendix 2: Answers.

**Question 1.** Define differential gene expression.

**Question 2.** Explain the significance of induced pluripotent stem (iPS) cells that function like pluripotent inner cell mass (ICM) cells.

**Question 3.** Describe three strategies for establishing differential gene expression during development.

**Question 4.** Outline the general steps of differential gene expression induced by concentration-dependent morphogens.

**Question 5.** Which strategy for differential gene expression do *Saccharomyces cerevisiae* cells use in the regulation of mating-type switching? Name the relevant mRNA or protein used in the strategy.

**Question 6.** Which strategy for differential gene expression do *Bacillus subtilis* cells use when the forespore influences gene expression in the mother cell? Name the relevant mRNA or protein used in the strategy.

**Question 7.** The Dorsal protein controls the dorsoventral patterning for the early embryo of *Drosophila melanogaster*. Explain how the number and type of Dorsal-binding sites in 5' regulatory DNAs relate to thresholds of gene expression.

**Question 8.** Describe an experiment that showed that the 3' UTR of *bicoid* and *oskar* mRNAs is required for proper localization in the *Drosophila* oocyte.

**Question 9.** Explain how Bicoid and Nanos set up a gradient of Hunchback protein in the embryo to ensure proper division of the embryo into segments.

**Question 10.** Consider *eve* stripes 1–7. Suggest a reporter assay to test that the enhancer for stripe #2 in the 5' regulatory region of the *eve* gene is necessary and sufficient for proper expression of stripe #2.

**Question 11.** Review Figure 21–20 for the wild-type expression of the *eve* stripes. If the enhancer for *eve* stripes #3 and #7 in the 5' regulatory region of the *eve* gene is deleted, predict the pattern of *eve* stripes in the embryo.

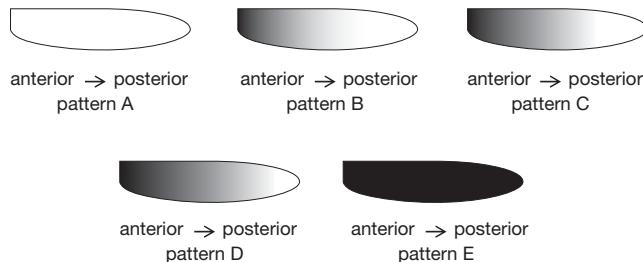
**Question 12.** Review Figure 21–21. Predict why the concentration of Bicoid and Giant drop sharply rather than gradually from the anterior to posterior position along the embryo.

**Question 13.** Explain how *Drosophila* legs could be expressed instead of antennae.

**Question 14.** Why are many signaling proteins considered toolkit genes?

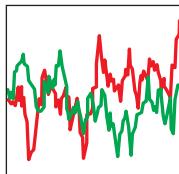
**Question 15.** *Bicoid* mRNA is maternally derived and localized to the anterior of early *Drosophila* embryos. Bicoid protein is a concentration-dependent transcriptional activator of the *hunchback* gene.

Depicted below are embryos in which a reporter is expressed under the control of the wild-type *hunchback* promoter. The darkened regions indicate reporter expression. Use these patterns to answer the following questions.



- Which expression pattern would you expect in embryos from a mother that does not express *bicoid* mRNA? Explain.
- Drosophila* flies are diploid organisms. This results in a pattern of reporter expression that most closely resembles pattern C. Would you expect this pattern to change if the mother had only one copy of the *bicoid* gene? If no, explain why not. If yes, explain and select the most likely expression pattern from the choices above.
- The *hunchback* promoter contains high-affinity and low-affinity binding sites for Bicoid protein. Which expression pattern would you expect if the *hunchback* promoter was mutated so that it now contains only the low-affinity binding sites for Bicoid protein and the mother has two copies of the *bicoid* gene? Explain your choice.

CHAPTER 22



## Systems Biology

TECHNOLOGICAL ADVANCES HAVE TRANSFORMED THE nature of molecular biology. It is now possible to identify every component—every gene and protein—engaged in a complex cellular process such as the differentiation of a naive stem cell into heart muscles. Before the advent of large-scale DNA-sequencing technologies and proteomics methods, molecular biologists sought to obtain general principles from the systematic dissection of just a subset of the total components—those believed to be the key rate-limiting regulatory agents of the process under study. The ability to identify and characterize every component of a process provides the opportunity for a new line of inquiry: what are the underlying design principles? In this chapter, we discuss the emerging discipline of systems biology, which has arisen from the marriage of traditional experimental molecular biology and computational analysis.

Molecular biology owes its success to tackling relatively simple systems, making it possible to investigate underlying mechanisms in great detail. This traditional approach has begun to give way, however, to more ambitious, holistic strategies in which higher and more complex levels of biological organization are examined by a combination of quantitative and high-throughput measurements, modeling, reconstruction, and theory. This interdisciplinary line of investigation has come to define the emerging field of systems biology. Systems biology draws on mathematics, engineering, physics, and computer science, as well as molecular and cellular biology. The objective is to describe the emergent properties of the web of interactions that govern the workings of living things and to do so in a manner that is quantitative and predictive. This approach can apply to biological systems operating at many levels, such as information transfer, signal transduction, cell division, and cytoskeleton dynamics. Here, and as appropriate for a text on the molecular biology of the gene, we focus on the systems biology of gene regulatory circuits. The hope is that this approach will reveal principles of gene control that cannot be understood from the study of individual components in isolation.

Systems biology is closely allied with another new field, that of **synthetic biology**. Like systems biology, synthetic biology seeks to elucidate design principles of biological circuits. However, synthetic biology attempts to do so by the creation of artificial networks that mimic the features of natural pathways of gene control. This approach enables us to test our models for how regulatory systems work. By virtue of their relative simplicity, such artificial networks can be analyzed in a more quantitative manner than the generally more complex regulatory circuits found in natural systems.

Systems biology is of high interest not only in what it tells us regarding the logic of gene control, but also in the context of evolution. The principal

### O U T L I N E

Regulatory Circuits, 776



Autoregulation, 776



Bistability, 780



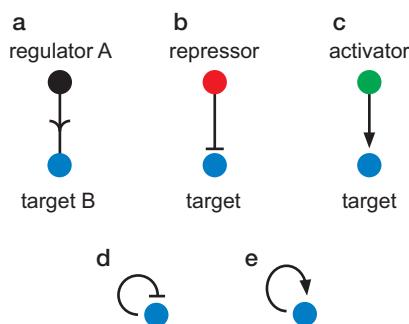
Feed-Forward Loops, 784



Oscillating Circuits, 786



Visit Web Content for Structural  
Tutorials and Interactive Animations



**FIGURE 22-1** Simple networks consisting of nodes and edges. (a) A simple switch. Two versions of the switch are shown with negative (b) and positive (c) signs. (d) Negative autoregulation; (e) positive autoregulation.

driving force in the evolution of higher organisms is, as we saw in Chapter 19, in the changes to networks that govern gene expression, not in the genes themselves. For example, animals have similar sets of genes but express these genes in different places and at different times (sometimes radically different). In other words, regulatory networks are relatively plastic in evolution, whereas the genes they control are relatively static.

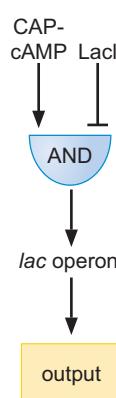
Here, we present a brief introduction to systems biology and synthetic biology with particular emphasis on natural and reconstructed circuits of gene control. We concentrate on the principles of the design of genetic circuits and on an intuitive understanding of the behavior of alternative wiring diagrams for gene control, but not on the detailed mathematics underlying much of this field. Systems biology is a new field, but, as we shall see, some of its principles derive from classic studies of gene control, particularly those presented in Chapter 18. In this chapter, however, we introduce the formalized language that helps extend these simple examples to a range of biologically diverse regulatory systems.

## REGULATORY CIRCUITS

**Regulatory circuits** can be described as simple networks consisting of **nodes** and **edges**. Nodes are the genes and are represented by dots; edges represent the regulation of one gene by the product of another and are shown by lines (Fig. 22-1a). Edges can convey directionality to indicate whether A regulates B or vice versa. Edges can also have signs to indicate whether the regulation is negative or positive. Thus, a line ending with a “ $\perp$ ,” extending from gene A to gene B, indicates that the gene A product is a negative regulator of gene B (Fig. 22-1b). Conversely, a line with an arrowhead extending from gene A to gene B indicates that the gene A product acts positively on the expression of gene B (Fig. 22-1c).

Let us begin with a simple two-node switch in which the product of gene A controls the expression of gene B (Fig. 22-1a). Thus, in response to a signal, the regulatory protein encoded by gene A triggers the expression of gene B. The regulatory protein can be a repressor; in which case, transcription is triggered by the presence of an inducer, which inactivates the repressor (Fig. 22-1b). Alternatively, the regulator can be an activator whose ability to trigger transcription occurs in response to a signaling molecule (Fig. 22-1c).

The lactose operon (described in Chapter 18, Fig. 18-6) is governed by two regulatory proteins and provides interwoven examples of both kinds of regulation: transcription is triggered by the presence of an inducer, which inactivates the LacI repressor, and by a rise in the concentration of cAMP, which promotes the binding of the CAP (catabolite activator protein) activator to DNA. Thus, the lactose operon is not a simple switch: its expression requires *both* the absence of repression and the presence of CAP bound by its ligand cAMP. This is the logic of an “AND gate,” a term from electrical engineering that denotes that two input conditions must be met in order for there to be an output. Here, the conditions are relief of repression and positive activation. AND gates are depicted by the symbol shown in Figure 22-2.



**FIGURE 22-2** The AND gate. The lactose operon is subject to the logic of an AND gate in which the output (transcription of the operon) requires both the presence of CAP-cAMP AND the absence of LacI repressor.

## AUTOREGULATION

Frequently, regulatory genes control their own transcription as well as the transcription of other target genes. This control is known as **autoregulation**,

and its sign can be either negative or positive, each with its own characteristic properties (Fig. 22-1d,e).

### Negative Autoregulation Dampens Noise and Allows a Rapid Response Time

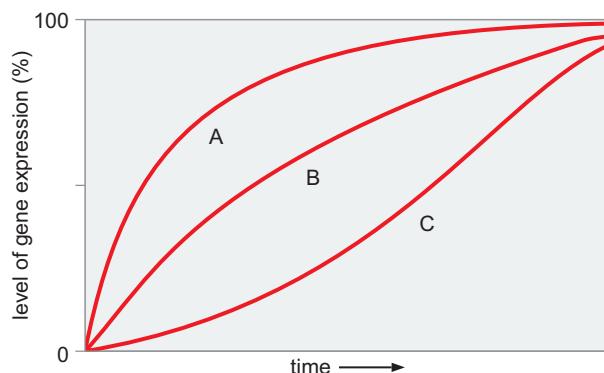
We consider negative autoregulation first; in this case, the gene for a repressor is negatively controlled by its own product. A classic example of negative autoregulation is the *cI* gene of bacteriophage  $\lambda$ , which we discussed in Chapter 18. (Recall that *cI* is also an example of positive autoregulation, as we discuss later.) Thus, the binding of the CI repressor to operator site  $O_{R3}$  blocks transcription of its own gene (Chapter 18, Fig. 18-27).

What is the biological significance of negative autoregulation, and why has it been repeatedly selected in evolution? One explanation was presented in Chapter 18: negative autoregulation is a homeostasis mechanism ensuring that the level of the regulatory protein is held at a constant level. Thus, should the level of CI fall sufficiently low to relieve repression of *cI* and other target genes, the resulting increase in transcription would raise the cellular concentration of repressor and restore repression. Conversely, should expression of the gene overshoot and produce more repressor than needed, then negative autoregulation will ensure that the gene is kept silent while the level of repressor falls through dilution during cell growth and division or by proteolytic degradation or both.

Negative autoregulation provides another, perhaps less obvious, benefit: rapid response time. Consider the opposing imperatives of producing repressor as rapidly as possible but not wastefully producing excess repressor. The use of a strong promoter would ensure rapid production but would, in steady state, lead to overaccumulation; on the other hand, the use of a comparatively weak promoter could achieve the proper level of repressor but would take a long time to do so. Negative autoregulation allows the best of both worlds: a relatively strong promoter can be used to drive rapid accumulation of the regulatory protein, whereas self-inhibition of transcription shuts off excess accumulation when the appropriate level of repressor is reached. Both mathematical modeling and experiments have confirmed that negative autoregulation allows a more rapid response for the same level of protein accumulation than simple regulation (Fig. 22-3).

### Gene Expression Is Noisy

Implicit in our discussion of the role of negative autoregulation in homeostasis is the concept of *noise* in gene expression. Until recently, it was assumed

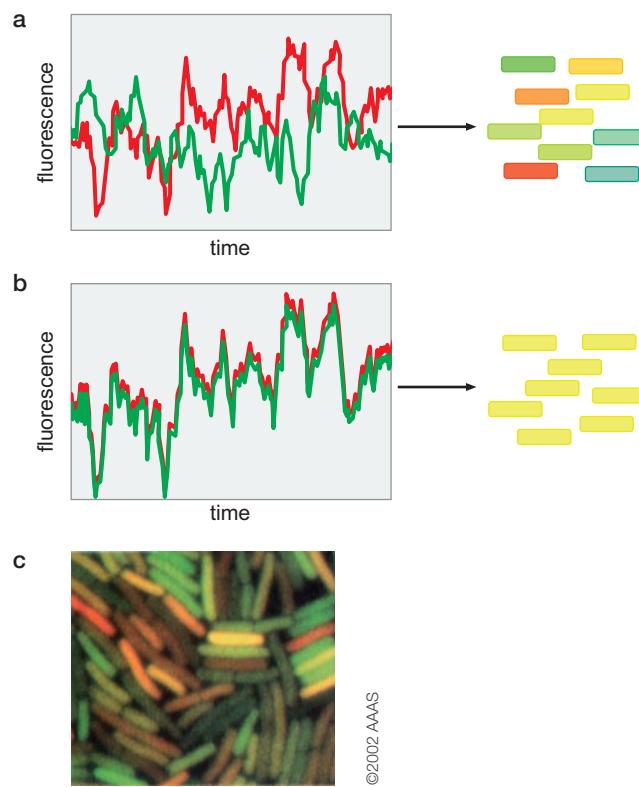


**FIGURE 22-3** Kinetics in response to an inducer. A simple switch (B), a negative autoregulatory switch (A), and a positive autoregulatory switch (C) respond with different kinetics to an inducing signal.

that the level of expression of a gene in a homogeneous population of cells is relatively constant from cell to cell. Now, however, we appreciate that the levels of gene expression vary substantially among individuals in a population and even between two copies of the same gene in the same cell. We therefore define **noise** as the variation in gene expression under seemingly uniform conditions. The existence of noise indicates that stochasticity influences the level of expression of individual genes. **Stochasticity** indicates that a process is characterized to some degree by randomness. As we shall see later, some regulatory motifs are designed to cope with noise, and other motifs are designed to exploit it.

Noise in gene expression comes from two sources—**intrinsic** and **extrinsic**; both lead to differences in gene expression in a population. **Intrinsic noise** refers to variation in the level of expression of individual genes within a cell and is due to stochastic events within the machinery for gene expression. A classic experiment that demonstrates intrinsic noise uses *Escherichia coli* cells harboring two copies of the same gene. One copy of the gene is joined to a reporter encoding a red fluorescent protein and the other to a reporter encoding a green fluorescent protein. Absent intrinsic noise, both gene copies should produce equal amounts of the red and green fluorescent proteins, and hence the cells should be yellow. What is observed instead is that many cells are conspicuously red and others conspicuously green (Fig. 22-4a). Thus, the level of expression of each gene is not identical within any given cell. That is, in some cells, one copy of the gene (e.g., the one tagged with the green fluorescent protein) is more actively expressed than the other copy (the one tagged with the red fluorescent protein); in other cells, the opposite is true.

**Extrinsic noise** refers to differences in gene expression between cells in a seemingly homogeneous population or to changes in gene expression in the same cell over time. This noise is likely caused by microheterogeneity in the



**FIGURE 22-4** Intrinsic versus extrinsic noise. Cells of *E. coli* harboring two copies of the same gene—in one case fused to a reporter generating a red fluorescent protein and in the other to a reporter generating a green fluorescent protein. (a) The predicted results if both genes varied in expression both over time and among individuals within the same cell (intrinsic noise). (b) The predicted results if the level of expression of both genes varied in synchrony over time because of extrinsic noise. (c) A fluorescent micrograph documents intrinsic noise from the observation, that in addition to yellow cells, some cells are red and others green. (Reprinted, with permission, from Elowitz M.B. et al. 2002. *Science* 297: 1184–1186. © AAAS.)

environment of individual cells or by fluctuations in the capacity of cells to perform transcription or protein synthesis over time. An example of extrinsic noise is illustrated in Figure 22-4b in which the level of expression of both genes is seen to vary over time. In this case, the level of expression of both genes in an individual cell varies in unison, rising for a period and then falling. This means that the overall capacity of individual cells to support the expression of some or all genes fluctuates with time. In the fluorescent micrograph shown in Figure 22-4c, we observe that some cells are yellow (meaning that they are expressing both the red and green reporting fusions), whereas others are red (meaning that they are expressing only the red reporter fusion) or green (meaning that they are expressing only the green reporter fusion), as a result of intrinsic noise.

Returning to negative autoregulation, we see that this regulatory motif helps cells cope with noise by allowing cells to compensate for variations in the level of expression of the autoregulated gene. The negative autoregulatory circuit governing bacteriophage  $\lambda$  CI synthesis is therefore said to be “robust.” **Robustness** indicates that the output of a regulatory circuit is insensitive to a particular parameter. Thus, the ability of the CI autoregulatory circuit to achieve a steady-state level of repressor is robust with respect to noise in the expression of the *cI* gene. As we shall see later, other regulatory motifs also help cells meet the challenge of coping with various sources of stochasticity, such as the fluctuations in signals that trigger gene expression.

Stochasticity is not a peculiarity of *E. coli*. Indeed, it is likely to be widespread among living things. For example, the coat pattern of a cat cloned by transfer of a somatic nucleus into an enucleated embryonic stem cell is not identical to that of the cat from which its genome was derived. Because both the clone and the cat from which it was derived are genetically identical, one might have expected that the cats would have identical coat patterns. That they do not suggests that the cascade of genetic events governing coat pattern is not wholly hardwired and must involve stochastic processes. As a second example, the fingerprints of identical twins are not identical.

### Positive Autoregulation Delays Gene Expression

Positive autoregulation occurs when an activator protein stimulates the transcription of its own gene (Fig. 22-1e). Once again, the *cI* gene of bacteriophage  $\lambda$  provides a classic but complex example: at low cellular concentrations, the CI repressor preferentially occupies the  $O_{R2}$  and  $O_{R1}$  operators that lie just upstream of the promoter ( $P_{RM}$ ) that drives *cI* transcription (Chapter 18, Fig. 18-26). CI protein sitting at  $O_{R2}$  contacts RNA polymerase to stimulate transcription, thereby promoting more CI synthesis. Of course, and as we have seen, when CI reaches high levels, it also occupies  $O_{R3}$  to repress transcription. Therefore, the *cI* gene is subject to both positive and negative autoregulation.

Now let us consider the case of a gene for an activator protein that is subject to positive autoregulation alone (see Fig. 22-1e). The steady-state accumulation of the gene product occurs when the rate of synthesis of the protein is in balance with the loss of the protein through degradation (should it be unstable) or its dilution through growth and division of the cell. Thus, “**steady state**” refers to a condition in which the level of the gene product varies only negligibly over time. The important point is that the time required to reach steady state after a gene is switched on is longer for the case of positive autoregulation than for the case of negative autoregulation or for no feedback at all (Fig. 22-3). Or, to be more precise, the time at which half-maximal accumulation occurs is longer for positive autoregulation than

for the alternative regulatory switches. This is because the rate of production, which increases over time, depends on the accumulation of the activator in the first place.

Positive autoregulation can be useful in biological processes that unfold slowly, such as development, which can benefit from the slow accumulation of proteins involved in morphogenesis. For example, in the ancient (or primordial) developmental process of sporulation in the bacterium *Bacillus subtilis* (to which we shall return later), the principal regulatory proteins that govern late events in spore formation (the alternative RNA polymerase  $\sigma$  factors  $\sigma^G$  and  $\sigma^K$ ) stimulate transcription of their own structural genes as well as the genes for morphogenetic proteins. Thus, the  $\sigma$  factors as well as the products of the genes they control accumulate slowly because the production of  $\sigma^G$  and  $\sigma^K$  depends on their own synthesis.

Positive autoregulation has an additional benefit. It is the basis for an extreme type of regulatory switch known as a “bistable switch,” as we shall explain later.

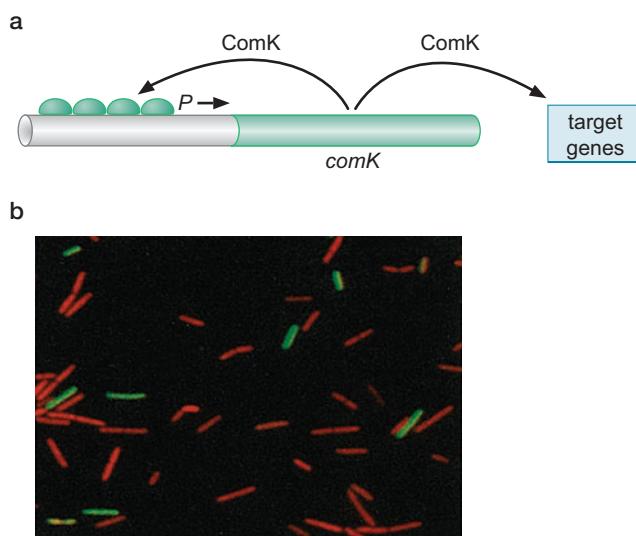
## BISTABILITY

All of the regulatory circuits we have considered so far are reversible in the sense that once the signal that turned a gene or genes ON is removed, the circuit switches back to the OFF state. In some cases, however, when the gene(s) is switched ON, it remains locked ON for relatively long periods of time. This is known as a **bistable switch**.

### Some Regulatory Circuits Persist in Alternative Stable States

A well-studied example of a bistable switch is the circuit that governs whether *B. subtilis* becomes genetically competent. **Competence** is a specialized state in which the bacterium has stopped growing and has acquired the capacity to take up naked DNA from its environment and incorporate homologous sequences into its genome by genetic recombination. The master regulator for competence is the DNA-binding protein ComK, an activator of approximately 100 genes, including its own (Fig. 22-5). What renders the switch stable is cooperativity in the binding of multiple ComK molecules to

**FIGURE 22-5** Bistability. (a) Bistability is governed by a positive autoregulatory switch in which the regulatory protein ComK stimulates transcription from the *comK* gene itself as well as from target genes. Cooperation among ComK molecules bound at the promoter creates a nonlinear response, which is hypersensitive to small, stochastic changes in the level of ComK. The switch is poised at a knife edge between flipping into an ON or an OFF state. (b) A classic example of bistability in which a population of *B. subtilis* cells (red) is either ON (green) or OFF for the expression of a reporter gene under the control of a promoter activated by the competence regulator ComK. The reporter encodes a protein that fluoresces green. (b, Reprinted, with permission, from Dubnau D. and Losick R. 2006. *Mol. Microbiol.* **61**: 564–572. © Blackwell Science.)



the promoter region for *comK*. As we saw in Chapter 18, Box 18-4 in the case of the  $\lambda$  repressor (which is itself responsible for a classic example of bistability to which we shall return later), cooperativity of this sort imparts nonlinearity on the output of the switch as a function of the concentration of the activator. In other words, the output is highly sensitive to changes in the level of ComK (the opposite of robustness).

Whether or not cells have the potential to turn on *comK* is governed by a regulatory pathway operating at the level of the proteolytic stability of the ComK protein. Nevertheless, the ultimate decision to activate *comK* is stochastic. That is, under conditions in which ComK is not subject to degradation, only some of the cells in the population become competent. This can be vividly seen using cells harboring a fluorescent reporter (the gene for the green fluorescent protein) for ComK-directed gene activity. Figure 22-5 shows that cells bifurcate into a subpopulation in which *comK* is ON and a subpopulation in which it is OFF. This is because the positive-feedback loop is poised on a knife edge between having insufficient ComK to switch *comK* ON and just enough (a threshold amount) to trigger the positive autoregulatory loop necessary to turn on ComK-controlled genes (see Box 22-1, Bistability and Hysteresis). Thus, noise in the expression of the *comK* gene resulting in small variations in the levels of ComK between cells enables the activator to reach a threshold concentration in some cells and not others. This example of positive autoregulation illustrates how noise in gene expression can be exploited to drive cells into alternative states.

Positive autoregulation is not the only basis for bistability. A switch that exists stably in two alternative states is also achieved by the use of mutually repressing repressors, that is, two repressors that negatively control each other's transcription. As mentioned above, bacteriophage  $\lambda$  provides a classic example of a bistable switch but one based on a double-negative regulatory circuit rather than positive autoregulation; the mutually antagonistic actions of the CI and Cro repressors together with cooperativity lock in the alternative lysogenic and lytic states of the virus (Chapter 18). Returning to the language of systems biology, we would say that bacteriophage  $\lambda$  has a two-node switch linked in both directions by negative edges.

Although numerous examples of bistable switches are found in bacteria, bistability is by no means limited to microbes. Thus, for example, during embryogenesis, the nematode *Caenorhabditis elegans* generates bilaterally symmetric gustatory neurons called “ASE left” and “ASE right” that express genes for alternative taste receptors. A double-negative-feedback loop that can be stably maintained in one state or another dictates whether a common precursor cell will express one set of receptors or the other. In this case, the switch is not thrown stochastically. Rather, upstream signals dictate in which direction the switch is thrown, whereas the double-negative-feedback loop subsequently locks the switch in its predetermined state.

### Bimodal Switches Vary in Their Persistence

Bistable switches, as we have seen, are bimodal in that they can persist for extended periods of time in alternative stable states. In the case of genetic competence and the phage  $\lambda$  genetic switch, the basis for bistability is self-reinforcing regulatory circuits coupled with cooperative binding of regulatory proteins to DNA. Some regulatory circuits that show bimodality are said to be excitable because they do not persist in alternative stable states. Like bistable systems, excitable systems involve a self-reinforcing circuit that causes a large and stereotypical response to a small perturbation. In

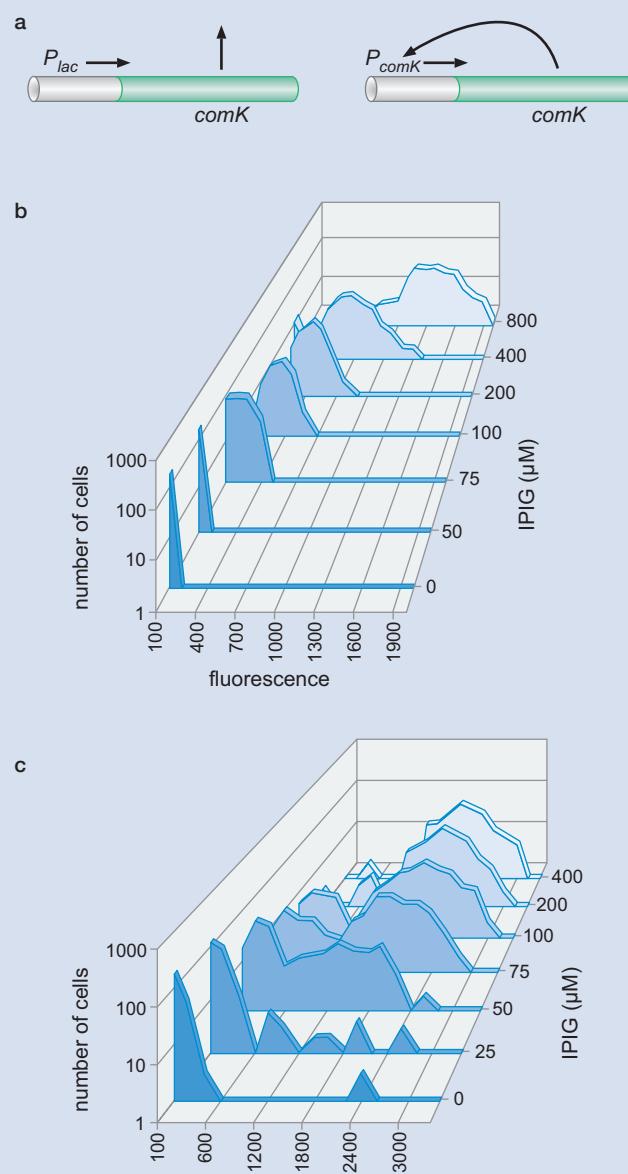
## ► KEY EXPERIMENTS

## Box 22-1 Bistability and Hysteresis

An experiment showing that positive autoregulation is the basis for the bistability of the *comK* switch is based on the use of a modified copy of *comK* that has been brought under the control of a promoter whose activity can be modulated up or down in response to an inducer (Box 22-1 Fig. 1a). In cells harboring the modified gene alone, no bistability is observed, and the level of ComK-directed gene expression increases in a more or less uniform manner in response to increasing levels of inducer, showing a unimodal distribution of expression levels among cells in the population at any give concentration of inducer (Box 22-1 Fig. 1b). However, in cells harboring both the modified gene and the normal autoregulated gene, increasing concentrations of inducer cause the cells to bifurcate into a subpopulation showing a low level of ComK activity and a subpopulation showing a high level of ComK activity (Box 22-1 Fig. 1c). In other words, production of ComK from the modified gene “primes the pump” for the autoregulated gene, causing the switch to be thrown ON in more and more cells as the level of ComK is increased.

Strictly speaking, the use of the term “bistability” requires that a switch show a property called **hysteresis**. Hysteresis is a kind of memory that implies that a switch that has been thrown ON under a particular set of conditions does not immediately switch OFF when those conditions are removed or reversed. Consider, for example, the hysteretic properties of ferromagnetic material. When exposed to a magnetic field, the material becomes magnetized and, importantly, remains so even when the external magnetic field is removed. Now let us return to our example of cells harboring both ComK and a modified copy of ComK that responds to an inducer. As we saw, adding more and more inducer causes the level of ComK to rise until it exceeds the threshold, causing the positive auto-regulatory switch to be thrown ON. Now consider what happens when we lower the level of inducer such that less and less ComK is produced from the engineered copy of the gene. We observe that as the level of inducer is lowered, ComK remains ON even at concentrations of inducer that were insufficient to throw the switch when inducer was increasing. In other words, ComK remembers that it is in the ON state even when the original conditions that switched it ON are reversed.

The switch governing the decision between the lysogenic and lytic modes of propagation of bacteriophage  $\lambda$  is also hysteretic. When the prophage is induced in response to a brief exposure of lysogenic cells to a DNA-damaging agent, the phage irreversibly enters the lytic mode of growth. That is, the phage does not re-enter the lysogenic state (resume synthesizing CI repressor) even after the inducing signal (the DNA-damaging agent) is removed. As a counterexample, when the lactose operon is switched ON by the presence of lactose, the operon returns to its OFF state when the inducer is removed from the medium.



**BOX 22-1 FIGURE 1** Bistability of the *comK* switch. (a) The experiment shows that positive autoregulation causes bistability. The panel shows a modified *comK* gene in which the *comK* promoter is replaced by a promoter that responds to the inducer IPTG (that of the *lac* operon). (b) A graded response occurs when *comK* is under the control of an IPTG-inducible promoter in cells harboring only the modified *comK* gene. (c) A bimodal distribution is seen when the positive autoregulation is left intact and the system is primed with a lactose-inducible copy of *comK*. Note that the cells in (b) and (c) harbored a fusion of the gene for the green fluorescent protein to a promoter under the control of ComK. (b,c, Reprinted, with permission, from Maamar H. and Dubnau D. 2005. *Mol. Microbiol.* **56**: 615–624, Fig. 4E,J. © Blackwell Science.)

excitable systems, however, the switch to an alternative state is transient and readily reverses.

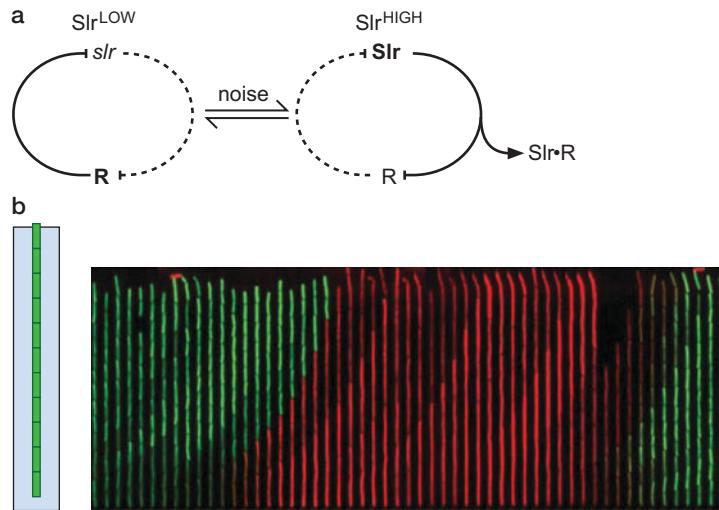
A classic example of an excitable system in biology is the action potential of a neuron. Neurons show a resting potential (typically  $-70$  mV) in which the concentration of cations is slightly higher outside the cell than inside, resulting in a net negative charge in the cytoplasm. Should the resting potential rise above a threshold ( $-55$  mV), protein channels in the membrane known as voltage-gated ion channels open, allowing sodium ions to flow into the cell. This inward flow of positive ions causes the membrane potential to rise still higher, triggering additional channels that had not yet opened to allow sodium ions into the cell. Eventually, this cascade of channel openings culminates in a peak of positive voltage ( $+40$  mV) inside the cell. High voltage, in turn, causes the sodium channels to close. Excess sodium ions are then pumped out of the neuron, restoring the membrane to its original resting state. Thus, a small perturbation in membrane potential triggers a large, programmed response but one that sets in motion its own rapid reversal to the original resting state.

Likewise, self-reinforcing regulatory circuits that are unable to sustain alternative states for extended periods of time or that set in motion a chain of events that causes the circuit to reverse can be considered to be excitable. Seen in this light, the genetic competence system discussed above can be considered to be excitable. Positive autoregulation by ComK creates a bistable switch that could maintain ComK at high levels for an extended period of time. However, superimposed on self-reinforcing synthesis of ComK is a negative-feedback loop that eventually leads to the proteolytic destruction of the activator protein. This negative-feedback circuit enables competent cells to exit their non-growing, competence state and return to a proliferative, vegetative state. The phage  $\lambda$  genetic switch, in contrast, can maintain the lysogenic state for many generations and hence is most properly considered to be bistable.

Competence is a low-probability event. Under competence-inducing conditions, only a small proportion of cells enter the non-growing state in which they can take up DNA. A much more robust example of excitability is shown by the very same bacterium when it is actively growing. Under conditions of steady-state growth, cells of *B. subtilis* exist in alternative states: individual swimming cells and chains of sessile cells. Importantly, the two cell types switch back and forth stochastically at a frequency of tens of cell generations. Why does *B. subtilis* produce a mixed cell population of two very different cell types? One hypothesis is that the bacterium has evolved to hedge its bets, not knowing how long current favorable conditions might last. The sessile chains can be thought of as settlers that stick to surfaces and exploit a currently favorable microenvironment, whereas the swimmers are foragers that swim off in search of new, favorable environments.

How do cells switch between the swimming and chaining states? At the heart of the switch are two regulatory proteins called SinR and SlrR (Fig. 22-6a). Like the CI/Cro switch of phage  $\lambda$ , which we considered in Chapter 18, the SinR and SlrR proteins are part of a double-negative loop. Unlike the phage  $\lambda$  loop, however, in which both regulatory proteins are repressors of each other's genes, SinR is a repressor of the gene for SlrR, but SlrR is an inhibitor of SinR that traps the repressor in a complex (the SlrR–SinR complex) in which it is unable to repress the gene for SlrR. Thus, one arm of the double-negative loop operates at the level of gene transcription and the other arm at the level of protein–protein interaction. The switch exists in two self-reinforcing states: an  $\text{SlrR}^{\text{LOW}}$  state in which the gene for SlrR is repressed by SinR and an  $\text{SlrR}^{\text{HIGH}}$  state in which the gene for SlrR is derepressed and SinR is held inactive by SlrR. Cells in the  $\text{SlrR}^{\text{LOW}}$  state express genes for

**FIGURE 22-6** An excitable circuit that governs switching between alternative cell states. (a) A cartoon of the double-negative loop governing motility and chaining. For simplicity, SlrR and SinR are abbreviated to Slr and R, respectively, in the cartoon. In the SlrR<sup>LOW</sup> state, SinR represses the gene for SlrR (left arm of loop), keeping SlrR levels low (shadowing). In the SlrR<sup>HIGH</sup> state, SlrR, which is at high levels (bold), sequesters SinR in a complex, derepressing the gene for SlrR. (b) A microfluidic channel that is closed at the bottom and open at the top as indicated by the cartoon at the left. As cells grow and divide, cells exit the channel at the top. The kymograph on the right shows a time-lapse series of micrographs taken at 5-min intervals. A motile cell (green) at the bottom left switches to the chaining state (red), giving rise to progeny that maintain the chaining state. Eventually, a chaining cell near the bottom switches back to the motility state (green), giving rise to progeny that maintain the motility state. (Panel b was kindly provided by T. Norman.)



motility and cell separation and hence are swimmers. Cells in the SlrR<sup>HIGH</sup> state, on the other hand, are repressed for motility and cell separation genes and hence grow as sessile chains. Cells in the SlrR<sup>HIGH</sup> state also express genes for an extracellular matrix that makes the chains sticky.

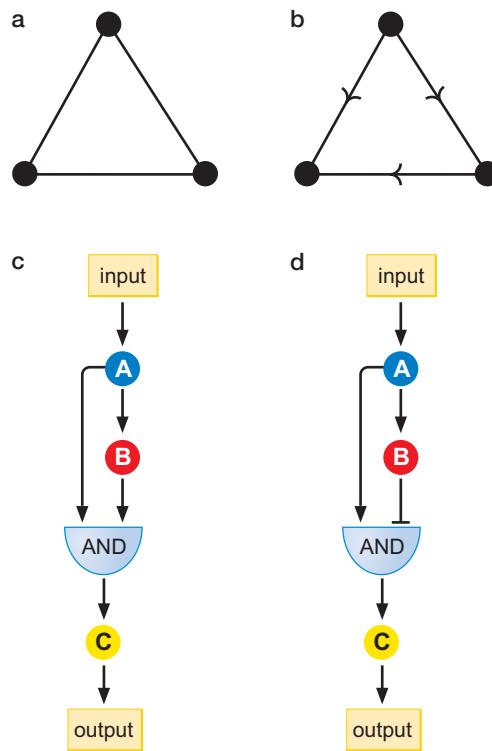
This switching can be visualized in real time using a microfluidic device in which cells are embedded in long channels, each the width of a bacterium. The cells carry a fluorescent reporter for a motility gene (green) and for a matrix gene (red) that is characteristic of the chaining state. The kymograph of Figure 22-6b shows a series of time-lapse micrographs of a single channel taken at 5-min intervals. A motile (green) cell at the bottom of the channel gave rise to progeny that expressed chaining genes (red). Subsequently, a cell in the chaining state near the bottom switches back to a cell expressing motility genes (green). Data collection from large numbers of time-lapse experiments show that cells persist for multiple generations in each state and switch from one to the other stochastically.

## FEED-FORWARD LOOPS

An important contribution from the field of systems biology is the finding that among the myriad kinds of simple regulatory circuits that are theoretically possible, only a small number are commonly found in nature. Evidently, certain circuits have beneficial properties that are favored by natural selection.

### Feed-Forward Loops Are Three-Node Networks with Beneficial Properties

A striking example of this is provided by networks that consist of three nodes (Fig. 22-7a). There are 13 possible ways of connecting three nodes with edges. These can be distinguished from each other by the direction of the edges, whether edges connect two or all three nodes, and whether pairs of nodes are connected by one or two edges. Remarkably, one of the 13 patterns, known as the **feed-forward loop** (Fig. 22-7b), is greatly overrepresented in nature. We refer to it as a “network motif” because it is a recurring theme in genetic circuits. The feed-forward network motif consists



**FIGURE 22-7** Types of networks. (a) A “three-node” network in which each node is a gene and the genes are joined to each other by edges. (b) The feed-forward loop, the most common three-node network found in nature. (c) A “coherent” form of the feed-forward loop in which both the direct and indirect edges leading to the target gene have a positive sign. (d) An “incoherent” form of the loop in which the direct edge has a positive sign and the indirect edge has a negative sign.

of a transcription factor A that controls the gene for a second transcription factor B (Fig. 22-7b). Both transcription factors, in turn, control the third gene in the motif C. Note that Figure 22-7b simply conveys the *direction* of regulation (e.g., node A controls node B), not the sign.

If signs (positive vs. negative control) are attributed to the directional edges, then eight kinds of feed-forward loops can be distinguished. Again, natural selection has favored two that are found more commonly than the others. In one of the favored feed-forward loop motifs (known as a “coherent motif”), both the direct and the indirect pathways leading to the target gene, representing the output, have the same sign (i.e., both A and B are activators) (Fig. 22-7c). In the other favored motif (known as an incoherent motif), the two pathways have different signs, with the target gene C being subject to positive control by A in the direct pathway and negative control by B in the indirect pathway (Fig. 22-7d). In both cases, expression of the target gene is subject to the logic of an AND gate; that is, transcription of C requires both A “AND” B in the former and A “AND NOT” B in the latter.

Because both motifs are favored among all other feed-forward loops and indeed among all possible three-node networks, it is reasonable to expect that they have favorable properties that have been the basis for their selection in evolution. Indeed, computational modeling and experiment reveal that each motif has characteristics that make them useful in regulatory circuits. For example, the coherent feed-forward loop has the property of requiring a sustained input in order for the target gene C to be transcribed (Fig. 22-7c). In other words, this kind of feed-forward loop is a persistence detector that only responds to a signal that is long-lived or persistent. This property derives from the fact that turning on the target gene depends on both the primary activator A and sufficient accumulation of the secondary activator B. Thus, the input signal must persist long enough for the secondary activator B to reach the threshold concentration needed to turn on the target gene C. In other words, by imposing a delay in the response to an input, the coherent feed-

forward loop helps the cell distinguish a true, sustained signal from a stochastic fluctuation (noise) in signal intensity.

The incoherent feed-forward motif has its own beneficial property (Fig. 22-7d). It is a pulse generator that causes gene expression to switch ON and then OFF. Thus, activator A turns on target gene C, but over time the accumulation of repressor B causes the target gene to turn OFF. Thus, the incoherent feed-forward loop is useful when gene expression is required for only a brief period of time.

### Feed-Forward Loops Are Used in Development

These insights reveal simplifying design principles in otherwise complex pathways of gene control. In some cases, a combination of coherent and incoherent feed-forward loops is used to produce elaborate patterns of gene activity. A dramatic example comes from the process of sporulation referred to above whose regulatory circuit is a linked series of coherent and incoherent feed-forward loops (Fig. 22-8). The coherent loops ensure that the input to the circuit is persistent and hence that development is not triggered at the wrong time or at the wrong place. Likewise, the incoherent loops are used to generate successive pulses of gene expression over the course of morphogenesis.

Yet another example is seen in the mechanisms that govern dorsoventral patterning in the *Drosophila* embryo. As discussed in Chapter 21, this process is initiated by the maternal regulatory protein Dorsal, which becomes distributed in a broad gradient. A direct target of Dorsal is the *twist* gene, which is activated at intermediate-high to high levels of the regulatory protein. Twist too is a regulatory protein, and it works in concert with Dorsal to activate a variety of target genes, such as *snail*. This regulatory motif is thus a clear example of a coherent feed-forward loop. In addition, however, *snail* encodes a transcriptional repressor, and many target genes of Dorsal and Twist are also repressed by Snail. Such target genes are thereby regulated by an incoherent feed-forward loop. Thus, the network of *dorsal*, *twist*, *snail*, and downstream genes consists, as in the case of bacterial sporulation, of linked coherent and incoherent feed-forward loops. In the case of *Drosophila* embryogenesis, the feed-forward loops are used to govern dorsoventral patterning. Thus, in the mesoderm, where the levels of Dorsal (and Twist) and hence Snail are high, targets of Snail-mediated repression are OFF, whereas in the neurogenic ectoderm, where the levels of Dorsal and hence Snail are low, these genes are ON.

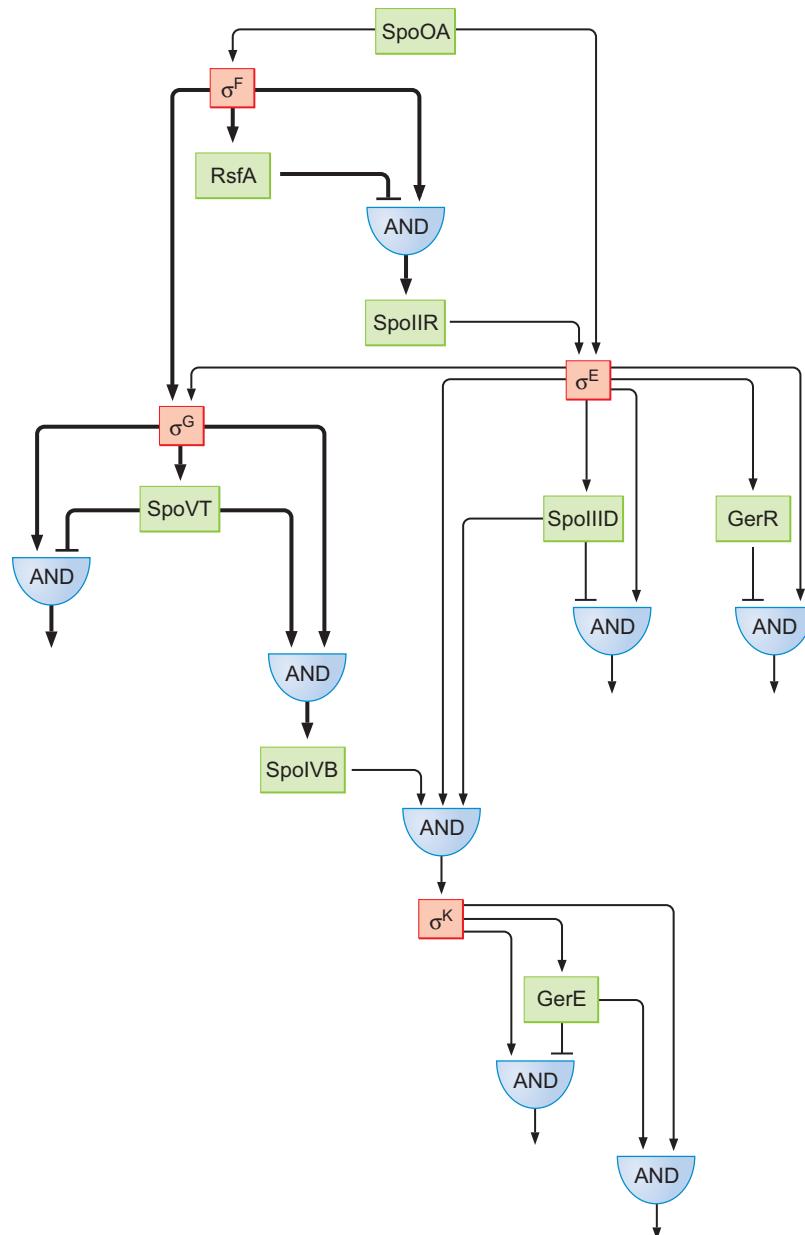
## OSCILLATING CIRCUITS

---

We generally think of regulation in terms of switching genes ON or OFF or adjusting their levels of expression. However, another kind of gene control of wide importance in biology is oscillation in which the expression of large numbers of genes is periodically UP-regulated and then DOWN-regulated at regular intervals over time. Elucidating the circuitry that governs this oscillatory behavior, and doing so in a quantitative manner, is one of the premier challenges of systems biology.

### Some Circuits Generate Oscillating Patterns of Gene Expression

A relatively simple example of an oscillating regulatory circuit is the cell cycle of the bacterium *Caulobacter crescentus* (Fig. 22-9). Here, the master

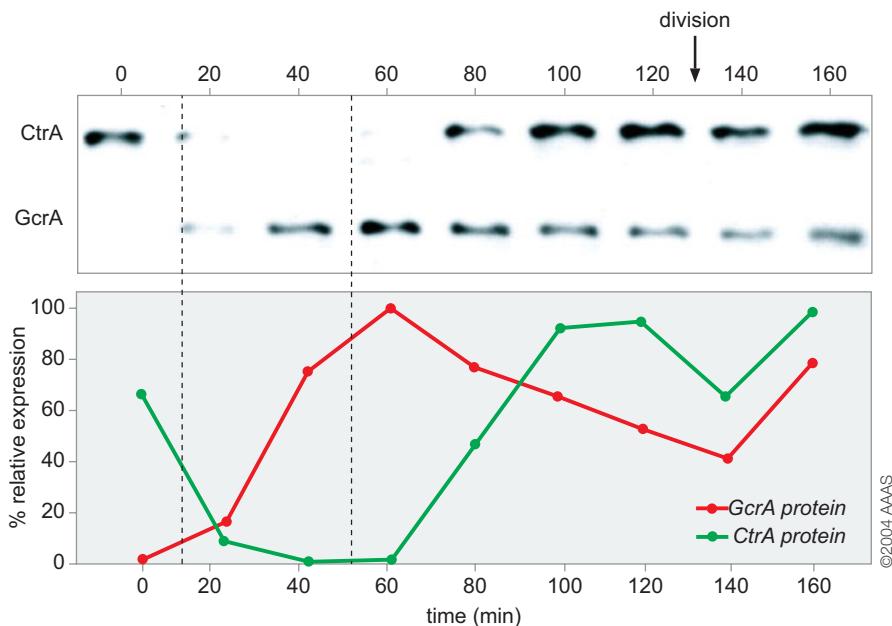


**FIGURE 22-8** The circuitry governing spore formation is a linked series of feed-forward loops. The names refer to regulatory or signaling proteins. Not shown for simplicity is that the  $\sigma^G$  and  $\sigma^K$  factors are subject to positive autoregulation. (Redrawn, with permission, from Wang S.T. et al. 2006. *J. Mol. Biol.* **358**: 16–37, Fig. 5. © Elsevier.)

regulators CtrA and GcrA rise and fall in abundance out of phase with each other in a periodic manner. Their alternating presence drives gene expression in an oscillatory pattern over the course of the cell cycle.

A well-known example of oscillatory behavior is the clock that drives the periodic expression of large numbers of genes at different times during the cycle of day and night. In flies and mammals, this circadian rhythm is governed in part by a negative-feedback loop involving the activator proteins Clock and Cycle and the autorepressor Per (Period). The Clock and Cycle proteins bind to the regulatory region for, and stimulate the transcription of, the *per* gene. When the Per protein accumulates to a critical level, it is able to counteract the action of Clock and Cycle and shut off its own synthesis. Once *per* is switched OFF in this manner, Per protein, which is proteolytically unstable, is depleted from the cell. This leads to a subthreshold level of the autorepressor, which is insufficient to block activation by Clock and Cycle. The *per* gene is thereby turned back ON. This ON/OFF cycle of

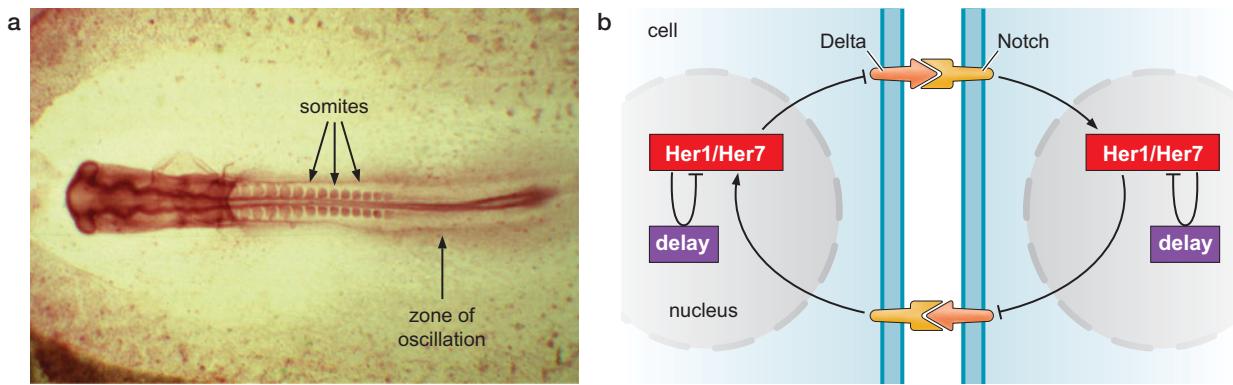
**FIGURE 22-9** The regulators CtrA and GcrA rise and fall in abundance out of phase with each other during the *Caulobacter* cell cycle. (Adapted, with permission, from Holtzendorff J. et al. 2004. *Science* 304: 983–987, Fig. 3B,C. © AAAS.)



*per* expression helps define the 24-h cycling of gene activity. It is critically dependent on the timing of Per protein synthesis and degradation. Changes in Per protein stability can change the frequency of oscillations to produce aberrant ON/OFF cycles once every 22 h or 26 h in place of the normal 24-h cycle. Nevertheless, just how the circadian clock maintains its 24-h cycle and does so in a robust manner is not fully understood and undoubtedly involves additional, yet to be elucidated mechanisms.

Interestingly, negative autoregulation also seems to be involved in another, unrelated example of periodic gene expression: the formation of somites in vertebrate embryos. Somites are condensed blocks of mesoderm cells that form the repeating muscle segments and vertebrae of the spinal column (Fig. 22-10a). They form in a head-to-tail manner and—in zebrafish at least—depend on the ON/OFF oscillating activities of the regulatory genes *her1* and *her7*. These genes are expressed cyclically in the future somite cells, up to the time when these cells are ready to differentiate and form a physical somite. As each new batch of cells matures, it halts its oscillation, in such a way that some cells become arrested at the peak of their oscillation cycle and others at the trough, in a regular spatial order that marks out the pattern of the forming somite. Just behind each newly established somite is the next group of future somite cells that go through the same process, involving another ON/OFF cycle of gene activity, thereby producing a new somite.

The oscillating ON/OFF expression of *her1* and *her7* in zebrafish has been subject to mathematical modeling and computer simulations. Her1 and Her7 are autorepressor proteins that are thought to bind to the regulatory regions of the *her1* and *her7* genes, shutting off transcription. But this repressed state of affairs lasts only a little while: once the Her1 and Her7 repressor proteins diminish below a critical threshold because of their depletion by proteolysis, the block to transcription is relieved, and a new cycle of protein synthesis begins, restoring the repressed state, and so on, in repeated cycles. The oscillating levels of the repressor gene products regulate expression of other genes so as to define the pattern of each new somite. In zebrafish, a new somite forms about every 30 min. The key feature of the model that explains the timing of somite formation is the delay between when the *her1* and *her7* genes



**FIGURE 22-10 Expression of somite genes in vertebrate development.** (a) Somites in the developing chick embryo are shown here. The arrows identify somites and the zone of oscillatory gene expression in which future somites will be generated. (Image kindly provided by Julian Lewis.) (b) Shown here is the model for generating and synchronizing oscillatory expression of somite genes in zebrafish. Oscillatory gene expression is governed by a negative-feedback loop involving autorepressors Her1 and Her7. Synchronization of the oscillations between cells is achieved by Delta/Notch signaling. Her1/Her7 inhibit production of the Delta ligand. Conversely, Notch signaling stimulates production of the Her proteins. (Adapted from a figure kindly provided by Julian Lewis.)

are switched ON and the accumulation of the autorepressors to a concentration sufficient to shut OFF their own synthesis (Fig. 22-10b).

The Her1/Her7 autoregulatory loop nicely explains how somite genes are expressed in an ON/OFF cycle in individual cells, but how is oscillatory gene expression in one cell kept in synchrony with that of other nearby cells in the prospective somite? Synchronization is achieved by an intercellular pathway of cell–cell signaling. In addition to repressing their own genes, Her1 and Her7 inhibit the expression of a gene coding for a cell-surface protein called Delta. Delta binds to the receptor protein Notch on neighboring cells (see Chapter 21). When activated by the Delta ligand, the Notch signaling system, in turn, stimulates the expression of the genes for Her1 and Her7. When Her protein levels are high in the ON/OFF cycle (and hence *her* gene expression low), production of the Delta ligand is low (Fig. 22-10b). As a consequence, Notch signaling and hence expression of *her* genes in adjacent cells is also low. Conversely, when the levels of Her1 and Her7 are low (and hence *her* gene expression high), Delta levels rise and thereby stimulate *her* gene expression in adjacent cells. In each case, the signal delivered from the neighbor via Notch collaborates with the cell’s internal Her1 and Her7 to keep the cell and its neighbor oscillating in synchrony. Thus, interlocking cycles of negative autoregulation and intercellular signaling generate and coordinate oscillatory behavior among the cells that give rise to the somite.

### Synthetic Circuits Mimic Some of the Features of Natural Regulatory Networks

A complementary approach to understanding the design principles that govern regulatory networks is to construct relatively simple circuits that mimic the features of natural systems, the goal of the field of synthetic biology. A dramatic example of successful circuit design that extends our discussion of oscillation is the “repressilator.” The repressilator is a three-node network that was created in *E. coli* and that consists of three regulatory proteins linked to each other in a circular fashion in which the sign of all

three edges is negative. The repressilator consists of the genes for the bacterial repressors  $\lambda$ CI, LacI, and TetR such that CI represses the gene for LacI, which, in turn, represses the gene for TetR, which, to complete the network, represses the gene for CI. One might have anticipated that such a three-node circuit would result in a low steady-state level of transcription of all three genes. Instead, however, the repressilator shows a striking oscillatory pattern of transcription with a periodicity of  $\sim 2$  h. Presumably, fluctuations in the levels of the three repressors due to noise in the expression of their genes prevent the system from achieving steady state and result instead in an oscillatory pattern of expression. Still, the oscillatory behavior of the repressilator is far less robust than that of the natural systems considered above, which highlights the fact that the synthetic circuit is inadequate in mimicking the more intricate (but not yet fully elucidated) circuitry of natural oscillators.

Several other networks have been created synthetically that show diverse stereotyped patterns of behavior. One example is a library of artificial circuits created from multiple transcription factors and multiple promoters in a variety of combinations. Members of this circuit library respond differentially to different combinations of input signals. Another example comes from the construction of “sender” and “responder” strains that create banded patterns of gene expression on agar plates. The sender strain is in the center of the plate and produces a signaling molecule that diffuses out from the center to create a gradient. Each of two responder strains, which are present throughout the plate, responds differentially to high and low concentrations of the signaling molecule by producing distinguishable, chromogenic reporter proteins. As a result, one responder strain produces coloring in a halo pattern that is close to the sender cells, and the other produces a halo that is further away from the sender cells.

## SUMMARY

---

Systems biology is a newly emerging field that seeks to describe complex levels of biological organization by using a combination of quantitative and high-throughput measurements, modeling, reconstruction, and theory. When applied to regulatory circuits, systems biology attempts to reveal principles of gene control that cannot be understood from the study of individual components in isolation. The complementary field of synthetic biology also seeks to elucidate design principles, but it attempts to do so by the creation of artificial regulatory networks that mimic features of natural circuits.

Transcription networks consist of nodes, which represent genes, and edges, which represent the regulation of one gene by another. In a simple, two-node regulatory motif, one gene controls the expression of another, and this regulation can be either negative or positive. Another simple motif is autoregulation, in which a gene regulates its own expression. Negative autoregulation, in which a gene represses its own expression, has the property of dampening noise, which is the variation in gene expression under seemingly uniform conditions. Positive autoregulation has the property of allowing steady-state expression to be reached slowly. An extreme form of positive autoregulation is the bistable switch in which a gene can be either OFF or ON for long periods of time.

Another common motif in regulatory networks is the feed-forward loop. A feed-forward loop is a three-node motif in which a regulatory gene (gene A) governs both the expression of a target gene and the expression of a second regulatory gene (gene B). This second regulatory gene also controls the expression of the target gene. Thus, in a feed-forward loop, gene A controls the expression of the target gene both directly and indirectly via gene B. The expression of the target is subject to an AND gate in that expression is subject to two conditions: in one case, the presence of both activators and in the other, the presence of the activator and the absence of the repressor.

Some regulatory circuits in nature generate oscillating cycles of gene expression as observed in the cell cycle, development, and circadian rhythms. The design of these circuits is such that the appearance of one regulatory protein leads to its own disappearance and the appearance of a second regulatory protein. The second regulatory protein, in turn, causes its own disappearance and the reappearance of the first regulatory protein, thereby generating a continuing ON/OFF cycle of gene expression. A synthetic network consisting of three repressors linked in tandem in a circular circuit mimics natural oscillators in that it generates a cyclic pattern of gene expression but not with the robustness of natural oscillators.

The methods used in systems biology permit the systematic identification of every component engaged in a complex cellular process. The ability to obtain such information is prompting a paradigmatic shift in the way biologists analyze data. Instead of asking *how* a process works, it is now possible to ask *why* it is organized in a particular fashion. Looking ahead, the insights gained from systems biology in combina-

tion with the increasing sophistication of synthetic biology may some day make it possible to create artificial cells with the minimal circuitry for self-propagation. If so, then the future holds the prospect of artificial cells with tailor-made features, such as the capacity to efficiently metabolize pollutants, recycle waste materials, convert sunlight into fuel, or combat human disease.

## BIBLIOGRAPHY

### Books

Alon U. 2006. *An introduction to systems biology: Design principles of biological circuits*. Chapman & Hall/CRC, Boca Raton, Florida.

### Systems Biology

Alon U. 2007. Network motifs: Theory and experimental approaches. *Nat. Rev. Genet.* **8**: 450–461.

Bintu L., Buchler N.E., Garcia H.G., Gerland U., Hwa T., Kondev J., and Phillips R. 2005. Transcriptional regulation by the numbers: Models. *Curr. Opin. Genet. Dev.* **15**: 116–124.

Bintu L., Buchler N.E., Garcia H.G., Gerland U., Hwa T., Kondev J., Kuhlman T., and Phillips R. 2005. Transcriptional regulation by the numbers: Applications. *Curr. Opin. Genet. Dev.* **15**: 125–135.

Crosson S., McAdams H., and Shapiro L. 2004. A genetic oscillator and the regulation of cell cycle progression in *Caulobacter crescentus*. *Cell Cycle* **3**: 1252–1254.

Dubnau D. and Losick R. 2006. Bistability in bacteria. *Mol. Microbiol.* **61**: 564–572.

Endy D. 2005. Foundations for engineering biology. *Nature* **438**: 449–453.

McAdams H.H., Srinivasan B., and Arkin A.P. 2004. The evolution of genetic regulatory systems in bacteria. *Nat. Rev. Genet.* **5**: 169–178.

McGrath P.T., Viollier P., and McAdams H.H. 2004. Setting the pace: Mechanisms tying *Caulobacter* cell-cycle progression to macroscopic cellular events. *Curr. Opin. Microbiol.* **7**: 192–197.

Raser J.M. and O’Shea E.K. 2005. Noise in gene expression: Origins, consequences, and control. *Science* **309**: 2010–2013.

Prinzak D. and Elowitz M.B. 2005. Reconstruction of genetic circuits. *Nature* **438**: 443–448.

Vilar J.M., Guel C.C., and Leibler S. 2003. Modeling network dynamics: The *lac* operon, a case study. *J. Cell Biol.* **161**: 471–476.

## QUESTIONS

### MasteringBiology®

For instructor-assigned tutorials and problems, go to MasteringBiology.

For answers to even-numbered questions, see Appendix 2: Answers.

**Question 1.** What do nodes and edges represent in regulatory circuits? How are nodes and edges depicted?

**Question 2.** Explain what is meant by the term “AND gate” in terms of regulation.

**Question 3.** Describe the plot for negative autoregulation (level of gene expression [%] over time) shown in Figure 22-3. Explain why negative regulation is selected for in evolution.

**Question 4.** Describe the relationship between noise and stochasticity.

**Question 5.** Consider the experiment in which the expression of two copies of the same gene are measured using the green fluorescent protein reporter for the first copy of the gene and red fluorescent protein reporter for the second copy of the gene in *E. coli* cells. Provide an example of intrinsic versus extrinsic noise observed for this system.

**Question 6.** Explain why a regulatory circuit under negative autoregulation is described as robust.

**Question 7.** In Figure 22-3, what portion of the positive autoregulation curve represents when the output reaches steady state?

Explain how steady state is reached when the gene expression is subject to positive autoregulation.

**Question 8.** What property of ComK binding to the promoter for *comK* allows this regulatory circuit to be a bistable switch?

**Question 9.** What type of regulatory circuit controls *Bacillus subtilis* cells switching between the swimming and chaining states? How long do cells persist in the swimming or chaining states?

**Question 10.** Consider a light-activated gene. In the presence of persistent light, the gene turns on and soon after turns off. Which type of feed-forward loop do you expect regulates this gene? Explain your choice.

**Question 11.** What is a circadian rhythm and how is it regulated?

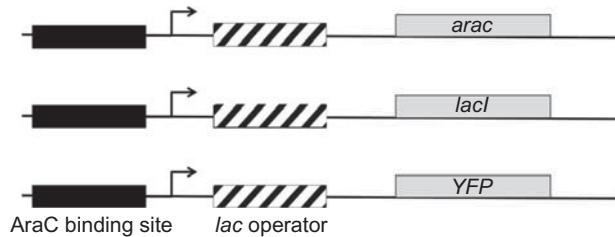
**Question 12.** Using edges and nodes, name and draw a regulatory circuit that represents the expression of a regulatory protein if turned on quickly and maintained at a constant level.

**Question 13.** Using edges and nodes, draw a regulatory circuit that represents the synthetic repressor. Name the genes at the nodes and describe the pattern of expression from the repressor.

**Question 14.** The *ara* operon controls the expression of several genes including *araC*. In the presence of the sugar arabinose

and the DNA-binding protein AraC, gene expression is turned on. The AraC protein is involved in positive autoregulation.

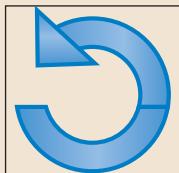
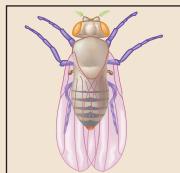
As depicted below, researchers constructed a circuit consisting of a series of artificial promoters (each containing an AraC binding site and a *lac* operator). Remember that the presence of arabinose and AraC promotes transcription of downstream genes, *lacI* encodes the LacI repressor, and binding of LacI to the *lac* operator turns off transcription (even in the presence of arabinose and AraC).



- A. If arabinose is added to a culture of bacteria containing all three constructs, an oscillatory pattern of YFP expression is observed. Assume that a small amount of AraC is present. Briefly explain how YFP expression is turned on following addition of arabinose.
- B. After the initial increase in YFP production following addition of arabinose, how is YFP turned off?
- C. Once YFP expression is off, how is it turned on again?
- D. How do you think the oscillation would be affected if a *small* amount of IPTG (an inducer of the *lac* operon) was added to the medium along with arabinose?
- E. Draw the three-node circuit, labeling the nodes (AraC, LacI and YFP) and including the appropriate edges ( $\longrightarrow$  to represent activation and  $\longrightarrow$  to represent repression).

P A R T      6

# APPENDICES



## O U T L I N E

---

**APPENDIX 1**  
Model Organisms, 797



**APPENDIX 2**  
Answers, 831

## PHOTOS FROM THE COLD SPRING HARBOR LABORATORY ARCHIVES



**Robert Horvitz, 1990 Symposium on The Brain.** Horvitz (left) began working on the worm *Caenorhabditis elegans* as a postdoc in Sydney Brenner's lab at Cambridge in the mid 1970s, before continuing with that model system in his own lab at MIT. The Nobel Prize in Physiology or Medicine in 2002 was awarded for work on the worm, and it was shared by Horvitz (for his work defining genes that controlled programmed cell death) with Brenner himself, who had established the system, and another of his postdocs, John Sulston. In this photo, Horvitz is shown with two members of his lab at the time, Elizabeth Sawin (now a co-director of Climate Interactive) and Asa Abeliovich (a neurobiologist at Columbia).



**Mary Lyon and Rudolf Jaenisch, 1985 Symposium on Molecular Biology of Development.** Lyons and Jaenisch both work with mice. Lyons discovered the phenomenon of mammalian X-chromosome inactivation, the mechanism by which female mammalian cells achieve dosage compensation (Chapter 20). Jaenisch was influential in developing techniques to create transgenic mice and also techniques of therapeutic cloning.



**Michael Ashburner, 1970 Symposium on Transcription of Genetic Material.** Ashburner is a long time champion of the model organism *Drosophila*, with research interests covering many aspects of the structure and function of the *Drosophila* genome. He was part of the consortium that sequenced the fly genome with Craig Venter's company Celera Genomics, about which experience he wrote a short book called *Won for All*. He is here seen kissing the hand of Barbara Hamkalo, at the time a postdoc at Harvard, now a professor of molecular biology at University of California, Irvine, interested in the structure of heterochromatin (Chapter 8).

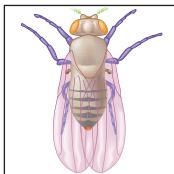


**Dale Kaiser, 1985 Symposium on Molecular Biology of Development.** Kaiser contributed much to the early studies of bacteriophage  $\lambda$  propagation (Chapter 18). One aspect of this work led him to recognize that DNA molecules with complementary single-strand ends can readily be joined together, a finding critical to the development of recombinant DNA technologies.



**Barbara McClintock and Harriet Creighton, 1956 Symposium on Genetic Mechanisms: Structure and Function.** McClintock, the maize geneticist, appeared in an earlier photo, in the Part 3 opener. Creighton worked with McClintock as a student, and together they published an important paper correlating chromosomal crossing over during meiosis with genetic exchange. Creighton spent the last 30 years of her career teaching Botany at Wellesley College, from which she had herself graduated in 1929.

*This page intentionally left blank*



# Model Organisms

A WELL-KNOWN ADAGE IN MOLECULAR BIOLOGY is that fundamental problems are most easily solved in the simplest and most accessible system in which the problem can be addressed. For this reason, over the years molecular biologists have focused their attention on a relatively small number of so-called model organisms. Among the most important of these in order of increasing complexity are *Escherichia coli* and its phage, the T phage and phage  $\lambda$ ; baker's yeast *Saccharomyces cerevisiae*; the mustard-like weed, *Arabidopsis thaliana*; the nematode *Caenorhabditis elegans*; the fruit fly *Drosophila melanogaster*; and the house mouse *Mus musculus*.

What is it that model systems have in common? An important feature of all model systems is that they can be manipulated and studied genetically with the use of the many traditional and new powerful tools of molecular genetics. A second common feature is that the study of each model system attracted a critical mass of investigators. This meant that ideas, methods, tools, and strains could be shared among scientists investigating the same organism, facilitating rapid progress.

For example, beginning in the 1940s a circle of scientists gathered around Max Delbrück, Salvador Luria, and Alfred D. Hershey, spending the summers at the Cold Spring Harbor Laboratories in New York studying the multiplication of the T phage of *E. coli*. These scientists, called the Phage Group, were among those who were important in establishing the field of molecular biology. Many of the members of the Phage Group were physicists attracted to phage, not only because of their relative simplicity, but because the large numbers of phage that could be studied in each experiment generated results that were quantitative and statistically significant. By the late 1950s Cold Spring Harbor offered an annual phage course, where ever-growing numbers of investigators came to learn the new system. This was a case where focusing on the same model organism guaranteed faster progress than would have been made if these individuals had studied many different organisms.

The choice of a model organism depends on what question is being asked. When studying fundamental issues of molecular biology, it is often convenient to study simpler unicellular organisms or viruses. These organisms can be grown rapidly and in large quantities and typically allow genetic and biochemical approaches to be combined. Other questions—for example, those concerning development—can often only be addressed by using more complicated model organisms.

Thus, the T phage (and its best-known member, T4, in particular) proved to be an ideal system for tackling fundamental aspects of the nature of the

## O U T L I N E

- Bacteriophage, 798
    -
  - Bacteria, 802
    -
  - Baker's Yeast, *Saccharomyces cerevisiae*, 808
    -
  - Arabidopsis*, 811
    -
  - The Nematode Worm, *Caenorhabditis elegans*, 816
    -
  - The Fruit Fly, *Drosophila melanogaster*, 819
    -
  - The House Mouse, *Mus musculus*, 825
    -
- Visit Web Content for Structural Tutorials and Interactive Animations

gene and information transfer. Meanwhile, yeast, with its powerful mating system for genetic analysis, became the premier system for elucidating fundamental aspects of the eukaryotic cell. Evolutionary conservation of both proteins and general regulatory mechanisms from fungi to higher cells has meant that discoveries made in yeast frequently hold true for humans. The nematode and the fruit fly also offer well-developed genetic systems for tackling problems that cannot be effectively addressed in lower organisms, such as development and behavior. Finally, the mouse, though less facile to study than nematodes and fruit flies, is a mammal and hence the best model system for gaining insights into human biology and human disease.

In this chapter, we describe some of the most commonly studied experimental organisms and present the principal features and advantages of each as a model system. We also consider the kind of experimental tools that are available for studying each organism and some of the biological problems that have been studied in each case. This chapter is not intended as a comprehensive presentation of all the model organisms that have had an important impact in molecular biology.

## BACTERIOPHAGE

---

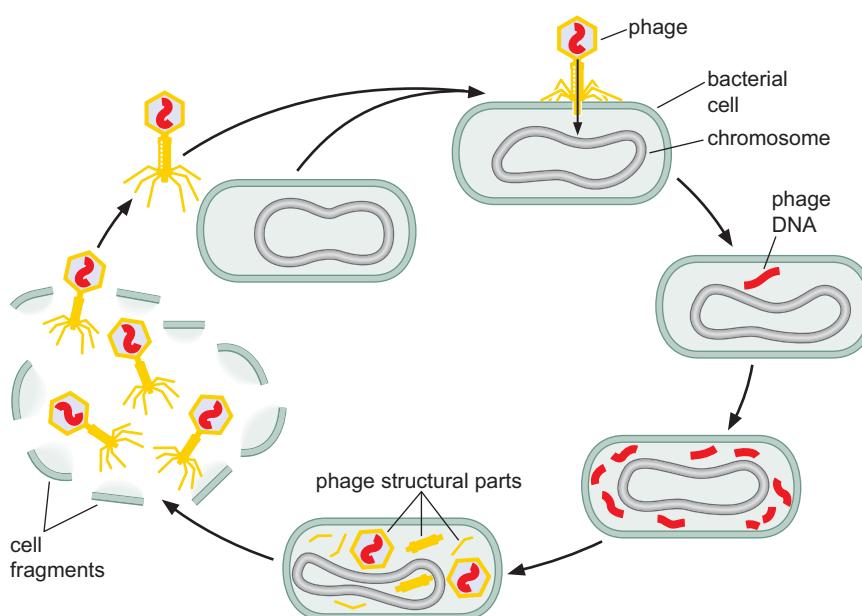
Bacteriophage (and viruses in general) offer the simplest system to examine the basic processes of inheritance. Their genomes, typically small, are replicated—and the genes they encode expressed—only after being injected into a host cell (in the case of phage, a bacterial cell). The genome can also undergo recombination during these infections.

Because of the relative simplicity of the system, phage were used extensively in the early days of molecular biology—indeed, they were vital to the development of that field. Even today they remain a system of choice when studying the basic mechanisms of DNA replication, gene expression, and recombination. In addition, they have been important as vectors in recombinant DNA technology (Chapter 7) and are used in assays for assessing the mutagenic activity of various compounds.

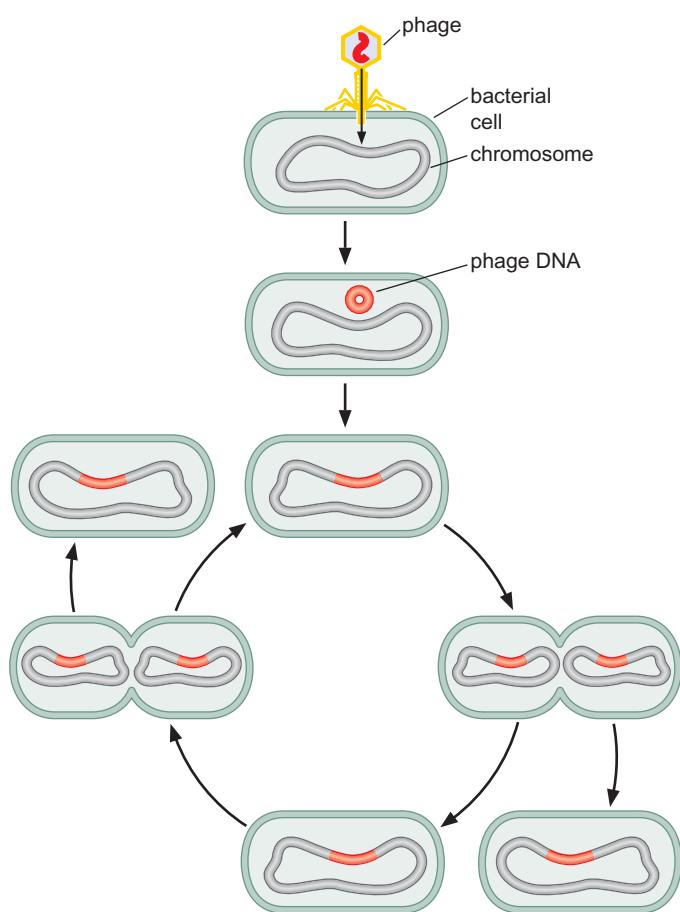
Phage typically consist of a genome (DNA or RNA, most commonly the former) packaged in a coat of protein subunits, some of which form a head structure (in which the genome is stored) and some a tail structure. The tail attaches the phage particle to the outside of a bacterial host cell, allowing the genome of the phage to be passed into that cell. There is specificity here: each phage attaches to a specific cell surface molecule (usually a protein) and so only cells bearing that “receptor” can be infected by a given phage.

Phage come in two basic types—**lytic** and **temperate**, terms that describe their mode for replication. The former, examples of which include the T phage, grow only lytically. That is, as shown in Figure A-1, when the phage infects a bacterial cell, its DNA is replicated to produce multiple copies of its genome (up to several hundred copies) and expresses genes that encode new coat proteins. These events are highly coordinated to ensure new phage particles are constructed before the host cell is lysed to release them. The progeny phage are then free to infect further host cells.

Temperate phage (such as phage  $\lambda$ ) can also replicate lytically. But they can adopt an alternative developmental pathway called **lysogeny** (Fig. A-2). In lysogeny, instead of being replicated, the phage genome is integrated into the bacterial genome, and the coat protein genes are not expressed. In this integrated, repressed state the phage is called a **prophage**. The prophage is replicated passively as part of the bacterial chromosome at cell division, and so both daughter cells are lysogens. The lysogenic state can be



**FIGURE A-1** The lytic growth cycle of a bacteriophage. The phage particle sticks to the outer surface of a suitable bacterial host cell (one bearing the appropriate receptor) and injects its genome, usually a DNA molecule. That DNA is replicated, and the genes expressed to produce many new phage. Once the progeny phage are assembled into mature particles, the bacterial cell is lysed, and the progeny is released to infect another host cell.



**FIGURE A-2** The lysogenic cycle of a bacteriophage. The initial steps of infection are the same as seen in the lytic case (see Fig. A-1). But once the DNA has entered the cell, it is integrated into the bacterial chromosome where it is passively replicated as part of that genome. Also, the genes encoding the coat proteins are kept switched off. The integrated phage is called a prophage. The lysogen can be stably maintained for many generations, but it can also switch to the lytic cycle efficiently under appropriate circumstances. See Chapter 18 for a fuller description of these matters.



**FIGURE A-3** Plaques formed by phage infection of a lawn of bacterial cells. In the case shown, the plaques are produced by a lytic T phage. (Reprinted, with permission, from Stent G.S. 1963. *Molecular biology of bacterial viruses*, Fig. 1. © W.H. Freeman.)

maintained in this way for many generations but is also poised to switch to lytic growth at any time. This switch from the lysogenic to lytic pathway, called **induction**, involves excision of the prophage DNA from the bacterial genome, replication, and the activation of genes needed to make coat proteins and to regulate lytic growth (shown in Fig. 18-20).

### Assays of Phage Growth

For bacteriophage to be useful as an experimental system, methods are needed to propagate and quantify phage. Propagation is needed to generate material (high titer phage stocks for use in experiments or for DNA extraction). Phage are typically propagated by growth on a suitable bacterial host in liquid culture. Thus, for example, a vigorously growing flask of bacterial cells can be infected with phage. After a suitable time, the cells lyse, leaving a clear liquid suspension of phage particles.

To quantify the numbers of phage particles in a solution, a plaque assay is used (Fig. A-3). This is done as follows: phage are mixed with, and adsorb to, bacterial cells into which they inject their DNA. The mix is then diluted, and those dilutions are added to “soft agar,” which contains many more (and uninfected) bacterial cells. These mixtures are poured onto a hard agar base in a petri dish, where the soft agar sets to form a jelly-like top layer in which the bacterial cells are suspended; some are infected, but most are not. The plates are then incubated for several hours to allow bacterial growth and phage infection to take their course.

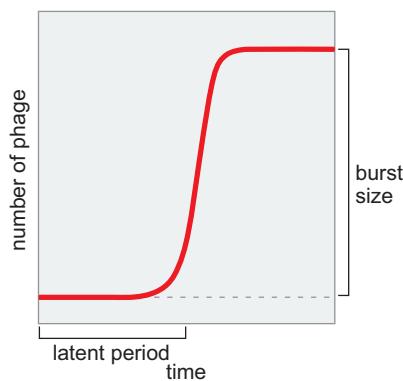
Each infected cell (from the original mix) will lyse during subsequent incubation in the soft agar. The consistency of the agar allows the progeny phage to diffuse, but not far, so they infect only bacterial cells growing in the immediate vicinity. Those cells, in turn, lyse, releasing more progeny, which again infect local cells, and so on. The result of multiple rounds of infection is formation of a **plaque**, a circular clearing in the otherwise opaque lawn of densely grown uninfected bacterial cells. This is because the uninfected bacterial cells grow into a dense population within the soft agar, whereas those bacterial cells located in areas around each initial infection are killed off, leaving a clear patch. Knowing the number of plaques on a given plate, and the extent to which the original stock was diluted before plating, makes it trivial to calculate the number of phage in that original stock.

### The Single-Step Growth Curve

This classic experiment revealed the life cycle of a typical lytic phage and paved the way for many subsequent experiments that examined that life cycle in detail. The essential feature of this procedure is the synchronous infection of a population of bacteria and the elimination of any reinfection by the progeny. This allows the progress of a single round of infection to be followed (Fig. A-4).

Phage were mixed with bacterial cells for 10 minutes. This time period is long enough for bacterial cells to adsorb the phage, but it too short for infection to progress much further. This mixture is then diluted (with fresh growth media) by a factor of 10,000. This dilution ensures that only those cells that bound phage in the initial incubation will contribute to the infected population; also, it ensures that progeny phage produced from those infections will not find host cells to infect.

The diluted population of infected cells is then incubated to allow infection to proceed. At intervals, a sample can be removed from the mixture and



**FIGURE A-4** The single-step growth curve. As described in the text, the single-step growth curve reveals the length of time it takes a phage to undergo one round of lytic growth and also the number of progeny phage produced per infected cell. These are the latent period and burst size, respectively.

the number of free phage counted using a plaque assay. Initially that number is very low (comprising just the phage from the initial infection that did not infect a cell before being diluted).

Once sufficient time has elapsed for infected cells to lyse and release their progeny, a big increase in the number of free phage is detected. (This takes about 30 minutes for the lytic phage T4.) The time lapse between infection and release of progeny is called the **latent period**, and the number of phage released is called the **burst size**.

### Phage Crosses and Complementation Tests

Being able to count the number of phage within a population allows researchers to measure whether a given phage derivative can grow on a given bacterial host cell (and the efficiency with which it does so—e.g., the burst size). Also, the plate assay allows certain types of phage derivatives to be distinguished because of the different plaque morphologies they produce. Differences in host range and plaque morphologies were very often the result of genetic differences between otherwise identical phage. In the early days of molecular biology, this provided genetic markers in a system in which they could be analyzed, enabling researchers to ask how genetic information is encoded and functions.

The ability to perform mixed infections—in which a single cell is infected with two phage particles at once—makes genetic analysis possible in two ways. First, it allows one to perform phage crosses. Thus, if two different mutants of the same phage (and thus harboring homologous chromosomes) coinfect a cell, recombination—and thus genetic exchange—can occur between the genomes. The frequency of this genetic exchange can be used to order genes on the genome. A high recombination frequency indicates that the mutations are relatively far apart, whereas a low frequency indicates that the mutations are located close to each other. The large numbers of phage particles that can be used in such experiments ensure that even very rare events will occur (recombination between two very closely positioned mutations) as long as there is a way to screen for—or better still, select for—the rare event. Second, coinfection also allows one to assign mutations to complementation groups; that is, one can identify when two or more mutations are in the same or in different genes. Thus, if two different mutant phage are used to coinfect the same cell and as a result each provides the function that the other was lacking, the two mutations must be in different genes (complementation groups). If, on the other hand, the two mutants fail to complement each other, then that can be taken as evidence that the two mutations are likely located in the same gene.

### Transduction and Recombinant DNA

Phage crosses and complementation tests allow the genetics of the phage themselves to be analyzed. These same vehicles and techniques can, however, also be used to investigate the genetics of other systems. Initially these observations were restricted to bacterial genes inadvertently picked up during an infection (as we describe later). With the advent of recombinant DNA techniques in the 1970s, however, these studies were extended to DNA from any organism.

During infection, a phage might occasionally (and accidentally) pick up a piece of bacterial DNA. The most common way in which a phage

picks up a section of the host DNA is when a prophage excises from the bacterial chromosome during induction of a lysogen. That process involves a site-specific recombination event (see Chapter 12), and if that event occurs at slightly the wrong position, phage DNA is lost and bacterial DNA included. As long as that exchange does not eliminate part of the phage genome required for propagation, the resulting recombinant phage can still grow and can be used to transfer the bacterial DNA from one bacterial host to another. This process is known as **specialized transduction**. The bacterial DNA included in the specialized transducing phage is amenable to the same kind of genetic analysis as is possible for the phage itself.

Because of its ability to promote specialized transduction, it was natural that phage  $\lambda$  was chosen as one of the original cloning vectors (Chapter 7). Thus, by eliminating many of the sites for a particular restriction enzyme and leaving only one (insertion vector) or two (replacement vector) in a region of the phage not essential for lytic growth,  $\lambda$  can be made to accept the insertion (in vitro) of DNA from any source. That DNA can be propagated and analyzed much more easily than it could in its organism of origin. The restriction endonuclease sites in  $\lambda$  were eliminated by repeatedly selecting phage that plated with higher and higher efficiencies on strains expressing the restriction system in question. By enriching for resistance to endonuclease in this way, and then, in vitro, mapping which sites were lost and which retained, the desired derivative was identified.

Many different  $\lambda$  vectors were developed, all differing in the restriction sites used and in how recombinant phage could be identified. One selection system worked as follows: a  $\lambda$  derivative was derived in which a solitary restriction site was retained within the *cI* gene, the gene that encodes the repressor (see Chapter 18). In the parent vector, therefore, this gene is intact and the phage can, if it chooses, form a lysogen; the phage, therefore, forms turbid plaques. When a piece of DNA is inserted at this site, however, the resulting recombinant phage has a disrupted *cI* gene, cannot form lysogens, and so forms only clear plaques.

This change in plaque morphology provides an easy way of distinguishing recombinant from nonrecombinant phage. Moreover, this approach can be made into a selection (rather than a screen) if the bacterial strain used is an *hfl* strain (see Box 16-5). On that strain, any phage that can form a lysogen invariably does so. Thus, only recombinant phage produce plaques on the *hfl* strain.

## BACTERIA

---

The attraction of bacteria such as *E. coli* or *Bacillus subtilis* as experimental systems is that they are relatively simple cells and can be grown and manipulated with comparative ease. Bacteria are single-celled organisms in which all of the machinery for DNA, RNA, and protein synthesis is contained in the same cellular compartment (bacteria have no nucleus).

Bacteria usually have a single chromosome—typically much smaller than the genome of higher organisms. Also, bacteria have a short generation time (the cell cycle can be as short as 20 minutes) and a genetically homogeneous population of cells (a clone) can easily be generated from a single cell. Finally, bacteria are convenient to study genetically because, on the one hand, they are haploid (which means that the phenotypes of mutations, even recessive mutations, manifest readily), and, on the other hand, because genetic material can be conveniently exchanged between bacteria.

Molecular biology owes its origin to experiments with bacterial and phage model systems. Up until the famous fluctuation analysis experiments of Luria and Delbrück in 1943, the study of bacteria (bacteriology) had remained largely outside the realm of traditional genetics. Taking a statistical approach, Luria and Delbrück demonstrated that bacteria can undergo a change in which they become resistant to infection by a particular phage. Critically, they showed that this change arises spontaneously, rather than as a response (adaptation) to the phage. Thus, like other organisms, bacteria can inherit traits (e.g., sensitivity or resistance to a phage), and occasionally this inheritance can undergo a spontaneous change (mutation) to an alternative inheritable state. The experiments of Luria and Delbrück showed that, like other organisms, bacteria exhibit genetically determined characteristics. But because of their simplicity, bacteria would be ideal experimental systems in which to elucidate the nature of the genetic material and the trait-determining factors (genes) of Gregor Mendel.

### Assays of Bacterial Growth

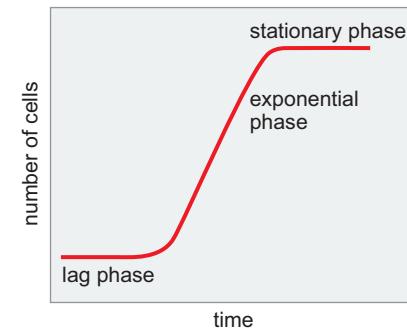
Bacteria can be grown in liquid or on solid (agar) medium. Bacterial cells are large enough ( $\sim 2 \mu\text{m}$  in length) to scatter light, allowing the growth of a bacterial culture to be monitored conveniently in liquid culture by the increase in optical density. Actively growing bacteria that are dividing with a constant generation time increase in numbers exponentially. They are said to be in the **exponential phase of growth**. As the population increases to high numbers of cells, the growth rate slows and bacteria enter the **stationary phase** (Fig. A-5).

The number of bacteria can be determined by diluting the culture and plating the cells on solid (agar) medium in a petri dish. Single cells grow into macroscopic colonies consisting of millions of cells within a relatively brief period of time. Knowing how many colonies are on the plate and how much the culture was diluted makes it possible to calculate the concentration of cells in the original culture.

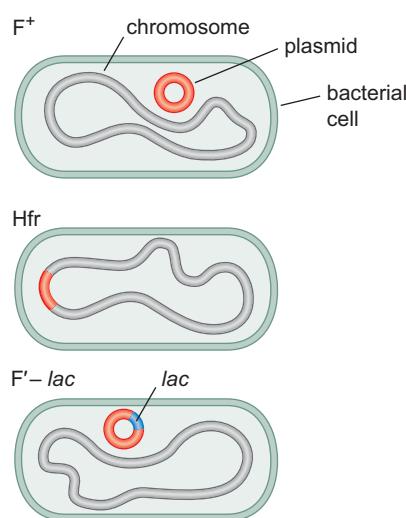
### Bacteria Exchange DNA by Sexual Conjugation, Phage-Mediated Transduction, and DNA-Mediated Transformation

A principal advantage of bacteria as a model system in molecular biology is the availability of facile systems for genetic change. Genetic exchange makes it possible to map mutations, to construct strains with multiple mutations, and to build partially diploid strains for distinguishing recessive from dominant mutations and for carrying out *cis-trans* analyses.

Bacteria often harbor autonomously replicating DNA elements known as **plasmids** (Fig. A-6). Some of these plasmids, such as the fertility plasmid of *E. coli* (known as the **F-factor**) are capable of transferring themselves from one cell to another. Thus, a cell harboring an F-factor (which is said to be  $F^+$ ) can transfer the plasmid to an  $F^-$  cell. F-factor-mediated conjugation is a replicative process. Thus, the  $F^+$  cell transfers a copy of the F-factor, while still retaining a copy, such that the products of conjugation are two  $F^+$  cells. Sometimes the F-factor integrates into the chromosome and as a consequence mobilizes conjugative transfer of the host chromosome to an  $F^-$  cell. A strain harboring such an integrated F-factor is said to be an **Hfr** (for high frequency recombinant) **strain** and is enormously useful for carrying out genetic exchange.



**FIGURE A-5** Bacterial growth curve. As described in the text, bacterial cells, such as *E. coli*, can grow very rapidly when not overcrowded and when propagated in well oxygenated rich medium. This phase of growth is called the exponential phase because the cells are replicating exponentially. Once the number of cells gets too high, and the culture becomes very dense, growth tails off into the so-called stationary phase. Cells taken from the stationary phase and diluted to low density in fresh medium will again enter exponential phase growth, but only after a lag phase. The rate of cell number increases in each of these phases is shown.



**FIGURE A-6** The three forms of F-plasmid-carrying cells.  $F^+$  cells harbor a single copy of the F-plasmid which replicates as an independent minichromosome. In an Hfr strain, the F-plasmid is integrated into the bacterial chromosome and is replicated as part of that larger molecule. In an  $F'$ -strain, an F-plasmid that had previously been integrated into the host chromosome excises, bringing with it a region of adjacent host DNA. All three cell types can be transferred to a recipient  $F^-$  cell. If the donor cell is an  $F^+$  strain, it copies and transfers just the F-plasmid; if an  $F'$ , it copies and transfers the F-plasmid along with the incorporated host DNA; if an Hfr, it copies and transfers varying amounts and parts of the host chromosome, depending on the site of integration and the duration of mating. Once in the recipient, chromosomal DNA from the host is available for recombination, and hence genetic exchange, with the genome of the recipient cell.

Precisely which parts of the host chromosome are transferred during any given example of this exchange varies for two reasons. First, different Hfr strains have the F-plasmid integrated at different locations within the host chromosome. Transfer of the host chromosome into the recipient cell takes place linearly, starting with that region of the chromosome closest to one end of the integrated F-plasmid. Thus, where the plasmid is integrated determines which part of the chromosome is transferred first. Also, it is rare that the entire chromosome gets transferred before mating is broken off. Thus, genes far from the transfer start point are transferred with low frequency, and distant genes may never get transferred in a given mating. Note that a complete copy of the integrated F-factor is transferred last, if at all.

A third and extremely important form of the F-factor is the  $F'$ -plasmid. The  $F'$  is a fertility plasmid that contains a small segment of chromosomal DNA, which is transferred along with the plasmid from cell to cell with high frequency. For example, one such  $F'$  of historic importance is  $F'-lac$ , an F-factor that contains the lactose operon.  $F'$ -factors can be used to create partially diploid strains that have two copies of a particular region of the chromosome. This was precisely how François Jacob and Jacques Monod created partially diploid strains for carrying out their *cis-trans* analyses of mutations in the lactose operon repressor gene and the operator site at which the repressor binds (see Box 18-3).

The F-factor can undergo conjugation only with other *E. coli* strains; however, certain other conjugative plasmids are promiscuous and can transfer DNA to a wide variety of unrelated strains—even to yeast. Such promiscuous conjugative plasmids provide a convenient means for introducing DNA, including DNA that has been modified by recombinant DNA technology, into bacterial strains that are otherwise lacking in their own systems of genetic exchange.

Yet another powerful tool for genetic exchange is phage-mediated transduction (Fig. A-7). **Generalized transduction** is mediated by phage that occasionally package a fragment of chromosomal DNA during maturation of the virus rather than viral DNA. When such a phage particle infects a cell, it introduces the segment of chromosomal DNA from its previous host in place of infectious viral DNA. The injected chromosomal DNA can recombine with the chromosome of the infected host cell, effecting the permanent transfer of genetic information from one cell to another. This kind of transduction is called generalized transduction because any segment of host chromosomal DNA can be transferred from one cell to another. Depending on the size of the virion, some generalized transducing phages transduce only a few kilobases of chromosomal DNA, whereas others transduce well over 100 kb of DNA.

Another kind of phage-mediated transduction is called **specialized transduction**, as already mentioned. This process involves a lysogenic phage such as  $\lambda$  that has incorporated a segment of chromosomal DNA in place of a segment of phage DNA. Such a specialized transducing phage can, upon infection, transfer this bacterial DNA to a new bacterial host cell.

Finally, we come to the case of DNA-mediated transformation, which we described in Chapter 7. Certain experimentally important bacterial species (for example, *B. subtilis* but not *E. coli*) possess a natural system of genetic exchange that enables them to take up and incorporate linear, naked DNA (released or obtained from their siblings) into their own chromosome by recombination. Often the cells must be in a specialized state known as “genetic competence” to take up and incorporate DNA from their environment. Genetic competence is especially useful as it is possible to use recombinant DNA technology to modify a cloned segment of chromosomal

DNA and then have it taken up and incorporated into the chromosomes of competent recipient cells.

### Bacterial Plasmids Can Be Used as Cloning Vectors

As we have seen, bacteria frequently harbor circular DNA elements known as plasmids that can replicate autonomously. Such plasmids can serve as convenient vectors for bacterial DNA as well as foreign DNA. Indeed, the initial (and successful) attempts to clone recombinant DNA involved a plasmid (*pSC101*) of *E. coli* that contains a unique restriction site for *Eco*RI into which DNA could be inserted without impairing the capacity of the plasmid to replicate (Chapter 7).

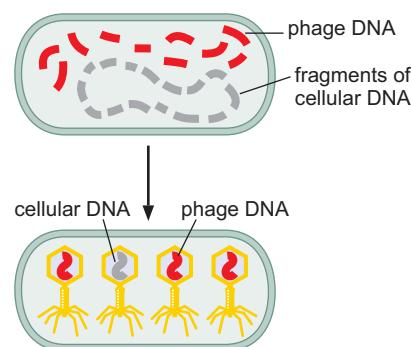
### Transposons Can Be Used to Generate Insertional Mutations and Gene and Operon Fusions

As we discussed in Chapter 12, **transposons** are not only fascinating genetic elements in their own right but are enormously useful tools for carrying out molecular genetic manipulations in bacteria. For example, transposons that integrate into the chromosome with low-sequence specificity (i.e., with a high degree of randomness), such as *Tn5* and *Mu*, can be used to generate a library of insertional mutations on a genome-wide basis (Fig. A-8).

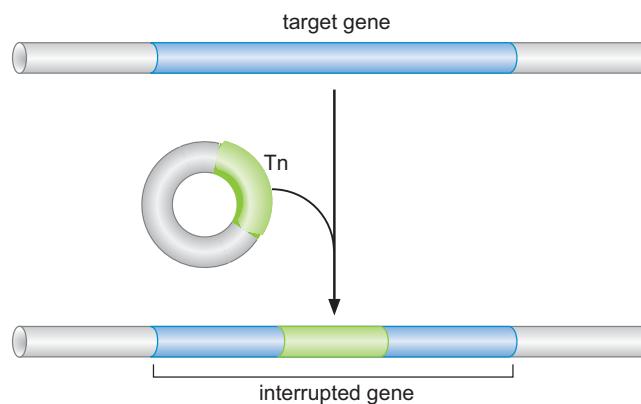
Such mutations have two important advantages over traditional mutations induced by chemical mutagenesis. One advantage is that the insertion of a transposon into a gene is more likely to result in complete inactivation (a null mutation) of the gene (when such is desired) than a simple nucleotide substitution created by a mutagen. The second advantage is that, having inactivated the gene, the presence of the inserted DNA makes it easy to isolate and clone that gene. Even more simply, with the appropriate DNA primers, the identity of the inactivated gene can be determined by DNA sequence analysis from chromosomal DNA harboring the transposon insertion.

Transposons can also be used to create gene and operon fusions on a genome-wide basis. Modified transposons have been created that harbor a reporter gene such as a promoter-less *lacZ* (e.g., *Tn5lac*). When this transposon inserts into the chromosome (in the appropriate orientation), transcription of the reporter is brought under the control of the disrupted target gene. Such a fusion is known as an operon or transcriptional fusion (Fig. A-9).

Other fusion-generating transposons have been created that harbor a reporter gene lacking both a promoter and sequences for the initiation of

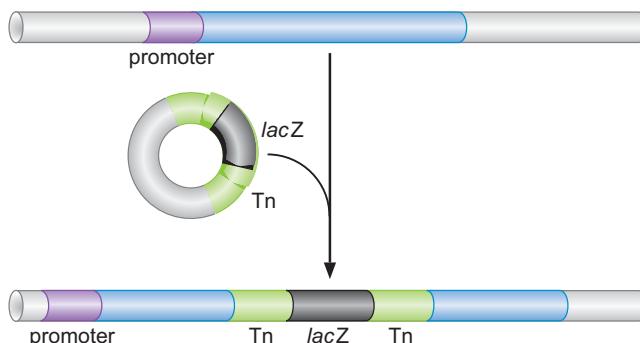


**FIGURE A-7** Phage-mediated generalized transduction. As described in the text, during some phage infections, the host chromosome is fragmented, and segments of that DNA can be packaged in the phage particles instead of the replicated phage DNA. This host DNA is thereby delivered to another cell in the same way as the phage genome ordinarily would. Once in the new host, the DNA can be recombined with the chromosome found there, promoting genetic exchange.



**FIGURE A-8** Transposon-generated insertional mutagenesis. The transposon, carried into a cell on a plasmid, can then transpose from that vehicle into the host genome. Because of the high density of coding regions (genes) on a typical bacterial chromosome, the transposon will very often insert into a gene. A marker carried on the transposon (such as antibiotic resistance) allows cells harboring insertions to be isolated. Knowing the sequence at the ends of the transposon, and of the genome into which it has inserted, makes identifying its location straightforward.

**FIGURE A-9** Transposon-generated *lacZ* fusions. The method of transposon mutagenesis outlined in Fig. A-8 can be modified to allow insertion of a reporter gene (e.g., *lacZ*) into any region of the genome. This allows expression of a host gene (the one in which the transposon–*lacZ* fusion is inserted) to be assessed simply by measuring the level of expression of *lacZ* in that strain.



translation. In these cases, expression of the reporter requires both that it be brought under the transcriptional control of the target gene and that it be introduced into the reading frame of the target gene so that it can be translated properly. A fusion in which the reporter is joined both transcriptionally and translationally to the target gene is known as a gene fusion.

### Studies on the Molecular Biology of Bacteria Have Been Enhanced by Recombinant DNA Technology, Whole-Genome Sequencing, and Transcriptional Profiling

With the advent of recombinant DNA technologies, such as DNA cloning, the availability of whole-genome sequences and methods for studying gene transcription on a genome-wide basis have, of course, revolutionized molecular biological studies of higher cells. But these same technologies have had an impact on the study of bacterial model systems as well, especially when used in conjunction with the traditional tools of bacterial genetics. For example, the development of tailor-made derivatives of transposons for creating gene fusions is facilitated by recombinant DNA methodologies. As another example, the use of genetic competence in combination with recombinant methods for creating precise mutations and gene fusions has expanded the kinds and number of molecular genetic manipulations. The availability of microarrays representing all of the genes in a bacterium has made it possible to study gene expression on a genome-wide basis. In combination with the tools described above, the function of genes identified as being expressed under a particular set of conditions can be rapidly and conveniently elucidated. Methods for rapidly identifying proteins that interact with each other (such as two-hybrid analysis; see Box 19-1), which have had a great impact in yeast and other eukaryotic systems, are also powerful tools for elucidating networks of interactions among bacterial proteins. The availability of whole-genome sequences and promiscuous conjugative plasmids has created opportunities for carrying out molecular genetic manipulations in bacterial species that otherwise lack sophisticated, traditional tools of genetics.

### Biochemical Analysis Is Especially Powerful in Simple Cells with Well-Developed Tools of Traditional and Molecular Genetics

Since the earliest days of molecular biology, bacteria have occupied center stage for biochemical studies of the machinery for DNA replication, information transfer, and gene regulation, among many other topics. There are several reasons for this. First, large quantities of bacterial cells can be grown

in a defined and homogeneous physiological state. Second, the tools of traditional and molecular genetics make it possible to purify protein complexes harboring precisely engineered alterations or to overproduce and thereby obtain individual proteins in large quantities. Third, and of great importance, the machinery for carrying out DNA replication, gene transcription, protein synthesis, and so forth is much simpler (having far fewer components) in bacteria than in higher cells, as we have seen repeatedly in this text. Thus, elucidating fundamental mechanisms proceeds more rapidly in bacteria in which fewer proteins need to be isolated and in which mechanisms are generally more streamlined than in higher cells.

### Bacteria Are Accessible to Cytological Analysis

Despite their apparent simplicity and the absence of membrane-bound cellular compartments (e.g., a nucleus and a mitochondrion), bacteria are not simply bags of enzymes, as had been thought for many decades. Instead, as we now know, proteins and protein complexes have characteristic locations within the cell. Even the chromosome is highly organized inside bacteria. Despite their small size, bacteria are accessible to the tools of cytology, such as immunofluorescence microscopy for localizing proteins in fixed cells with specific antibodies, fluorescence microscopy with the green fluorescent protein for localizing proteins in living cells, and fluorescence *in situ* hybridization (FISH) for localizing chromosomal regions and plasmids within cells. The applications of such methods have provided invaluable insights into several of the molecular processes considered in this text. For example, we now know that the replication machinery of the bacterial cell is relatively stationary and is localized to the cell center (Chapter 9). This finding tells us that the DNA template is threaded through a relatively stationary replication “factory” during its duplication as opposed to the traditional view in which the DNA polymerase traveled along the template like a train on a track. As another example, the application of cytological methods have taught us (again contrary to the traditional view) that during replication the two newly duplicated origin regions of the chromosome migrate toward opposite poles of the cell. Cytological methods are an important part of the arsenal for molecular studies on the bacterial cell.

### Phage and Bacteria Told Us Most of the Fundamental Things about the Gene

Molecular biology owes its origin to experiments with bacterial and phage model systems. Indeed, as we saw in Chapter 2, groundbreaking work with a pneumococcus bacterium led to the discovery that the genetic material is DNA. Since then, experiments with *E. coli* and its phage have led the way, as we have seen throughout this book. For example, the experiment of Hershey and Martha Chase convinced people that the genetic material of phage is DNA; the experiment of Matthew Meselson and Franklin W. Stahl proved that DNA replicates semiconservatively in *E. coli*; the phage crosses of Francis H. Crick and Sydney Brenner (Chapter 16) revealed that the genetic code is built of triplet codons; the elegant genetic studies carried out by Charles Yanofsky in *E. coli* demonstrated genetic colinearity; and the work of Jacob and Monod (see Box 18-2) uncovered the fundamental strategies of gene regulation. There are countless other examples where, by choosing these simplest of systems, fundamental processes of life were understood.

An important example comes from the classic work of Seymour Benzer, who examined intensely a single genetic locus in phage T4, called *rII*. Wild-type T4 is capable of growing in either of two strains of *E. coli* known as B and K, but *rII* mutants grow only in strain B. This makes it possible to detect wild-type phage (arising, e.g., from recombination between two different *rII* mutants) at frequencies of less than 0.01%. That is, a single wild-type phage can be detected among 10,000 *rII* mutant phage when plated on a lawn of strain K bacteria where only the rare recombinant will form a plaque.

Taking advantage of this seemingly arcane property of *rII* mutations, Benzer carried out recombination experiments between pairs of *rII* mutants and was thereby able to map the order of such mutations at a high level of resolution (approaching or reaching that of the nucleotide base pair). He also devised a “complementation” test (discussed above) for showing that the *rII* locus comprises two adjacent genes. Benzer introduced the term **cis-tron** to describe the gene (based on the words *cis* and *trans*). As an aside, it is interesting to note that it was this work that enabled this same locus to be exploited by Crick and Brenner in their genetic studies on the genetic code.

### Synthetic Circuits and Regulatory Noise

In recent years, *E. coli* and other bacterial species have become models for studying new aspects of gene expression. In particular, noise in the regulation of genetic circuits has been identified by looking at expression in many individual cells within genetically identical populations. The variations seen, and the biological consequences of these variations, have allowed people to better understand the advantages as well as the problems of noise in gene networks. These studies are aided by advances in reporter and imaging technology, but again rely on the basic advantages of the bacterial scale and life cycle. (See Chapter 22 for examples from *Bacillus subtilis*.)

Synthetic biological technologies have also largely been pioneered in bacterial systems. Many novel gene regulatory circuits have been constructed and these afford new ways of studying basic characteristics of such networks as well as offering the potential to design new strains with useful functions—such as cells that can digest oil slicks. Recently it was even demonstrated that an entire bacterial genome could be constructed artificially from numerous synthesized fragments and, once constructed, could function well enough to maintain a living cell.

## BAKER'S YEAST, *SACCHAROMYCES CEREVISIAE*

---

Unicellular eukaryotes offer many advantages as experimental model systems. They have relatively small genomes compared to other eukaryotes (see Chapter 8) and a similarly smaller number of genes. Like *E. coli*, they can be grown rapidly in the laboratory (~90 minutes per cell division under ideal conditions), allowing cloned populations to be propagated from a single precursor cell. Despite this simplicity, yeast cells have the central characteristics of all eukaryotic cells. They contain a discrete nucleus with multiple linear chromosomes packaged into chromatin, and their cytoplasm includes a full spectrum of intracellular organelles (e.g., mitochondria) and cytoskeletal structures (such as actin filaments).

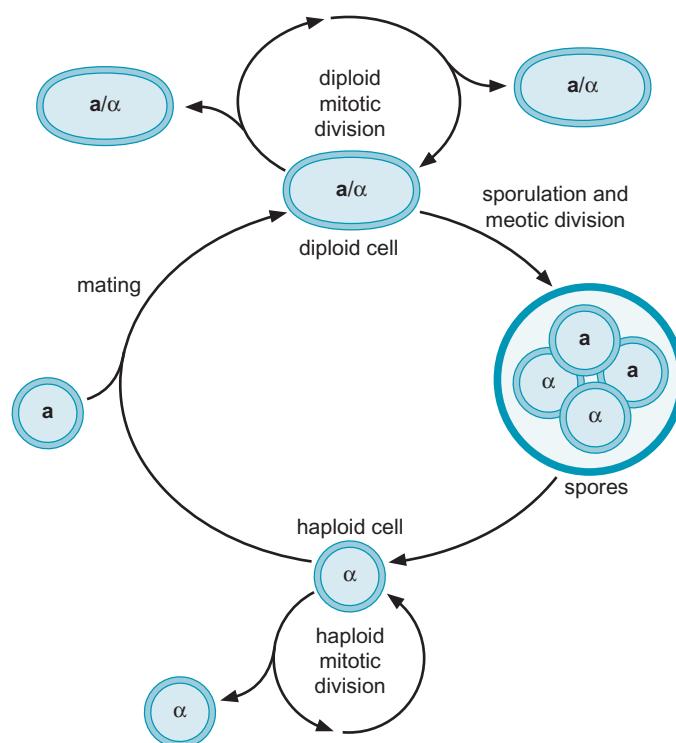
The best studied unicellular eukaryote is the budding yeast *S. cerevisiae*. Often referred to as brewer's or baker's yeast because of its use as a fermenting agent, *S. cerevisiae* has been intensely studied for more than 100 years. In experiments in the 1860s, Louis Pasteur identified this yeast as the catalyst

for fermentation (prior to Pasteur's work, sugar was believed to break down spontaneously into alcohol and carbon dioxide). These studies eventually led to the identification of the first enzymes and the development of biochemistry as an experimental approach. The genetics of *S. cerevisiae* has been studied since the 1930s, resulting in the characterization of many of its genes. Thus, like *E. coli*, *S. cerevisiae* allows investigators to attack fundamental problems of biology using both genetic and biochemical approaches.

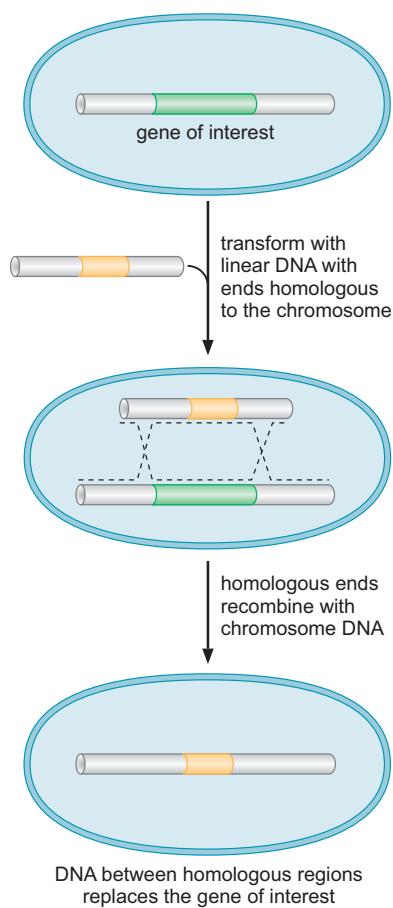
### The Existence of Haploid and Diploid Cells Facilitates Genetic Analysis of *S. cerevisiae*

*S. cerevisiae* cells can grow in either a haploid state (one copy of each chromosome) or diploid state (two copies of each chromosome) (Fig. A-10). Conversion between the haploid and diploid states is mediated by mating (haploid to diploid) and sporulation (diploid to haploid). There are two haploid cell types called **a** and **α** cells. When grown together, these cells mate to form **a/α** diploid cells. Under conditions of reduced nutrients, **a/α** diploids undergo meiotic division (see Chapter 8) to generate a structure known as the ascus that contains four haploid spores (two **a** spores and two **α** spores). When growth conditions improve, these spores can germinate and grow as haploid cells or mate to re-form **a/α** diploids.

In the laboratory, these cell types can be manipulated to perform a variety of genetic assays. Genetic complementation can be performed by simply mating two haploid strains, each of which contains one of the two mutations whose complementation is being tested. If the mutations complement each other, the diploid will be a wild type for the mutant phenotype. To test the function of an individual gene, mutations can be made in haploid cells in which there is only a single copy of that gene. For example, to ask if a given gene is essential for cell growth, the gene can be deleted in a haploid. Only deletions of nonessential genes can be tolerated by haploid cells.



**FIGURE A-10** The life cycle of the budding yeast *S. cerevisiae*. *S. cerevisiae* exists in three forms. Two haploid cell types, **a** and **α**, and the diploid product of mating between these two. Replication of these different cell types, mating and sporulation, are shown.



**FIGURE A-11 Recombinational transformation in yeast.** Any region of the yeast genome can readily be replaced by the sequence of choice. The DNA to be inserted is flanked with DNA sequences homologous to the sequences flanking the region in the chromosome to be replaced. When the donor fragments are introduced to the cell, high levels of homologous recombination in this organism ensure a high frequency of recombination with the chromosome, resulting in the genetic exchange shown. The inserted DNA may differ from the resident sequence by as little as a single base pair, or at the other extreme, it can be very different in length and sequence. Thus, very elaborate genetic modifications can be achieved.

## Generating Precise Mutations in Yeast Is Easy

The genetic analysis of *S. cerevisiae* is further enhanced by the availability of techniques used to precisely and rapidly modify individual genes. When linear DNA with ends homologous to any given region of the genome is introduced into *S. cerevisiae* cells, very high rates of homologous recombination are observed resulting in the replacement of chromosomal sequences with DNA used in the transformation (Fig. A-11). This property can be exploited to make precise changes within the genome. This approach can be used to precisely delete the coding region of an entire gene, change a specific codon in an open reading frame, or even change a specific base pair in a promoter. The ability to make such precise changes in the genome allows very detailed questions concerning the function of particular genes or their regulatory sequences to be pursued with relative ease.

## *S. cerevisiae* Has a Small, Well-Characterized Genome

Because of its rich history of genetic studies and its relatively small genome, *S. cerevisiae* was chosen as the first eukaryotic (nonviral) organism to have its genome entirely sequenced. This landmark was accomplished in 1996. Analysis of the sequence ( $1.3 \times 10^6$  bp) identified approximately 6000 genes and provided the first view of the genetic complexity required to direct the formation of a eukaryotic organism.

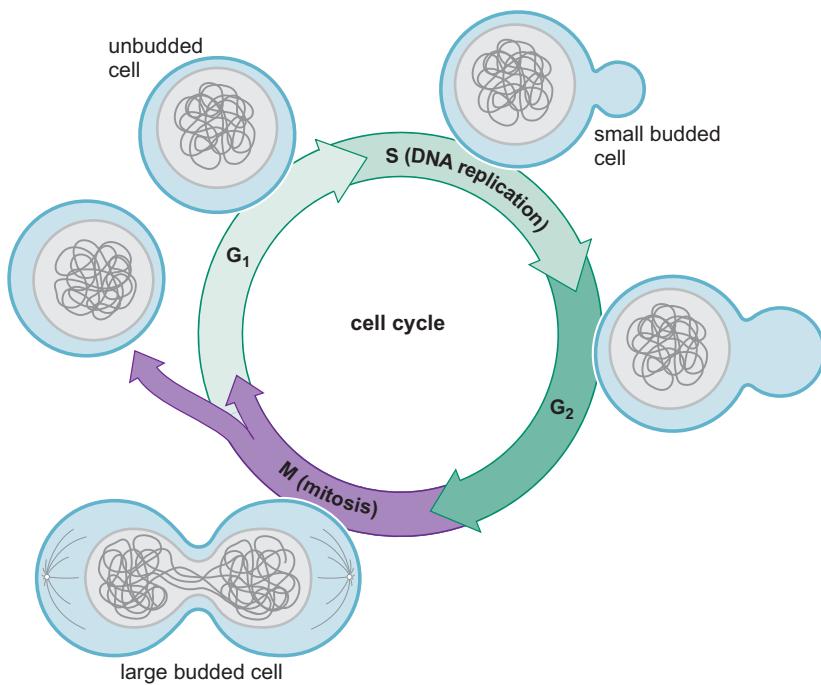
The availability of the complete genome sequence of *S. cerevisiae* has allowed “genome-wide” approaches to studies of this organism. For example, DNA microarrays that include sequences from each of the approximately 6000 *S. cerevisiae* genes have been used extensively to characterize patterns of gene expression under different physiological conditions. Indeed, the levels of gene expression in *S. cerevisiae* cells have now been tested for hundreds of different conditions, including different carbon sources (such as glucose vs. galactose), cell types, and growth temperatures. These findings are not only useful to determine the expression of individual genes but have also led to the grouping of genes into coordinately regulated sets, which all respond similarly to changes in conditions.

Other genome-wide resources include a library of 6000 strains, each deleted for only one gene. Greater than 5000 of these strains are viable as haploids, indicating that the majority of yeast genes are nonessential under the ideal growth conditions in the laboratory. This collection of strains has allowed the development of new genetic screens in which every gene in the *S. cerevisiae* genome can be tested individually for its role in a particular process. The use of microarrays has also allowed the genome-wide mapping of binding sites for transcriptional regulators using chromatin immunoprecipitation techniques (see Chapter 7).

## *S. cerevisiae* Cells Change Shape as They Grow

As *S. cerevisiae* cells progress through the cell cycle, they undergo characteristic changes in shape (Fig. A-12). Immediately after a new cell is released from its mother, the daughter cell appears slightly elliptical in shape. As the cell progresses through the cell cycle, it forms a small “bud” that will eventually become a separate cell. The bud grows until it reaches a size slightly smaller than the “mother” cell from which it arose. At this point the bud is released from the mother and both cells start the process again.

Simple microscopic observation of *S. cerevisiae* cell shape can provide a lot of information about the events occurring inside the cell. A cell that lacks



**FIGURE A-12** The mitotic cell cycle in yeast. *S. cerevisiae* divides by budding. The development of a daughter bud through the mitotic cycle is shown and described in the text.

a bud has yet to start replicating its genome. This is because in a wild-type *S. cerevisiae* cell, the emergence of a new bud is linked to the initiation of DNA replication. Similarly, a growing cell with a very large bud is almost always in the process of executing chromosome segregation.

The powerful genetic, biochemical, and genomic tools available to study *S. cerevisiae* have made it a favored eukaryotic organism for the analysis of basic molecular and cell-biological questions. Studies of *S. cerevisiae* have made fundamental contributions to our understanding of eukaryotic transcription and gene regulation, DNA replication, recombination, translation, and splicing. Genetic studies in baker's yeast have identified proteins involved in all of these events. Perhaps most importantly, the proteins and genes identified as critical to these fundamental events in *S. cerevisiae* are almost always conserved in other eukaryotes including human. Thus, what is learned with this simple model eukaryote is almost always relevant to the same events in the more complex organisms.

## ARABIDOPSIS

Plant science has the longest history of all the life sciences, with its roots in agriculture and botanical medicine: through Mendel, plant science laid the foundations for genetics; and through Charles Darwin, Barbara McClintock, William Bateson, and others, for cytogenetics, development, physiology, and evolution. Plant science continues to make important advances in fundamental areas like RNA interference, while impacting the economy and the environment as it always has. In the last few decades, the humble mustard-like weed, *A. thaliana*, has emerged as a model system on a parallel with *Drosophila*, *C. elegans*, and the mouse. Even more so than its animal counterparts, *Arabidopsis* illustrates most key aspects of plant biology, especially among the angiosperms (flowering seed plants). And just as maize revolutionized plant genetics in the 20th century, *Arabidopsis* promises

to revolutionize plant genomics and most aspects of plant biology into the future.

### *Arabidopsis* Has a Fast Life Cycle with Haploid and Diploid Phases

Like yeast, all plants have both haploid and diploid life cycle phases, which are named according to their products—the diploid phase (like yeast) supports meiosis to generate spores and is therefore named the “sporophyte” (spore-bearing plant). These haploid spores germinate to give rise to the haploid phase, from which gametes of each sex differentiate, and so the haploid phase is known as either the male or female gametophyte. The gametes fuse during fertilization to generate diploid zygotes. The relative length of these phases varies—mosses spend most of their time in the gametophyte phase, whereas *Arabidopsis* and other higher plants spend their time as sporophytes, and ferns lie somewhere in between. In flowering plants, the gametophyte phase is very short, consisting of only two to three mitotic divisions, and the germ line arises from flowers that develop on the adult plant rather than being sequestered in the embryo like animals. Most plants (like *Arabidopsis*) are hermaphrodites and give rise to haploid gametophytes of both sexes from differentiated flowers or floral parts where male and female meiosis occur (see Fig. A-13). But some plants are dioecious (individual sexes) and can even have differentiated sex chromosomes, although such species are rare.

Seed plants, like *Arabidopsis*, go through additional phases after fertilization—embryogenesis and dormancy, giving rise to the eponymous seed. As in mammals, the embryo is nurtured by extraembryonic tissues, which terminally differentiate and go no further. But unlike animals, these extraembryonic tissues (known as the endosperm “inner seed”) are the product

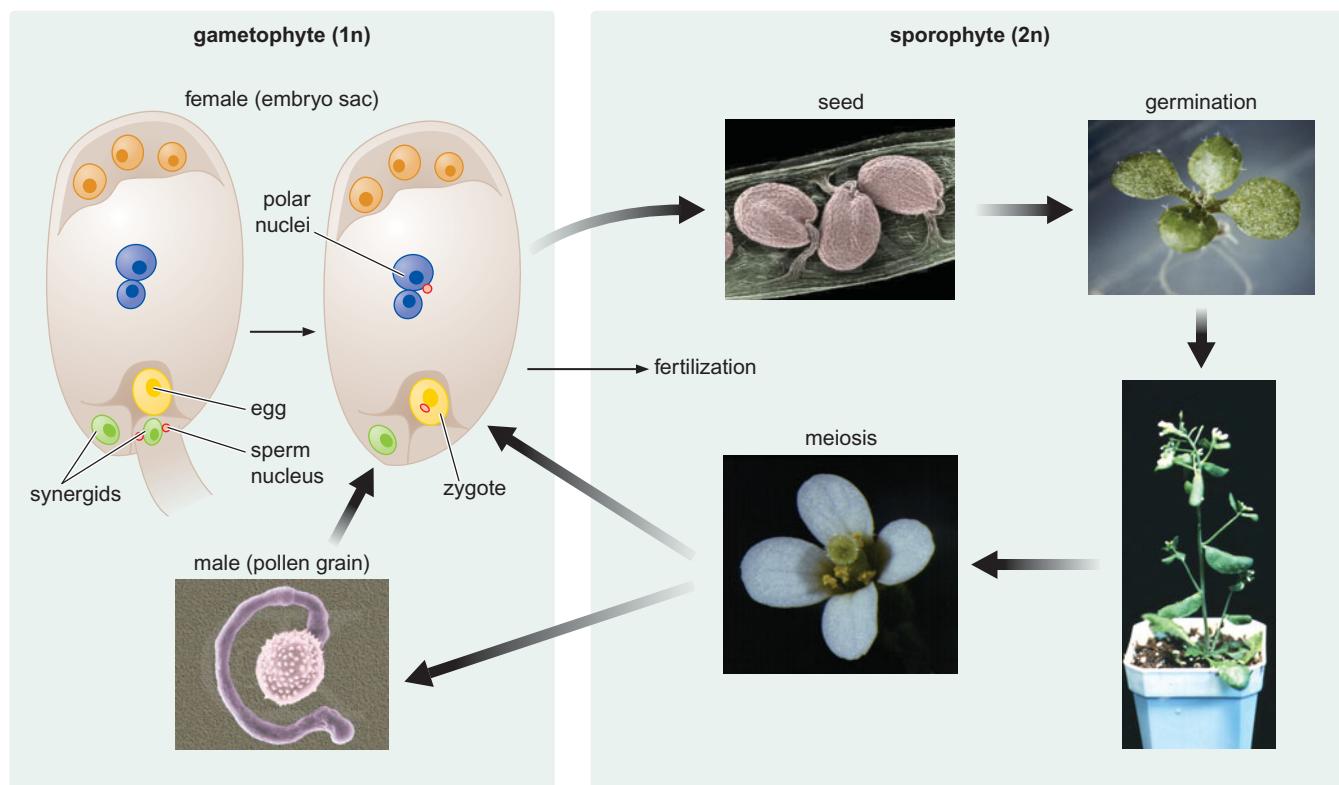


FIGURE A-13 The life cycle phases of *Arabidopsis*. (Courtesy of Rob Martienssen.)

of an independent fertilization, between a second haploid sperm cell and the diploid “central” cell on the female side, which itself is formed by fusion of two haploid sisters of the egg cell. Rapid division of the triploid nucleus is followed by cellularization and starch and protein accumulation, which provides important nutrients to the embryo. The endosperm is ephemeral in many plants (such as *Arabidopsis*), being gobbled up by the embryo as it grows, but can survive until germination in other plants, providing starch in staple crops like wheat and maize.

### *Arabidopsis* Is Easily Transformed for Reverse Genetics

Infection with the soil bacterium, *Agrobacterium tumefaciens*, and its relatives leads to the induction of tumorous growths (galls) because of the transfer of hormone biosynthesis genes from the bacterial Ti (tumor-inducing) plasmid into the chromosomal DNA of the host plant. The tumor-inducing genes are found in the transfer DNA (T-DNA) portion of the plasmid, flanked by directly repeated border sequences required for transfer. By replacing the tumor-inducing genes with genes of interest, it is possible to transform plants. *Arabidopsis* can be transformed by simply spraying the plants with, or dipping them into, a concentrated culture of *Agrobacterium* in a surfactant solution to promote infection. Transient infection occurs almost immediately and is useful for transient expression studies, but stable transformation is thought to occur several days or weeks later, possibly on infection of the female gametophyte, before fertilization. By including a selectable marker gene (for various types of herbicide resistance), it is possible to select transformed plants by germinating seed on media, or soil, containing herbicides.

The efficiency of *Arabidopsis* transformation is so high that it can be used for mutagenesis; random insertion of hundreds of thousands of T-DNAs in individual plants, followed by amplification and sequencing of the insertion sites, has resulted in numerous collections of plants with disruptions in most of the genes in the genome. These insertions can be used for “reverse” genetics in exactly the way that deletions are used in yeast. Further, by including reporter genes, or strong enhancers, on the T-DNA, these insertions can be used to report the expression of the genes in which they integrate or else activate them in cells in which they are not normally expressed. By including transposable elements in the T-DNA, it is possible to generate large numbers of transposon hops without the need for additional transformation and to generate derivative alleles, revertants, and mosaics. Because transformation was much more difficult in rice and maize, transposons have been the major tool for this “reverse genetics” approach in these plants.

### *Arabidopsis* Has a Small Genome That Is Readily Manipulated

The *Arabidopsis* genome includes only 105 Mb of euchromatic DNA, about 15 Mb of sequenced heterochromatin, and an additional 15–25 Mb of satellite repeats and rDNA, making a total of about 140 Mb. Most of the sequenced heterochromatin flanks each of the five centromeres, although smaller regions of heterochromatin (knobs) are found on chromosome arms. Sequencing of the euchromatic portion, and much of the heterochromatin, resulted in the sequence of 99% of the 29,000 *Arabidopsis* genes. Sequencing of many other plant genomes has revealed that several rounds of genome duplication (polyploidy) have occurred in the “eudicots” a major branch of the angiosperm evolutionary tree that includes *Arabidopsis*. The most recent duplication was only a few million years ago, so that about 25% of *Arabidopsis* genes have retained a functional homolog, resulting in sub-

stantial genetic redundancy and complicating reverse genetic strategies. On the other hand, forward genetics has been very powerful in *Arabidopsis*, perhaps in part because of this redundancy, which allows heavy doses of mutagens (such as ethylmethane sulfate [EMS], ethylnitrosourea [ENS], or irradiation) to be used without killing the plants, so that relatively small numbers of mutagenized seed can achieve saturation. Seed can be mutagenized directly and recessive mutations recovered by simply allowing the seed to germinate and self-pollinate.

The availability of the genome sequence and several polymorphic strains has made positional cloning of mutations identified by forward genetics extremely straightforward. EMS mutagenesis can even be used in a reverse genetics strategy, known as **tilling**, in which DNA from mutagenized plants is screened for point mutations in genes of interest. With the emergence of very-high-throughput sequencing methods, this strategy is likely to become even more practical and can recover a full spectrum of allelic variation in each gene. The availability of the genome sequence has enabled a host of other genomic technologies, such as tiling microarrays, high-throughput protein localization, and proteomic technologies, to name but a few.

RNA interference via small RNA (19–30 nucleotides) is an important endogenous and exogenous mechanism for regulating genes and was first discovered in plants (see Chapter 19). In *Arabidopsis* at least three classes of small RNA—microRNA (miRNA); *trans*-acting, short interfering RNA (tasiRNA); and siRNA associated with repeats—differ in size and biogenesis, but all can regulate genes by matching their sequence and promoting “slicing” via endonuclease activity, translational arrest, or chromatin and DNA modification. These small RNAs are derived from single-strand “hairpin” structure precursors or from double-stranded RNA that is the product of RNA-dependent RNA polymerase. Genomic methodology using RNA interference in *Arabidopsis* includes VIGS (virally induced gene silencing), cosuppression, hairpin silencing, and artificial miRNA.

## Epigenetics

Epigenetic variation is generally defined as “mutations” that are chromosomally inherited but do not involve a change in nucleotide sequence. These “epimutations” are usually reversible at a significantly higher frequency than regular mutations and are associated with chemical modification of DNA and associated proteins (especially histones). Plants have been at the forefront of epigenetics research for several decades, and *Arabidopsis* is no exception. Like mammalian genomes, but unlike those of yeast, worms, and flies, plant genomes are heavily modified by cytosine methylation, which, along with histone modification, has epigenetic consequences for expression of both genes and repetitive elements found in the genome. These modifications are guided by a variety of factors, including RNA interference, resulting in the phenomenon of RNA-dependent DNA methylation, first discovered in plants, and RNA-dependent histone modification, which also occurs in fission yeast and other organisms.

When silencing of a given gene differs between male and female germ lines, imprinting results in expression from (usually) the maternally inherited allele. Imprinted expression is prevalent in the extraembryonic endosperm tissue, reminiscent of imprinting in the mammalian placenta. In these well-studied examples, demethylation of imprinted genes occurs in the central cell, resulting in maternal expression in the endosperm.

Epigenetic effects are often influenced by the environment, and in a dramatic example, plants remember the cold of winter by flowering in the

following spring. This memory is induced by cold, retained by clonally propagated cells, but erased by meiosis, resulting in the familiar flowering habit of crops like winter wheat. In *Arabidopsis*, this process (vernalization) is regulated by RNA processing and histone modification, and involves the polycomb complex, also involved in cellular memory in animals.

### Plants Respond to the Environment

Unlike animals, plants are rooted to the spot and cannot flee environmental assault, resulting in properties not usually found in animals—such as grazing tolerance. The innate immune system, first molecularly characterized as the “gene-for-gene” response in plants, includes many components conserved in animals, but it is highly diversified and can recognize viruses, microbes, worms, insects, and even other plants. In addition to this “biotic” stress, plants must withstand and respond to “abiotic” stress, including changes in light intensity, circadian rhythm, nutrient, and salt and water stress, to name but a few. Many of these environmental triggers have profound effects on development—for example, by inducing or delaying flowering to optimize seed production.

Light plays a central role in plant biology, because of the photosynthetic chloroplast, which is derived from an ancient symbiotic prokaryote and responsible for most of the organic carbon fixed in the biosphere. Even in photosynthetic research, *Arabidopsis* is replacing classical physiological models—such as tobacco and spinach—because of the ease of genetic and genomic manipulation.

### Development and Pattern Formation

Plant development has influenced crop domestication and breeding, and therefore human history, more than any other aspect of plant biology, with dramatic innovations affecting inflorescence architecture, seed shattering, and leaf shape, selected by ancient farmers and sophisticated breeders alike. Cauliflower, popcorn, and kale each differ by only a handful of genes from progenitor species that would only be recognized as weeds to modern-day farmers. Because flowering plants are a recent evolutionary group, many of the genes responsible have since been identified using *Arabidopsis* as a model.

More generally, plants and animals diverged from a common but unicellular ancestor, so that multicellular development evolved independently in each kingdom. Therefore we see that essential general principles, such as the central importance of transcription factors and signaling hierarchies (peptides, hormones, and receptors), are recognized and present in each kingdom, whereas specific molecules are only rarely conserved. Some mechanisms, such as cell cycle and MAP (mitogen-activated protein) kinase cascades, are very familiar, but most are distinct. For example, homeotic and heterochronic identities are specified by transcription factors and miRNA in both lineages, but the molecules are not conserved, involving mostly MADS (MCM1, agamous, deficiens, and serum response) transcription factors in plants and *Hox* genes in animals. Intercellular communication involves hormones in both kingdoms, but these have only general similarities (with the exception, perhaps, of plant and animal steroids). Indeed, the highly connected supracellular vascular system of plants allows macromolecules, such as mRNA, small RNA, and transcription factors themselves, to pass directly between cells, whereas this phenomenon has only rarely been observed in animals.

Common developmental mechanisms, then, are likely to have had a function in unicellular or oligocellular ancestors, and may have been co-opted to serve similar functions independently. Perhaps conserved epigenetic mechanisms, such as the Polycomb system, served functions in genome organization, genome defense, chromosome biology, and cellular differentiation, rather than multicellular transcriptional memory, in the ancestral unicellular eukaryote. *Arabidopsis* is playing a major role in identifying these conserved functions within and between kingdoms.

## THE NEMATODE WORM, *CAENORHABDITIS ELEGANS*

---

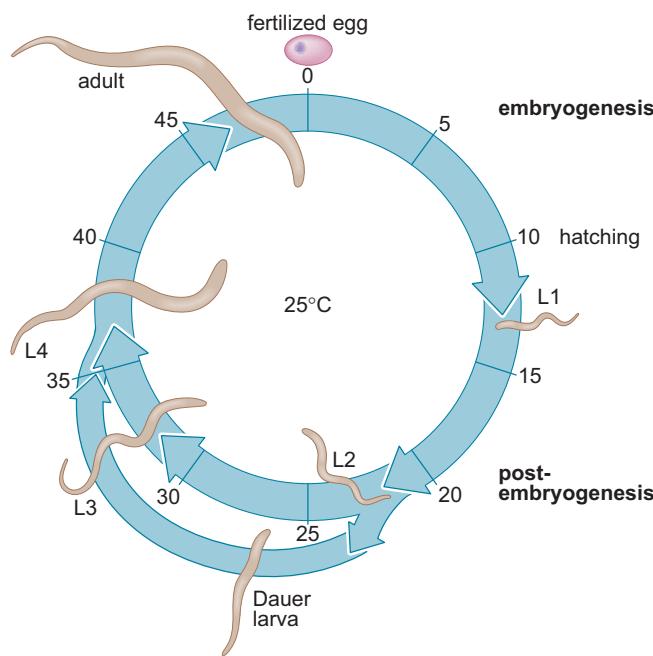
Brenner, after making seminal contributions in molecular genetics, identified a small metazoan in which to study the important questions of development and the molecular basis of behavior. Learning from the success of molecular genetic studies in phage and bacteria, he wanted the simplest possible organism that had differentiated cell types, but that was also amenable to microbiological-like genetics. In 1965 he settled on the small nematode worm *C. elegans* because it contained a variety of suitable characteristics. These include a rapid generation time to enable genetic screens, hermaphrodite reproduction producing hundreds of “self-progeny” so that large numbers of animals could be clonally generated, sexual reproduction so that genetic stocks could be constructed by mating, and a small number of transparent cells so that development could be followed directly.

Brenner set two ambitious initial goals that would be essential for the long-term success of this endeavor. One was a complete physical mapping of all cells (and cell–cell interactions) by reconstructing serial section electron micrographs (completed by John White in 1986). The second goal was mapping of the entire cell lineage of the animal (completed by John Sulston in 1983). This revealed how each cell in the adult worm arose during development and showed how progeny cells were related to each other in the final differentiated animal. Seven years later Brenner established the genetics of the new model organism with the isolation of more than 300 morphological and behavioral mutants. These defined more than 100 complementation groups mapping to six linkage groups. Nearly 30 years later there are 400 laboratories worldwide that study *C. elegans*. Because of its simplicity and experimental accessibility, it is now one of the most completely understood of all metazoans.

### *C. elegans* Has a Very Rapid Life Cycle

*C. elegans* is cultured on petri dishes and fed a simple diet of bacteria. They grow well at a range of temperatures, growing twice as fast at 25°C than at 15°C. At 25°C fertilized embryos complete development in 12 hours and hatch into free-living animals capable of complex behaviors. The hatchling worm passes through four juvenile or larval stages (L1–L4) over the course of 40 hours to become a sexually mature adult (Fig. A-14).

The adult hermaphrodite can produce up to 300 self-progeny over the course of about 4 days or can be mated with rare males to produce up to 1000 hybrid progeny. The adult lives for about 15 days. Under stressful conditions (low food, increased temperatures, or high population density), the L1 stage animal can enter an alternative developmental pathway leading to what is called a **dauer**. Dauers are resistant to environmental stresses and can live many months while waiting for environmental conditions to improve. The study of mutants that fail to enter the dauer stage, or that enter

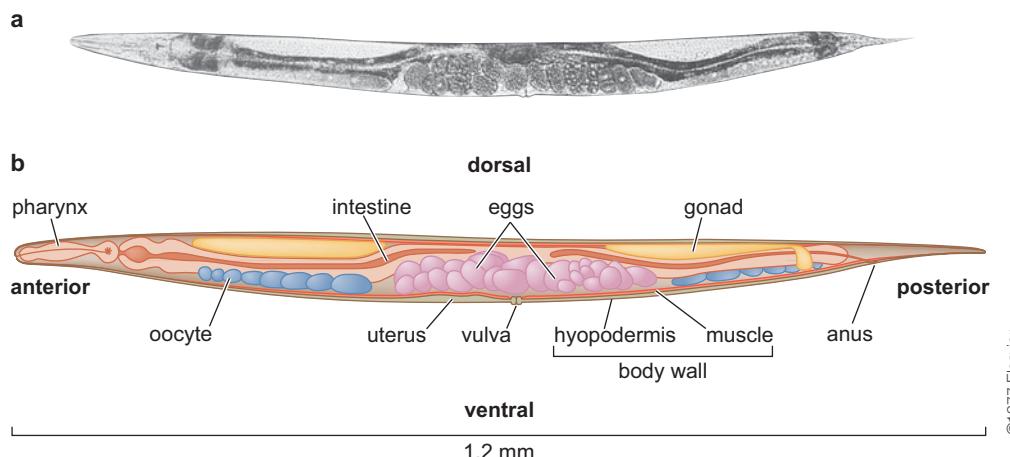


**FIGURE A-14** The life cycle of the worm, *C. elegans*. Shown is the life cycle in hours of development, from first-stage juvenile to adult, as described in the text. The alternative developmental stage—a dauer—is also shown.

it inappropriately, have identified genes expressed in specific neurons that function to sense environmental conditions, genes expressed throughout the animal that control body growth, and genes that control life span. Activation of these latter genes in the adult can dramatically extend the life span of the animal and homologs of these genes have been implicated in life extension in mammals.

### *C. elegans* Is Composed of Relatively Few, Well-Studied Cell Lineages

*C. elegans* has a simple body plan (Fig. A-15). The prominent organ in the adult hermaphrodite is the gonad, which contains the proliferating and differentiating germ cells (sperm and oocytes), fertilization chamber



**FIGURE A-15** The body plan of the worm. (a) A section through an adult hermaphrodite worm is shown. (b) The various organs are identified in the sketch below (b) and are described in the text. (a, Reprinted, with permission, from Sulston J.E. and Horvitz H.R. 1977. *Dev. Biol.* **56**: 110–156. © Elsevier.)

(spermatheca), and uterus for temporary storage of young embryos. The embryos pass from the uterus to the outside through the vulva, a structure formed postembryonically from 22 epidermal cells. Mutations that disrupt the formation of the vulva do not interfere with production of embryos but do prevent the eggs from being laid. Consequently, the embryos develop and hatch inside the uterus. The hatched worms then devour their mother and become trapped inside her skin (cuticle layer) forming a “bag of worms.” This readily identified phenotype has allowed the isolation of hundreds of vulva-less mutants identifying scores of genes that function to control the generation, specification, and differentiation of the vulval cells, indicating that the construction of this simplified “organ” requires 10s to 100s of genes. Among these genes are components of a highly conserved receptor tyrosine kinase signaling pathway that controls cell proliferation.

Many of the mammalian homologs of these genes are oncogenes and tumor-suppressor genes that when altered can lead to cancer. In *C. elegans*, mutations that inactivate this pathway eliminate vulva development because the vulval cells are never generated, whereas mutations that activate this pathway cause overproliferation of the vulva precursor cells, resulting in a multiple vulva phenotype. Because the animal is transparent and the vulva is generated from only 22 cells, it is possible to describe the mutant defect with cellular resolution such that the type of mutation can be associated with a specific cellular transformation. Furthermore, cell-autonomous versus cell-nonautonomous function of particular gene products can be distinguished.

### The Cell Death Pathway Was Discovered in *C. elegans*

Although the process of cell death had been recognized in a variety of developmental contexts (e.g., formation of digits and loss of tadpole tails), the demonstration that cell death is “programmed” and is a genetically regulated process were major insights provided by the genetic and experimental tractability of *C. elegans*. Early analysis of cell lineages noted that the same set of cells died in every animal, suggesting that cell death was under genetic control. The first cell death defective (*ced*) mutants isolated were defective for the consumption of the cell corpse by neighboring cells; thus in the mutants cell corpses persisted for many hours. Using these *ced* mutants, H. Robert Horvitz and his colleagues isolated many additional *ced* mutants that failed to produce persistent cell corpses. These mutants proved to be defective at initiating the cell death program. Analysis of the *ced* mutants showed that, in all but one case, developmentally programmed cell death is cell autonomous—that is, the cell commits suicide. In males, a cell known as the linker cell is killed by its neighbor. The molecular identification of the *ced* genes provided the means to identify proteins in mammals that carry out essentially the identical biochemical reactions to control cell death in all animals; in fact, expressing human homologs in *C. elegans* can substitute for a mutated *ced* gene. Cell death is as important as cell proliferation in development and disease and is the focus of intense research to develop therapeutics for the control of cancer and neurodegenerative diseases.

### RNAi Was Discovered in *C. elegans*

In 1998 a remarkable discovery was announced. The introduction of double-stranded RNA (dsRNA) into *C. elegans* silenced the gene homologous to

the dsRNA (for a full discussion of gene silencing, see Chapter 20). Although this phenomenon was recognized in other model organisms as well, the ease of both the genetic and developmental manipulation of *C. elegans* (e.g., the generation of mutants incapable of carrying out silencing) ensured *C. elegans* was at the forefront of the elucidation of this form of gene silencing. This unexpected discovery and subsequent analysis of RNA interference (RNAi) is significant in three respects. One is that RNAi appears to be universal because introduction of dsRNA into nearly all animal, fungal, or plant cells leads to homology-directed mRNA degradation. Indeed, much of what we know about RNAi comes from studies in plants (Chapter 20). The second was the rapidity with which experimental investigation of this mysterious process revealed the molecular mechanisms (see Fig. 20-10). The third significant aspect of RNAi is how widely exploited it has become to manipulate gene expression in multiple organisms. This offers a “systematic strategy” to identify genes involved in a given process rather than one might also do through traditional genetics. Thus, a library of interfering RNAs can be directed against all the genes within an organism, and by correlating specific phenotypes with particular RNAs from the library, one can track down the genes that play a role in that phenotype. This approach can be applied even in organisms not amenable to traditional genetics. These investigations intersected with the analysis of another RNA-mediated gene regulatory process that involves tiny endogenous miRNAs that have been shown to regulate gene expression in plants and animals, coordinate genome rearrangements in ciliates, and regulate chromatin structure in yeast. The first two miRNAs were discovered in genetic screens in *C. elegans*. A fraction of these worm miRNAs is conserved in flies and mammals, where their functions are just beginning to be revealed. Recent studies suggest that the human genome may contain 1000s of miRNA genes.

## THE FRUIT FLY, *DROSOPHILA MELANOGLASTER*

---

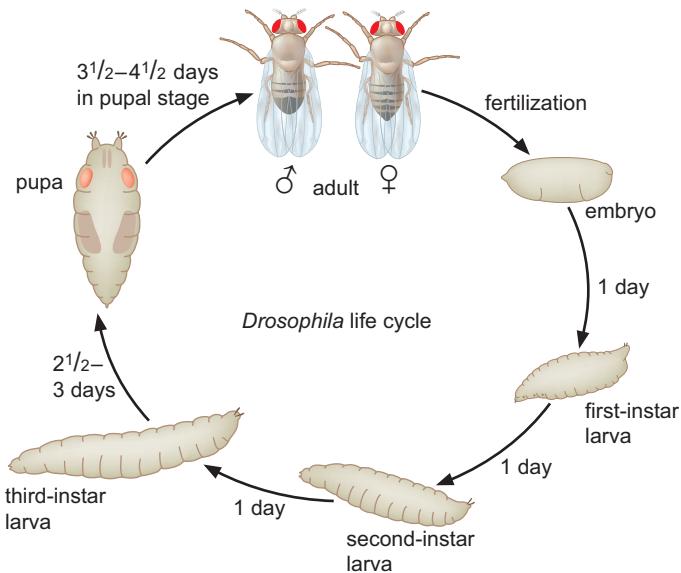
We are approaching the 100th anniversary of the fruit fly as a model organism for studies in genetics and developmental biology. In 1908 Thomas Hunt Morgan and his research associates at Columbia University placed rotting fruit on the window ledge of their laboratory in Schermerhorn Hall. Their goal was to isolate a small, quickly reproducing animal that could be cultured in the lab and used to study the inheritance of quantitative traits, such as eye color. Among the menagerie of creatures that were captured, the fruit fly emerged as the animal of choice. Adults produced large numbers of progeny in just 2 weeks. Culturing was done in recycled milk bottles using an inexpensive concoction of yeast and agar.

### *Drosophila* Has a Rapid Life Cycle

The salient features of the *Drosophila* life cycle are a very rapid period of embryogenesis, followed by three periods of larval growth prior to metamorphosis (Fig. A-16). Embryogenesis is completed within 24 hours after fertilization and culminates in the hatching of a first-instar larva. As we discussed in Chapter 21, the early periods of *Drosophila* embryonic development exhibit the most rapid nuclear cleavages known for any animal. A first-instar larva grows for 24 hours and then molts into a larger, second-instar larva. The process is repeated to yield a third-instar larva that feeds and grows for 2–3 days.

**FIGURE A-16** The *Drosophila* life cycle.

The various stages of development of the fly, shown here, are described in the text.

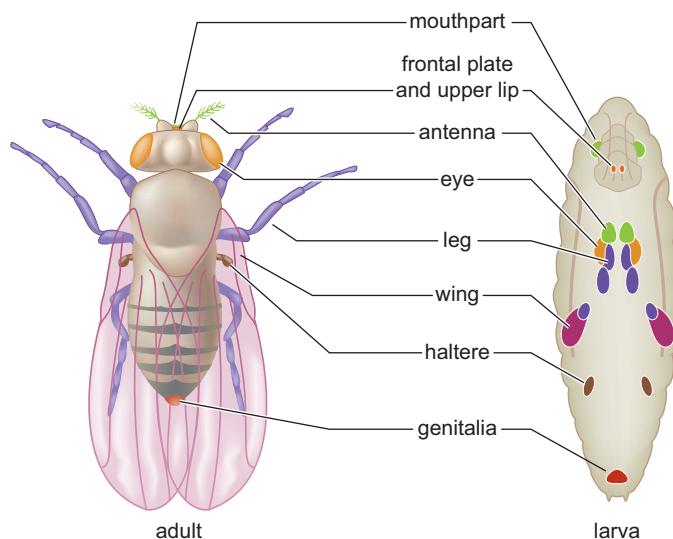


One of the key processes that occurs during larval development is the growth of the imaginal disks, which arise from invaginations of the epidermis in mid-stage embryos (Fig. A-17). There is a pair of disks for every set of appendages (e.g., a set of foreleg imaginal disks and a set of wing imaginal disks). There are also imaginal disks for eyes, antennae, the mouthparts, and genitalia. Disks are initially small and composed of fewer than 100 cells in the embryo but contain tens of thousands of cells in mature larvae. The development of the wing imaginal disk has become an important model system for understanding how gradients of secreted signaling molecules such as Hedgehog and Dpp (TGF- $\beta$ ) control complex patterning processes. Imaginal disks differentiate into their appropriate adult structures during metamorphosis (or pupation).

### The First Genome Maps Were Produced in *Drosophila*

In 1910 the Morgan lab identified a spontaneous mutant male fly that had white eyes rather than the brilliant red seen for normal strains. This single

**FIGURE A-17** Imaginal disks in *Drosophila*. The position of various imaginal disks in the larva are shown on the right. On the left are shown the limbs and the organs they form in the adult fly. These disks are initially formed as small groups of cells in the embryo but have grown to tens of thousands of cells in the mature larva. These disks develop into their respective adult structures during pupation.



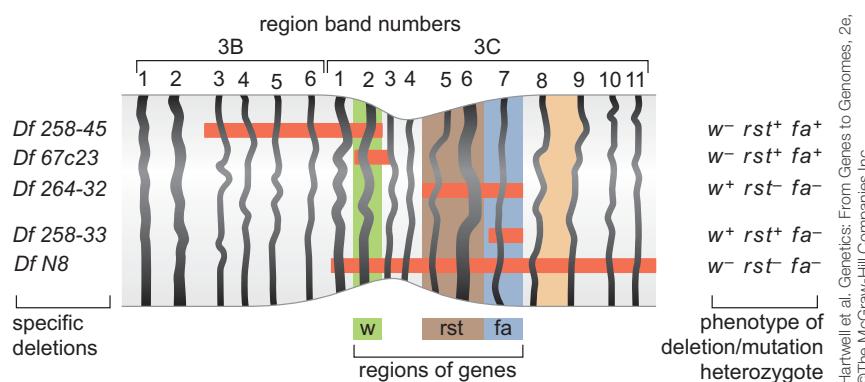
fly launched an incisive series of genetic studies that led to two major discoveries: genes are located on chromosomes, and each gene is composed of two alleles that assort independently during meiosis (see Mendel's first law; Chapter 1). The identification of additional mutations led to the demonstration that genes located on separate chromosomes segregate independently (Mendel's second law), whereas those linked on the same chromosome do not.

An undergraduate at Columbia University, Alfred H. Sturtevant (a member of the Morgan lab), developed a simple mathematical algorithm for mapping the distances between linked genes based on recombination frequencies. The simplicity and power of this work had an enormous impact that fundamentally changed genetics and provided the first demonstration that genes are physically defined and ordered entities along the chromosomes. By the 1930s, extensive genetic maps were produced that identified the relative positions of numerous genes controlling a variety of physical characteristics of the adult, such as wing size and shape and eye color and shape.

Hermann J. Muller, another scientist trained in the Morgan fly lab, provided the first evidence that environmental factors, such as ionizing radiation, can cause chromosome rearrangements and genetic mutations. Large-scale "genetic screens" are routinely performed by feeding adult males a mutagen, such as EMS, and then mating them with normal females. The F<sub>1</sub> progeny are heterozygous and contain one normal chromosome and one random mutation. A variety of methods are used to study these mutations, as described later.

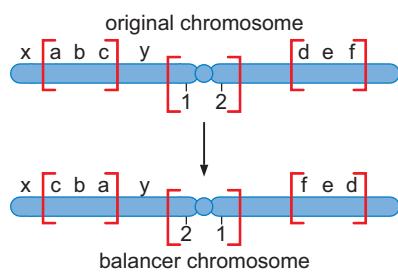
In addition to its remarkable fecundity (a single female can produce thousands of eggs) and rapid life cycle, the fruit fly was found to possess several very useful features that guaranteed it a sustained and prominent role in experimental research. It contains only four chromosomes: two large autosomes, chromosomes 2 and 3, a smaller X chromosome (which determines sex), and a very small fourth chromosome. Calvin B. Bridges—yet another of Muller's colleagues—discovered that certain tissues in *Drosophila* larvae undergo extensive endoreplication without cell division. In the salivary gland, this process produces remarkable giant chromosomes composed of approximately 1000 copies of each chromatid. Bridges used these **polytene chromosomes** to determine a physical map of the *Drosophila* genome (the first produced for any organism) (Fig. A-18).

Bridges identified a total of approximately 5000 "bands" on the four chromosomes and established a correlation between many of these bands and the locations of genetic loci identified in the classical recombination maps. For example, female fruit flies that are heterozygous for the recessive *white*



**FIGURE A-18** Genetic maps, polytene chromosomes, and deficiency mapping. Endoreplication in the absence of cytokinesis generates enlarged chromosomes in some tissues of the fly, most notably the salivary glands where the giant chromosomes are composed of a thousand chromatids. It was possible, for the first time, to correlate the occurrence of genes for certain traits with specific physical segments of chromosomes. For example, white eye flies were correlated with deletions in the 3C region of the X chromosome. (With permission from Hartwell L. et al. 2004. *Genetics: From genes to genomes*, 2nd ed., p. 816, Fig. D-4. © McGraw-Hill.)

Hartwell et al. *Genetics: From Genes to Genomes*, 2nd ed.  
© The McGraw-Hill Companies Inc.



**FIGURE A-19 Balancer chromosome.** Balancer chromosomes (bottom panel) contain a series of inversions when compared with the original, parental chromosome (top panel). In this diagram, a hypothetical chromosome has two arms. The left arm of the balancer chromosome has an internal inversion that reverses the order of genes a, b, and c in the original chromosome. Similarly, the arm on the right of the balancer chromosome has an inversion that reverses the order of genes d, e, and f. In addition, there might be an inversion centered around the centromere, in this case reversing the order of genes 1 and 2. The balancer chromosome thus has a significantly different order of genes when compared with the original. As a result, there is a suppression of recombination between the chromosomes in heterozygotes containing one copy of each.

mutation exhibit normal red eyes. However, similar females that contain the *white* mutation and a small deletion in the other *X* chromosome, which removes polytene bands 3C2–3C3, exhibit white eyes. This is because there is no longer a normal, dominant copy of the gene. This type of analysis led to the conclusion that the *white* gene is located somewhere between polytene bands 3C2 and 3C3 on the *X* chromosome.

A variety of additional genetic methods were created to establish the fruit fly as the premiere model organism for studies in animal inheritance. For example, **balancer chromosomes** were created that contain a series of inversions relative to the organization of the native chromosome (Fig. A-19). Critically, such balancers fail to undergo recombination with the native chromosome during meiosis. As a result, it is possible to maintain permanent cultures of fruit flies that contain recessive, lethal mutations. Consider a null mutation in the *even-skipped* (*eve*) gene, which we discussed in Chapter 21. Embryos that are homozygous for this mutation die and fail to produce viable larvae and adults. The *eve* locus maps on chromosome 2 (at polytene band 46C). The null mutation can be maintained in a population that is heterozygous for a “normal” chromosome containing the null allele of *eve* and a balancer second chromosome, which contains a normal copy of the gene. Because the *eve* null allele is strictly recessive, these flies are completely viable. However, only heterozygotes are observed among adult progeny in successive generations. Embryos that contain two copies of the balancer chromosome die because some of the inversions produce recessive disruptions in critical genes. In addition, embryos that contain two copies of the normal chromosome die because they are homozygous for the *eve* null mutation.

### Genetic Mosaics Permit the Analysis of Lethal Genes in Adult Flies

**Mosaics** are animals that contain small patches of mutant tissue in a generally “normal” genetic background. Such small patches do not kill the individual because most of the tissues in the organism are normal. For example, small patches of *engrailed/engrailed* homozygous mutant tissue can be produced by inducing mitotic recombination in developing larvae using X-rays. When such patches are created in posterior regions of the developing wings, the resulting flies exhibit abnormal wings that have duplicated anterior structures in place of the normal posterior structures. The analysis of genetic mosaics provided the first evidence that *Engrailed* is required for subdividing the appendages and segments of flies into anterior and posterior compartments.

The most spectacular genetic mosaics are gynandromorphs (Fig. A-20). These are flies that are literally half male and half female. Sexual identity in flies is determined by the number of *X* chromosomes. Individuals with two *X* chromosomes are females, whereas those with just one *X* are males. (The *Y* chromosome does not define sexual identity in flies as it does in mice and humans: in flies, *Y* is only needed for the production of sperm.) Rarely, one of the two *X* chromosomes is lost at the first mitotic division following the fusion of the sperm and egg pronuclei in a newly fertilized *XX* embryo.

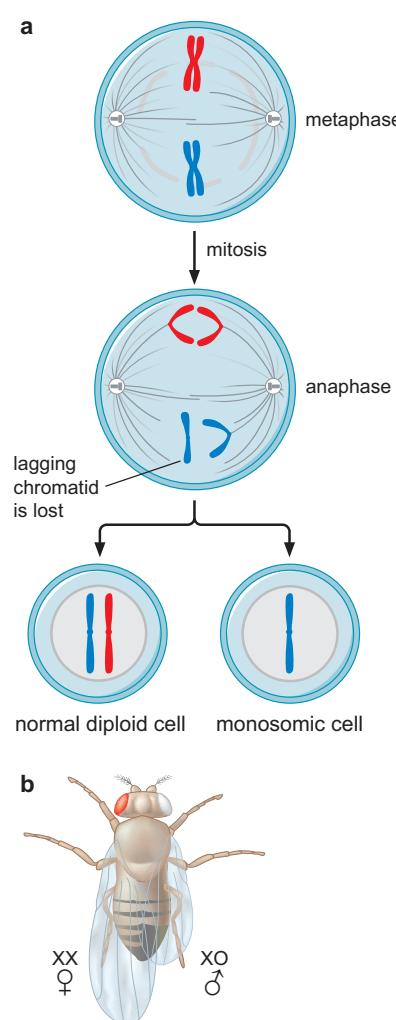
This *X* instability occurs only at the first division. In all subsequent divisions, nuclei containing two *X* chromosomes give rise to daughter nuclei with two *X* chromosomes, whereas nuclei with just one *X* chromosome give rise to daughters containing a single *X*. As we discussed in Chapter 21, these nuclei undergo rapid cleavages without cell membranes and then migrate to the periphery of the egg. This migration is coherent and there is little or no intermixing of nuclei containing one *X* chromosome with

nuclei containing two *X* chromosomes. Thus, half the embryo is male and half is female, although the “line” separating the male and female tissues is random. Its exact position depends on the orientation of the two daughter nuclei after the first cleavage. The line sometimes bisects the adult into a left half that is female and a right half that is male. Suppose that one of the *X* chromosomes contains the recessive white allele. If the wild-type *X* chromosome is lost at the first division, then the right half of the fly, the male half, has white eyes (the male half has only the mutant *X* chromosome), whereas the left half (the female side) has red eyes. (Remember that the female half has two *X* chromosomes and that one contains the dominant, wild-type allele.)

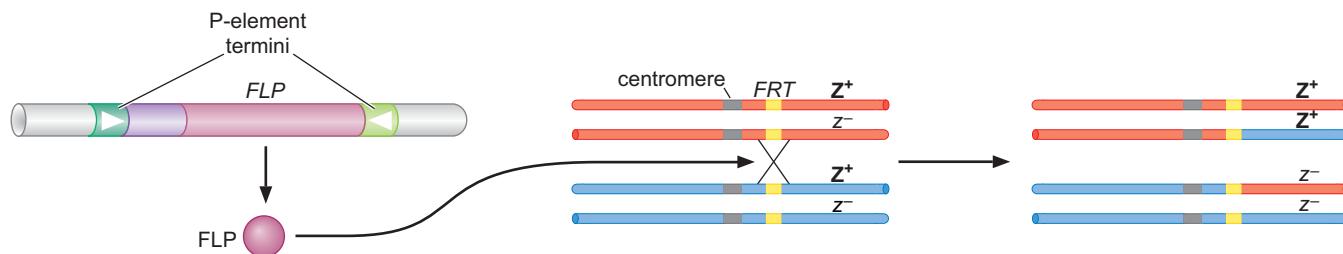
### The Yeast FLP Recombinase Permits the Efficient Production of Genetic Mosaics

What was not anticipated during the classical era of genetic analysis is the fact that *Drosophila* possesses several favorable attributes for molecular studies and whole-genome analysis. Most notably, the genome is relatively small. It is composed of only approximately 150 Mb and contains fewer than 14,000 protein coding genes. This represents just 5% of the amount of DNA that makes up the mouse and human genomes. As the fruit fly entered the modern era, several methods were established that improved some of the older techniques of genetic manipulation and also led to completely new experimental methods, such as the production of stable transgenic strains carrying recombinant DNAs.

As we discussed earlier, genetic mosaics are produced by mitotic recombination in somatic tissues. Initially, X-rays were used to induce recombination, although this method is inefficient and produces small patches of mutant tissue. More recently, the frequency of mitotic recombination was greatly enhanced by the use of the FLP recombinase from yeast (Fig. A-21). FLP recognizes a simple sequence motif, FRT, and then catalyzes DNA rearrangement (see Chapter 12). FRT sequences were inserted near the centromere of each of the four chromosomes using P-element transformation (see later discussion). Heterozygous flies are then produced that contain a null allele in gene *Z* on one chromosome and a wild-type copy of that gene on the homologous chromosome. Both chromosomes contain the FRT sequences. These flies are stable and viable as there is no endogenous FLP recombinase in *Drosophila*. It is, however, possible to introduce the recombinase in transgenic strains that contain the yeast FLP protein-coding sequence under the control of the heat-inducible *hsp70* promoter. Upon heat shock, FLP is synthesized in all cells. FLP binds to the FRT motifs in the two homologs containing gene *Z* and catalyze mitotic recombination (Fig. A-21). This method is quite efficient. In fact, short pulses of heat shock

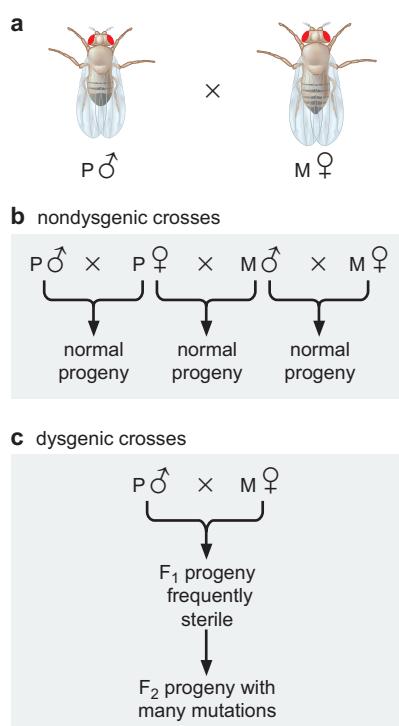


**FIGURE A-20** Gynandromorphs. Gynandromorph mutants are a particularly striking form of genetic mosaicism. (a) The blue *X* chromosome carries the recessive (*white*) mutation, whereas the red *X* chromosome has a normal dominant copy of the gene. The mutant is the result of *X* chromosome loss at the first mitotic division in an XX (female) fly as described in the text. (b) In the resulting mutant, one half of the fly is female, the other is male.



**FIGURE A-21** FLP-FRT. The use of this site-specific recombination system from yeast (described in Chapter 12) promotes high levels of mitotic recombination in flies. The recombination is controlled by expressing the recombinase in flies only when required.

are often sufficient to produce enough FLP recombinase to produce large patches of  $z^-/z^-$  tissue in different regions of an adult fly. FRT recognition sequences have been inserted throughout the *Drosophila* genome via P-transformation. It is now possible to create small deletions for just about any gene by inducing rearrangements between FRT sites flanking the gene of interest using the FLP recombinase.



**FIGURE A-22** Hybrid dysgenesis. P-element transposons reside passively in P strains because they express a repressor that keeps the transposons silent. When P strains are mated with an M strain lacking such a repressor, the transposons are mobilized within the pole cells and often integrate into genes required for germ cell formation. This explains the high frequency of sterility in the offspring from such crosses.

### It Is Easy to Create Transgenic Fruit Flies that Carry Foreign DNA

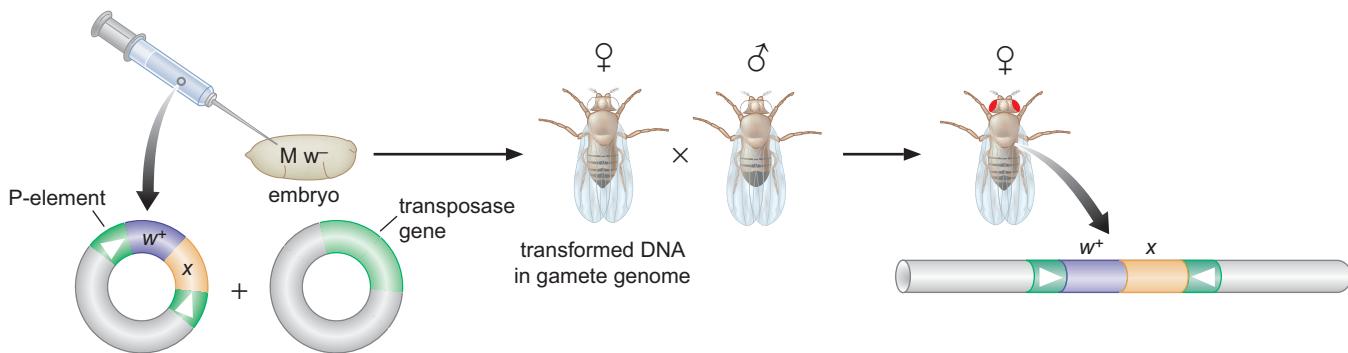
P-elements are transposable DNA segments that are the causal agent of a genetic phenomenon called **hybrid dysgenesis** (Fig. A-22). Consider the consequences of mating females from the “M” strain of *D. melanogaster* with males from the “P” strain (same species, but different populations). The F<sub>1</sub> progeny are often sterile. The reason is that the P strain contains numerous copies of the P-element transposon that are mobilized in embryos derived from M eggs. These eggs lack a repressor protein that inhibits P-element mobilization. P-element excision and insertion is limited to the pole cells, the progenitors of the gametes (sperm in males and eggs in females). Sometimes the P-elements insert into genes that are essential for the development of these germ cells, and, as a result, the adult flies derived from these matings are sterile.

P-elements are used as transformation vectors to introduce recombinant DNAs into otherwise normal strains of flies (Fig. A-23). A full-length P-element transposon is 3 kb in length. It contains inverted repeats at the termini that are essential for excision and insertion. The intervening DNA encodes both a repressor of transposition and a transposase that promotes mobilization. The repressor is expressed in the developing eggs of P strains. As a result, there is no movement of P-elements in embryos derived from females of the P strain (these contain P-elements). Movement is seen only in embryos derived from eggs produced by M-strain females, which lack P-elements.

Recombinant DNA is inserted into defective P-elements that lack the internal genes encoding repressor and transposase. This DNA is injected into posterior regions of early, precellular embryos (as we saw in Chapter 21, this is the region that contains the polar granules). The transposase is injected along with the recombinant P-element vector. As the cleavage nuclei enter posterior regions, they acquire both the polar granules and recombinant P-element DNA together with transposase. The pole cells bud off from the polar plasm and the recombinant P-elements insert into random positions in the pole cells. Different pole cells contain different P-element insertion events. The amount of recombinant P-element DNA and transposase is calibrated so that, on average, a given pole cell receives just a single integrated P-element. The embryos are allowed to develop into adults and then mated with appropriate tester strains.

The recombinant P-element contains a “marker” gene such as *white*<sup>+</sup> and the strain used for the injections is a *white*<sup>-</sup> mutant. The tester strains are also *white*<sup>-</sup>, so that any F<sub>2</sub> fly that has red eyes must contain a copy of the recombinant P-element. This method of P-element transformation is routinely used to identify regulatory sequences such as those governing *eve stripe 2* expression (which we discussed in Chapter 21). In addition, this strategy is used to examine protein-coding genes in various genetic backgrounds.

In summary, *Drosophila* offers many of the sophisticated tools of classical and molecular genetics that, as we have seen, are available in microbial model systems. One conspicuous exception has been the absence of



**FIGURE A-23** P-element transformation. P-elements can be used as vectors in the transformation of fly embryos. Thus, as discussed in the text, sequences of choice can be inserted into a modified P-element. A single copy of this recombinant molecule is stably incorporated into a single location of a fly chromosome.

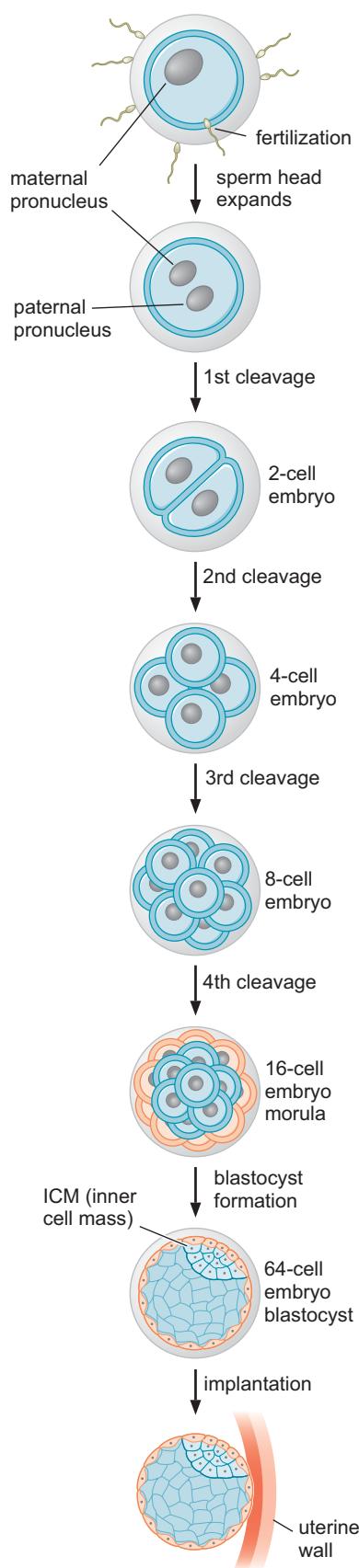
methods for precise manipulation of the genome by homologous recombination with recombinant DNA, such as in the creation of gene deletions. However, such methods were recently developed and are now being streamlined for routine use. Ironically, such manipulations are readily available, as we shall see, in the more complicated model system, the mouse. Nevertheless, because of the wealth of genetic tools available in *Drosophila* and the extensive groundwork of knowledge about this organism resulting from decades of investigation, the fruit fly remains one of the premier model systems for studies of development and behavior.

## THE HOUSE MOUSE, *MUS MUSCULUS*

By the standards of *C. elegans* and *Drosophila*, the life cycle of the mouse is slow and cumbersome. Embryonic development, or gestation, occurs over a period of 3 weeks and the newborn mouse does not reach puberty for another 5–6 weeks. Thus, the effective life cycle is roughly 8–9 weeks, more than five times longer than that of *Drosophila*. The mouse, however, enjoys a special status because of its exalted position on the evolutionary tree: it is a mammal and, therefore, related to humans. Of course, chimpanzees and other higher primates are closer to humans than mice, but they are not amenable to the various experimental manipulations available in mice.

Thus, the mouse provides the link between the basic principles, discovered in simpler creatures like worms and flies, and human disease. For example, the *patched* gene of *Drosophila* encodes a critical component of the Hedgehog receptor (Chapter 21). Mutant fly embryos that lack the wild-type *patched* gene activity exhibit a variety of patterning defects. The orthologous genes in mice are also important in development. Unexpectedly, however, certain *patched* mutants cause various cancers, such as skin cancer, in both mice and humans. No amount of analysis in the fly would reveal such a function. In addition, methods have been developed that permit the efficient removal of specific genes in otherwise normal mice. This “knockout” technology continues to have an enormous impact on our understanding of the basic mechanisms underlying human development, behavior, and disease. We briefly review the salient features of the mouse as an experimental system.

The chromosome complement of the mouse is similar to that seen in humans: there are 19 autosomes in mice (22 in humans), as well as X



**FIGURE A-24** Overview of mouse embryogenesis.

and Y sex chromosomes. There is extensive synteny between mice and humans: extended regions of a given mouse chromosome contain the same set of genes (in the same order) as the “homologous” regions of the corresponding human chromosomes. The mouse genome has been sequenced and assembled. As discussed in Chapter 21, the mouse has virtually the same complement of genes as those present in the human genome: each contains approximately 25,000 genes, and there is a one-to-one correspondence for more than 85% of these genes. Most, if not all, of the differences between the mouse and human genomes stem from the selective duplication of certain gene families in one lineage or the other. Comparative genome analysis confirms what we have known for some time: the mouse is an excellent model for human development and disease.

### Mouse Embryonic Development Depends on Stem Cells

Mouse eggs are small and difficult to manipulate. Like human eggs, they are just 100 microns in diameter. Their small size prohibits grafting experiments of the sort done in zebrafish and frogs, but microinjection methods have been developed for introducing recombinant DNA into the mouse germ line so as to create transgenic strains, as discussed later. In addition, it is possible to harvest enough mouse embryos, even at the earliest stages, for *in situ* hybridization assays and the visualization of specific gene expression patterns. Such visualization methods can be applied to both normal embryos and mutants carrying disruptions in defined genetic loci.

Figure A-24 shows an overview of mouse embryogenesis. The initial divisions of the early mouse embryo are very slow and occur with an average frequency of just once every 12–24 hours. The first obvious diversification of cell types is seen at the 16-cell stage, called the **morula** (Fig. A-24, panel 6). The cells located in outer regions form tissues that do not contribute to the embryo but instead develop into the placenta. Cells located in internal regions generate the inner cell mass (ICM). At the 64-cell stage, there are only 13 ICM cells, but these form all of the tissues of the adult mouse. The ICM is the prime source of embryonic stem cells, which can be cultured and induced to form any adult cell type upon addition of the appropriate growth factors. Human stem cells have become the subject of considerable social controversy, but offer the promise of providing a renewable source of tissues that can be used to replace defective cells in a variety of degenerative diseases such as diabetes and Alzheimer’s.

At the 64-cell stage (about 3–4 days after fertilization) the mouse embryo, now called a **blastocyst**, is finally ready for implantation. Interactions between the blastocyst and uterine wall lead to the formation of the placenta, a characteristic of all mammals except the primitive egg-laying platypus. After formation of the placenta, the embryo enters gastrulation, whereby the ICM forms all three germ layers: endoderm, mesoderm, and ectoderm. Shortly thereafter, a fetus emerges that contains a brain, a spinal cord, and internal organs such as the heart and liver.

The first stage in mouse gastrulation is the subdivision of the ICM into two cell layers: an inner hypoblast and an outer epiblast, which form the endoderm and ectoderm, respectively. A groove called the **primitive streak** forms along the length of the epiblast and the cells that migrate into the groove form the internal mesoderm. The anterior end of the primitive streak is called the **node**; it is the source of a variety of signaling molecules that are used to pattern the anterior–posterior axis of the embryo, including two secreted inhibitors of TGF- $\beta$  signaling, Chordin and Noggin. Double mutant mouse embryos that lack both genes develop into fetuses that lack head structures such as the forebrain and nose.

## It Is Easy to Introduce Foreign DNA into the Mouse Embryo

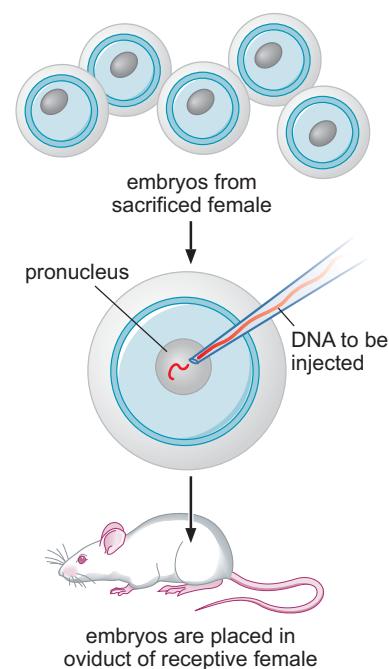
Microinjection methods have been developed for the efficient expression of recombinant DNA in transgenic strains of mice. DNA is injected into the egg pronucleus, and the embryos are placed into the oviduct of a female mouse and allowed to implant and develop. The injected DNA integrates at random positions in the genome (Fig. A-25). The efficiency of integration is quite high and usually occurs during early stages of development, often in one-cell embryos. As a result, the fusion gene inserts into most or all of the cells in the embryo, including the ICM cells that form the somatic tissues and germline of the adult mouse. Approximately 50% of the transgenic mice that are produced using this simple method of microinjection exhibit **germ-line transformation**; that is, their offspring also contain the foreign recombinant DNA.

Consider as an example a fusion gene containing the enhancer from the *Hoxb2* gene attached to a *lacZ* reporter gene. Embryos and fetuses can be harvested from transgenic strains carrying this reporter and stained to reveal the pattern of *lacZ* expression. In this case, staining is observed in the hindbrain (Fig. A-26). Transgenic mice have been used to characterize several regulatory sequences, including those that regulate the  $\beta$ -globin genes and *HoxD* genes. Both complex loci contain long-range regulatory elements (the LCR and GCR, respectively) that coordinate the expression of the different genes over distances of several hundred kilobases (see Chapter 19).

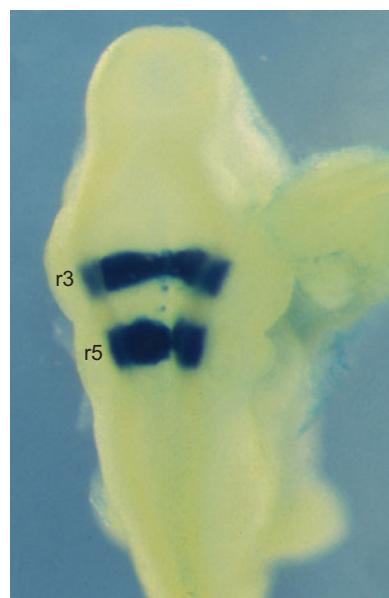
## Homologous Recombination Permits the Selective Ablation of Individual Genes

The single most powerful method of mouse transgenesis is the ability to disrupt, or “knock out,” single genetic loci. This permits the creation of mouse models for human disease. For example, the *p53* gene encodes a regulatory protein that activates the expression of genes required for DNA repair. It has been implicated in a variety of human cancers. When *p53* function is lost, cancer cells become highly invasive because of rapid accumulation of DNA mutations. A strain of mice has been established that is completely normal except for the removal of the *p53* gene. These mice, which are highly susceptible to cancer, die young. There is the hope that these mice can be used to test potential drugs and anticancer agents for use in humans. Although *Drosophila* contains a *p53* gene, and mutants have been isolated, it does not provide the same opportunity for drug discovery as does the mouse model.

Gene disruption experiments are done with embryonic stem (ES) cells (Fig. A-27). These cells are obtained by culturing mouse blastocysts so that ICM cells proliferate without differentiating. A recombinant DNA is created that contains a mutant form of the gene of interest (ES cells can also be generated from somatic cells by a newly devised procedure involving reprogramming with a small number of transcription factors [see Chapter 21, Box 21-1]). For example, the protein coding region of a given target gene is modified by deleting a small region near the beginning of the gene that removes codons for essential amino acids from the encoded protein

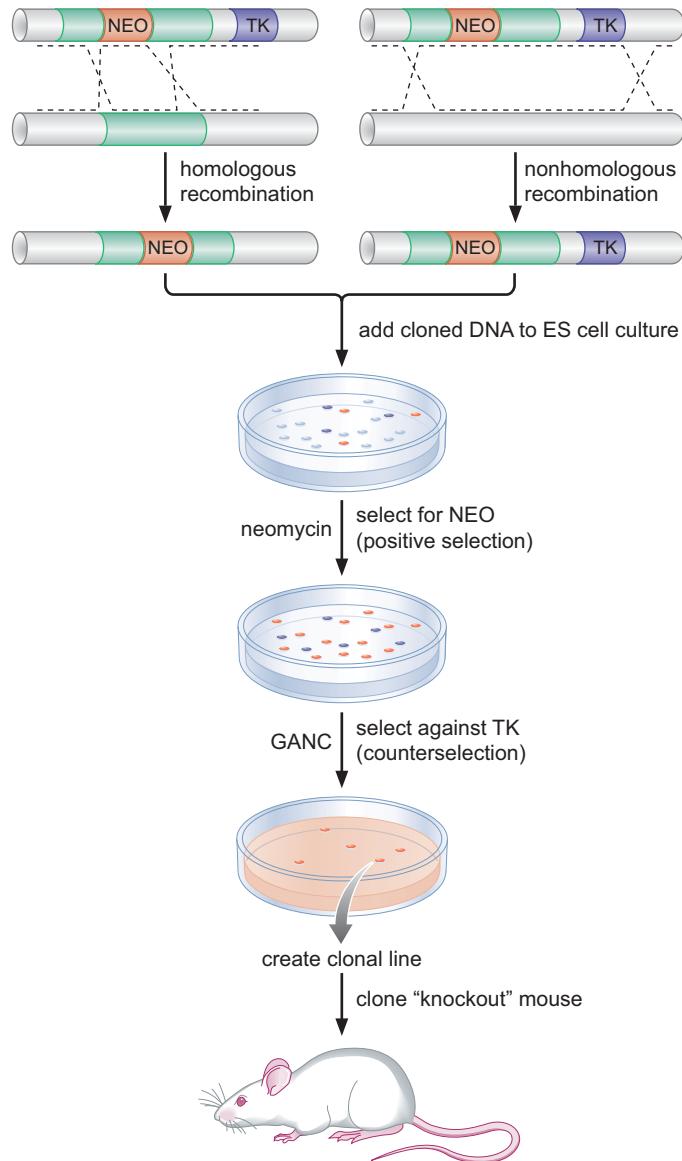


**FIGURE A-25** Creation of transgenic mice by microinjection of DNA into the egg pronucleus. One-cell embryos are obtained from a newly mated female mouse. Recombinant DNA is injected into the nucleus, and the embryo is then implanted into the oviduct of a surrogate. After several days, the embryo implants and ultimately forms a fetus that contains integrated copies of the recombinant DNA.



**FIGURE A-26** *In situ* expression patterns of embryos obtained from transgenic mice. A transgenic strain of mice was created that contains a portion of the *Hoxb2* regulatory region attached to a *lacZ* reporter gene. Embryos were obtained from transgenic females and stained to reveal sites of  $\beta$ -galactosidase (*LacZ*) activity. There are two prominent bands of staining detected in the hindbrain region of 10.5 day embryos. The embryo is displayed with the head up and the tail down. (Nonchev et al. 1996. *Proc. Natl. Acad. Sci.* **93**: 9339–9345, Fig. 1c.)

**FIGURE A-27** Gene knockout via homologous recombination. The figure outlines the method used to create a cell line lacking any given gene. Homologous recombination that occurs within a target gene (shown in green) results in the incorporation of NEO and disruption of that gene. Nonhomologous, or random, recombination can result in the incorporation of the disrupted gene containing NEO, and the gene encoding thymidine kinase (TK). Clones carrying both constructs survive exposure to neomycin, but the clones also carrying TK are subsequently counterselected by growth in gancyclovir (GANC). Clones containing the NEO insertion via homologous recombination are the only survivors. Once produced, these cells can be cloned and used to generate a complete mouse lacking that same gene (see Fig. A-25).



and causes a frameshift in the remaining coding sequence. The modified form of the target gene is linked to a drug-resistance gene, such as NEO, which confers resistance to neomycin. Only those ES cells that contain the transgene are able to grow in medium containing the antibiotic. The NEO gene is placed downstream of the modified target gene, but upstream of a flanking region of homology with the chromosome such that double recombination with the chromosome will result in the replacement of the target gene with the mutant gene and the drug resistance gene. (Alternatively, the NEO gene can be inserted into the target gene.)

There is, however, a high incidence of nonhomologous recombination in which recombination occurs illicitly at sites other than the endogenous gene. To enrich for homologous recombination events, the recombinant vector also contains a marker—the gene for the enzyme thymidine kinase (TK)—that can be subjected to counter selection by use of the drug gancyclovir, which is converted into a toxic compound by the kinase. The thymidine kinase gene is carried outside the region of homology with the chromosome in the vector. Hence, transformants in which the mutant gene has been incorporated into the chromosome by homologous

recombination will shed the thymidine kinase gene, but transformants in which incorporation into the chromosome occurred by illicit recombination will frequently contain the entire vector with the thymidine kinase gene and hence can be selected against.

As a result of this procedure, recombinant ES cells are obtained in which one copy of the target gene corresponds to the mutant allele. These recombinant ES cells are harvested and injected into the ICM of normal blastocysts. The hybrid embryos are inserted into the oviduct of a host mouse and allowed to develop to term. Some of the adults that arise from the hybrid embryos possess a transformed germ line and therefore produce haploid gametes containing the mutant form of the target gene. The ES cells that were used for the original transformation and homologous recombination assays give rise to both somatic tissues and the germ line. Once mice are produced that contain transformed germ cells, matings among siblings are performed to obtain homozygous mutants. Sometimes these mutants must be analyzed as embryos because of lethality. With other genes, the mutant embryos develop into full-grown mice, which are then examined using a variety of techniques.

### Mice Exhibit Epigenetic Inheritance

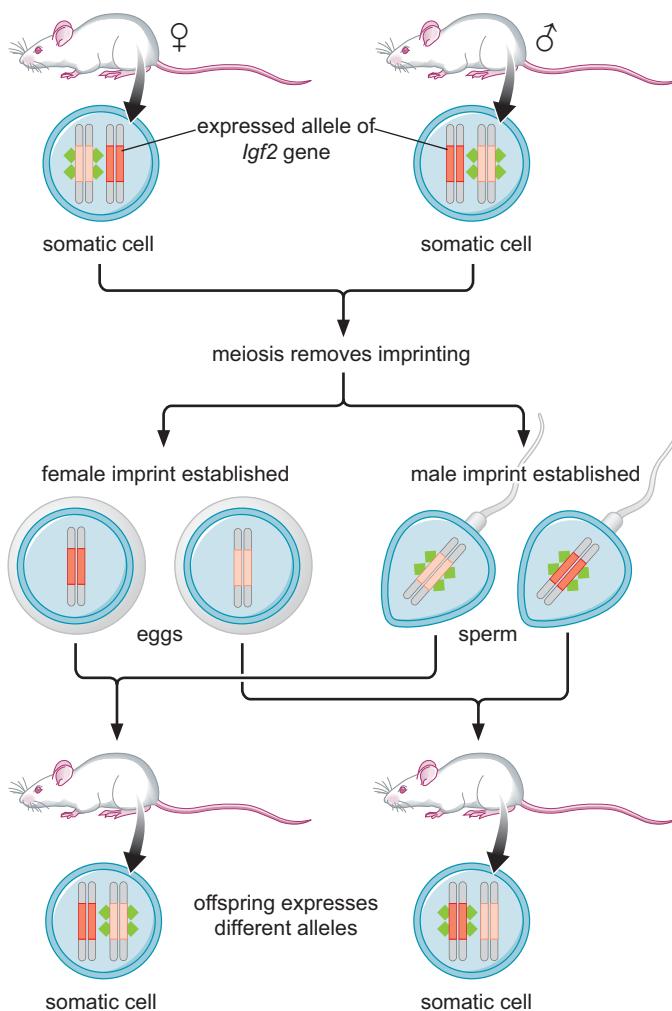
Studies on manipulated mouse embryos led to the discovery of a very peculiar mechanism of non-Mendelian, or epigenetic, inheritance. This phenomenon is known as **parental imprinting** (Fig. A-28). The basic idea is that only one of the two alleles for certain genes is active. This is because the other copy is selectively inactivated either in the developing sperm cell or the developing egg. Consider the case of the *Igf2* gene. It encodes an insulin-like growth factor that is expressed in the gut and liver of developing fetuses. Only the *Igf2* allele inherited from the father is actively expressed in the embryo. The other copy, although perfectly normal in sequence, is inactive. The differential activities of the maternal and paternal copies of the *Igf2* gene arise from the methylation of an associated silencer DNA that represses *Igf2* expression. During spermiogenesis, the DNA is methylated, and as a result, the *Igf2* gene can be activated in the developing fetus. The methylation inactivates the silencer. In contrast, the silencer DNA is not methylated in the developing oocyte. Hence, the *Igf2* allele inherited from the female is silent. In other words, the paternal copy of the gene is “imprinted”—in this case, methylated—for future expression in the embryo. This specific example is discussed in greater detail in Chapter 19.

There are approximately 30 imprinted genes in mice and humans. Many of the genes, including the preceding example of *Igf2*, control the growth of the developing fetus. It has been suggested that imprinting has evolved to protect the mother from her own fetus. The *Igf2* protein promotes the growth of the fetus. The mother attempts to limit this growth by inactivating the maternal copy of the gene.

We have considered how every organism must maintain and duplicate its DNA to survive, adapt, and propagate. The overall strategies for achieving these basic biological goals are similar in the vast majority of organisms and, therefore, may be examined rather successfully using simple organisms. It is, however, clear that the more intricate processes found in higher organisms, such as differentiation and development, require more complicated systems for regulating gene expression and that these can be studied only in more complex organisms. We have seen that a wide range of powerful experimental techniques can be used with success to manipulate the mouse and to explore various complex biological problems. As a result, the mouse has served as an excellent model system for studying devel-

**FIGURE A-28 Imprinting in the mouse.**

The permanent silencing of one allele of a given gene in a mouse. As outlined in the text, and described in detail in Chapter 19, imprinting ensures that only one copy of the mouse *Igf2* gene is expressed in each cell. It is always the copy carried on the paternal chromosome that is expressed.

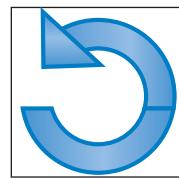


opmental, genetic, and biochemical processes that are likely to occur in more highly evolved mammals. The recent publication and annotation of the mouse genome has underscored the importance of the mouse as a model for further exploring and understanding problems in human development and disease.

## BIBLIOGRAPHY

- Burke D., Dawson D., and Stearns T. 2000. *Methods in yeast genetics*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
- Hartwell L.H., Hood L., Goldberg M.L., Reynolds A.E., Silver L.S., and Veres R.C. 2004. *Genetics: From genes to genomes*, 2nd ed. McGraw-Hill, New York.
- Miller J.H. 1972. *Experiments in molecular genetics*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
- Nagy A., Gertsenstein M., Vintersten K., and Behringer R. 2003. *Manipulating the mouse embryo*, 3rd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
- Sambrook J. and Russell D.W. 2001. *Molecular cloning: A laboratory manual*, 3rd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
- Snustad D.P. and Simmons M.J. 2002. *Principles of genetics*, 3rd ed. John Wiley and Sons, New York.
- Stent G.S. and Calendar R. 1978. *Molecular genetics: An introductory narrative*. W.H. Freeman and Co., San Francisco.
- Sullivan W., Ashburner M., and Hawley R.S. 2000. *Drosophila protocols*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
- Wolpert L., Beddington R., Lawrence P., Meyerowitz E., Smith J., and Jessell T.M. 2002. *Principles of development*, 2nd ed. Oxford University Press, Oxford.

## APPENDIX 2



# Answers

## CHAPTER 1

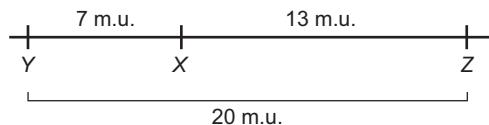
**Question 2.** False. A gene may have more than two alleles. One can determine the relationships among all of the alleles (e.g., by genetic crosses and mapping or sequencing).

**Question 4.** False. Some alleles display incomplete dominance as with the alleles of snapdragon. The F<sub>1</sub> generation shows an intermediate phenotype when true-breeding red and white snapdragons cross. A gene may have more than one allele, and the relationship between each allele relative to another has to be determined.

### Question 6.

- A. All pea plants with yellow seeds. All plants should be heterozygous (Yy) for the seed color gene. If yellow seeds are dominant, only the yellow seed phenotype will be observed.
- B. 3 yellow seeds: 1 green seed.
- C. 1 YY: 2 Yy: 1 yy.
- D. 2 heterozygotes (Yy): 2 homozygotes (YY and yy) or simplified 1:1.

### Question 8.



**Question 10.** From this information, you can say that *L* and *M* are linked and separated by 5 m.u. on one chromosome. A recombination frequency of 50% indicates that the genes are unlinked—

either far apart on the same chromosome or on different chromosomes. So gene *N* is >50 m.u. away from *L* and *M* or on a different chromosome.

**Question 12.** A mutation is a change in the sequence of DNA that is heritable. A low mutation rate allows organisms to adapt to changes in their environment over time. Mutations that enhanced the organism's ability to survive in a new environment would be passed on to a new generation.

**Question 14.** The value 15 must be the lowest value for observed recombinant progeny since a double crossover is the least likely event. From the information given, you know *pk* is in between *y* and *tri*, so 15 represents the total observed crossover between both *y* and *pk* and *pk* and *tri*.

To calculate the percent recombination (or m.u.), you divide the number of observed recombinants specific to a crossover between two genes of interest by the total number of progeny. For the observed recombinants, you add the value for the double crossovers because the single crossover did occur in those progeny.

Distance between *y* and *pk* =  $((X+15)/1000) * 100 = 23.0\%$  or 23.0 m.u.

*X*=215=the expected observed value for the total number of recombinant progeny representing a crossover between *y* and *pk*.

Distance between *pk* and *tri* =  $((X+15)/1000) * 100 = 18.4\%$  or 18.4 m.u.

*X*=169=the expected observed value for the total number of recombinant progeny representing a crossover between *pk* and *tri*.

## CHAPTER 2

**Question 2.** The backbone of DNA consists of sugar–phosphate groups. Therefore, the backbone is labeled with <sup>32</sup>P. Certain amino acids contain sulfur (cysteine and methionine). There-

fore, the protein is labeled with <sup>35</sup>S. The reverse labeling is not possible. There is no sulfur in DNA and no phosphorus in amino acids.

**Question 4.** The bases include cytosine, thymine, adenine, and guanine in DNA. The nitrogenous bases have a pyrimidine (cytosine and thymine) or purine (adenine and guanine) ring structure. Nucleosides include a nitrogenous base bound to a sugar (ribose in RNA, deoxyribose in DNA). Nucleotides include a nitrogenous base bound to a ribose or deoxyribose, which is bound to one, two, or three phosphate groups (mono-, di-, or triphosphate).

#### Question 6.

- Following one round of replication under the dispersive model, all DNA would contain half  $^{15}\text{N}$  and half  $^{14}\text{N}$ . So the resulting band after ultracentrifugation in a cesium chloride gradient would correspond to an intermediate band.
- Following one round of replication under the conservative model, half of the double-stranded DNA molecules would contain  $^{15}\text{N}$  and the other half of the double-stranded DNA molecules would contain  $^{14}\text{N}$ . So the resulting bands after ultracentrifugation in a cesium chloride gradient would correspond to one heavy band and one light band.
- The conservative model of replication can be eliminated in one round of bacterial replication, but two rounds of replication are required to distinguish between the dispersive and the semiconservative models. One round of replication under the conservative model would result in two double-stranded products, one labeled entirely with  $^{15}\text{N}$ , the other entirely with  $^{14}\text{N}$ , showing as two discrete bands (HH and LL) after centrifugation, which is not observed. But one round of replication, under either the dispersive or the semiconservative model, would produce a single band in the gradient (HL). A second round of replication, according to the semiconservative model, would produce two discrete bands (LL and HL), which are in fact observed; whereas the distributive model would predict only further intermediate bands, with each strand containing a mixture of (mostly) light and some heavy sequences. Thus, the semiconservative model is confirmed in the second round of replication.

**Question 8.** RNA (but not DNA) is located in the cytoplasm where protein synthesis occurs. The chemical structure of RNA is

similar to DNA (ribose instead of deoxyribose, uracil instead of thymine). RNA is synthesized from a DNA template.

**Question 10.** Polyribosome describes a group of ribosomes translating the same mRNA at the same time. Through polyribosomes, the translation of a specific protein is increased and the time to reach a certain level of that protein is decreased. This allows one mRNA to act as the template for multiple copies of the protein that can be made at the same time. This is useful because mRNA may not be at high levels or may have a short half-life (short-lived).

#### Question 12.



DNA polymerase is responsible for the duplication of DNA or DNA synthesis in the nucleus. RNA polymerase transcribes the DNA to mRNA in the nucleus. The ribosome is responsible for translation of RNA to protein in the cytoplasm. mRNA is the product in transcription and the template for protein synthesis (translation). tRNA acts as the adaptors during translation by reading the template and bringing in the appropriate amino acid. rRNA acts as a structural component of the ribosome as well as the catalytic component for peptide bond formation in translation.

#### Question 14.

- Protein synthesis is inhibited by the addition of DNase to about half of the level in the absence of DNase. The addition of more DNase does not increase the effect.
- Although DNA is not the direct template for protein synthesis, the two processes are connected through mRNA in the Central Dogma. In the presence of DNase, the DNA is destroyed. Without DNA, RNA polymerase has no template to make new mRNAs. mRNAs serve as the template for protein synthesis. mRNA can be short-lived, so the overall level of mRNA decreases indirectly as a result of the addition of DNase. Some mRNA must be present to see the observed level of protein synthesis in the presence of DNase.

## CHAPTER 3

**Question 2.** False. Enzymes lower the activation energy of a reaction. ( $\Delta G$  stays the same.)

**Question 4.** False. At  $25^\circ\text{C}$ , a 10-fold change in  $K_{\text{eq}}$  corresponds to about a 1.4-fold change in  $\Delta G$ .

#### Question 6.

- Hydrogen bonds between the bases in DNA.
- Covalent bond (peptide bond).

**Question 8.** Polar molecules have a dipole moment, whereas nonpolar molecules do not have a dipole moment. van der Waals interactions can include polar *and* nonpolar molecules.

#### Question 10.

$$K_{\text{eq}} = \frac{[\text{A}][\text{B}]}{[\text{AB}]} = \frac{[\text{A}] \times 2 \text{ mM}}{0.5 \text{ mM}} = 8.0 \times 10^5 \text{ mM},$$

$$[\text{A}] = 2.0 \times 10^5 \text{ mM.}$$

**Question 12.** Yes.  $\text{ATP} + \text{H}_2\text{O} \rightleftharpoons \text{ADP} + \text{P}_i \quad \Delta G = -7 \text{ kcal/mol}$  (Table 3-5).

Coupling the reaction to ATP hydrolysis gives an overall negative  $\Delta G$  value.

Overall reaction:  $\text{Glutamate} + \text{NH}_3 + \text{ATP} \rightleftharpoons \text{glutamine} + \text{ADP} + \text{P}_i \quad \Delta G = -3.6 \text{ kcal/mol.}$

#### Question 14.

- The side chain of tryptophan does not include hydrogen bond donor or acceptor.
- The side chain of glutamate includes a carboxylic acid capable of participating in hydrogen bonds.
- Although the numbers are small, there is a trend of arginine hydrogen bonding to guanine in the protein–DNA complexes.

## CHAPTER 4

---

### Question 2.

- A. 10 base pairs per helical turn  $\times$  4 helical turns =  $\sim$ 40 base pairs in 4 helical turns.

In solution: 10.5 base pairs per helical turn  $\times$  4 helical turns =  $\sim$ 42 base pairs in 4 helical turns.

Length = 3.4 nm per helical turn  $\times$  4 helical turns = 13.6 nm.

- B. 11 base pairs/helical turn  $\times$  4 helical turns =  $\sim$ 44 base pairs in 4 helical turns.

### Question 4.

#### G:C versus C:G major groove

The chemical group pattern for the major groove of a G:C base pair is AADH, whereas the pattern for the major groove of a C:G base pair is HDAA. The minor groove pattern is the same (ADA) for both base pairs.

A, hydrogen bond acceptor; D, hydrogen bond donor; H, non-polar hydrogens; M, methyl groups.

#### A:T versus G:C both

The chemical group pattern for the major groove of a A:T base pair is ADAM, whereas the pattern for the major groove of a G:C base pair is AADH. The pattern for the minor groove of a A:T base

pair is AHA, while the pattern for the minor groove of a G:C base pair is ADA.

#### A:T versus T:A major groove

The chemical group pattern for the major groove of a A:T base pair is ADAM, whereas the pattern for the major groove of a T:A base pair is MADA. The minor groove pattern is the same (AHA) for both base pairs.

### Question 6.

- Ai. Phosphodiester bond. (Mild treatment with DNase I will nick the DNA by cutting the sugar–phosphate backbone in DNA.)

- Aii. The DNA is originally supercoiled. After DNase I treatment, the DNA is no longer a cccDNA and becomes relaxed.

- B. Treatment with DNA ligase reseals the nick allowing for the formation of cccDNA again. The value of  $Lk^0$  is not an integer ( $10,000/10.5$ ). Lk, which is always an integer, does not equal  $Lk^0$ . Therefore, there must be some slight supercoiling.

### Question 8. N A Z X E

Diffraction pattern lines are perpendicular to lines in the letter.

## CHAPTER 5

---

**Question 2.** The most straightforward method to determine if the genetic material is DNA or RNA is to look for the presence of uracil in the sequence. If present, the virus contains RNA as the genetic material. If not, the genetic material is DNA. To determine if the genetic material is single-stranded or double-stranded, examine the percentages of each base. If the percentage of G equals the percentage of C and percentage of A equals the percentage of T, then the molecule is double-stranded DNA. For double-stranded RNA, the percentage of G=C and percentage of A=U. If the percentages do not show either pattern, then the genetic material is likely single-stranded.

**Question 4.** Uracil differs from thymine by the absence of a methyl group at the fifth carbon as in thymine. Both uracil and thymine base-pair with adenine. In DNA, spontaneous deamination of cytosine commonly occurs resulting in uracil. If replication takes place and uracil remains, a mutation occurs in the DNA (C:G to T:A). If uracil is naturally found in DNA, the uracil will not be recognized by DNA repair proteins as incorrect and mutations will occur.

**Question 6.** Secondary structure: stem, hairpins (stem and loop), and internal loop. Noncanonical base pairs: G:U and U:U.

**Question 8.** An RNA classified as a true ribozyme possesses a binding site for a specific substrate, a binding site for a cofactor, an active site for catalysis, and promotion of a reaction more than once per active site similar to a protein classified as an enzyme.

**Question 10.** The hammerhead must be divided into two separate parts. One part is capable of completing catalysis, whereas the other section of the RNA is the substrate. Because the substrate is not attached to the catalytic portion of the hammerhead, the substrate can be released to allow for a new molecule to bind. The hammerhead is now a true ribozyme, capable of completing many rounds of the reaction.

### Question 12.

- A. Adenosine.

- B. This nucleoside analog is not subject to hydrolysis in a strand of RNA like the other nucleosides. A methyl group is added to the 2'-hydroxyl of the ribose present in adenosine. When the methyl group is present, the oxygen can no longer become deprotonated at high pH and attack the scissile phosphate at the 3' position of the ribose in the RNA strand.

### Question 14.

- A. RNA is the catalyst or ribozyme. The protein moiety or RNA moiety alone do not cleave the substrate (reactions 2 and 3), but the reaction proceeds in the presence of protein and RNA. Because the substrate is cleaved (reaction 6) in the presence of the RNA moiety and spermidine (a nonspecific peptide), the RNA must be able to act as the catalyst. The protein moiety and spermidine are not able to complete catalysis without the RNA moiety (reaction 5). The protein moiety and spermidine help the RNA complete catalysis.

- B. The positively charged spermidine helps shield the repulsion between the negatively charged RNA catalyst and the negatively charged RNA substrate during the reaction.

## CHAPTER 6

---

**Question 2.** Ionic bonds form between oppositely charged groups. An ionic bond can form between the side chain of an acidic amino acid (aspartic acid or glutamic acid) and a basic amino acid (lysine, arginine, or histidine).

**Question 4.** In peptide bond formation, the carboxyl group of one amino acid covalently bonds with the amino group of another amino acid through the elimination of water. Because two molecules form a bond with the loss of water, the reaction is called a condensation or dehydration reaction (specific to the loss of water).

**Question 6.** Quaternary structure. Monomeric myoglobin has no quaternary structure; tetrameric hemoglobin has quaternary structure that is critical for its physiological function. Both proteins are globular, and their folded subunits are largely  $\alpha$ -helical; their secondary and tertiary structures are similar. Primary structure dictates secondary and tertiary structure; the primary structures of myoglobin and hemoglobin are therefore likely to be similar. (Note that even polypeptide chains with very dissimilar primary structures can have similar secondary and tertiary structures [e.g., plant hemoglobin and human hemoglobin]; but protein domains with related sequences always have the same folded structures, provided that the sequence similarity extends throughout the domain.)

**Question 8.**

- A. Disrupted, noncovalent bond.
- B. Disrupted, noncovalent bond.
- C. Not disrupted, covalent bond.
- D. Not disrupted, covalent bond. A  $\beta$  strand is a single unit of secondary structure; a  $\beta$  sandwich is an example of protein tertiary structure (a particular kind of folded domain).
- E. Disrupted, noncovalent bond.

**Question 10.** The two histidines and two cysteines are critical for coordination of the  $Zn^{2+}$ , which is in turn a critical stabilizing element for the very small, Zn-finger domain. Substituting alanine for any one of these four residues will eliminate  $Zn^{2+}$  binding and destabilize the domain, leading to loss of function.

**Question 12.** Enzymes catalyze (i.e., enhance the rate of) a reaction by lowering the energy needed to form the transition state and thereby lowering the energy barrier between reactants and products. Enzymes appear to do so in many cases because their active sites are complementary to the transition-state conformations of the reactants, rather than to the ground-state conformations—that is, there are favorable noncovalent interactions between the enzyme and the transition-state forms of its substrates.

**Question 14.**

- A. RRM (i.e., the sequence motifs characteristic of a type of RNA binding domain) are sequences of 80–90 amino acid residues that fold into a domain that recognizes a specific RNA. (Note that the term “RRM” is often misused to designate the domain; see Box 6-2, Glossary, for correct usage of the words “motif” and “domain.”)
- B. The data show that the amino-terminal RRM and the first three repeats are sufficient for full complementation. Neither the RRM plus one repeat nor the seven repeats plus the carboxy-terminal region is adequate to confer wild-type growth at 37°C.
- C. The results agree with the complementation assay, but they suggest that, *in vitro*, three repeats are not sufficient for full activity. Moreover, the form with RRM truncated, but the seven repeats and carboxy-terminal region intact, had no activity *in vitro*, but restored partial growth *in vivo*. These discrepancies suggest some redundancy of function, either among the Tif3 domains or with other components of the translation initiation complex.

## CHAPTER 7

---

**Question 2.**

- A. For a restriction enzyme that recognizes a 6-bp sequence, the frequency of finding that sequence in a given genome is 1 in  $4^6$  or 1 in 4096 bp.
- B. Yes. Even though the recognition sequences differ for *Xba*I and *Sal*I, the sticky ends can base-pair with each other because the single-stranded regions are complementary to each other.

**Question 4.** When performing a Southern blot, you detect a specific DNA sequence with a DNA probe. When performing a northern blot, you detect a specific mRNA sequence with a DNA probe. When performing a Southern blot, you digest the genomic DNA with a restriction enzyme, separate the DNA fragments by gel electrophoresis, transfer the DNA to a positively charged membrane, and detect a fragment of DNA that contains your DNA of interest with the probe. You perform a similar set of steps for a northern blot except that you do not digest the mRNA population. In a northern blot, you can detect the amount of a certain type of mRNA and can compare that to another sample produced under different experimental conditions.

**Question 6.** A genomic library consists of the complete set of DNA fragments, generated by restriction endonuclease digestion of the entire genome. A cDNA library, which consists only of expressed sequences in genomic DNA, is generated by reverse transcription of all mRNA in the cell. In each case, the resulting fragments of DNA are ligated into plasmid vectors. The human genome includes a large proportion of non-coding DNA including sequences that code for introns that are spliced out of the mRNA. cDNA libraries are useful for studying and expressing these gene-encoding sequences.

**Question 8.** Ion-exchange chromatography separates proteins based on charge. Gel-filtration chromatography separates proteins based on size. Affinity chromatography separates proteins based on interaction with a specific molecule, protein, or nucleic acid that is coupled to the beads.

**Question 10.**

- A. Only one end (strand) of the DNA is labeled so that nuclease digestion of the bound DNA fragment will produce, after gel

electrophoresis, a visible ladder of fragments extending from a single labeled end. Digestion of a strand labeled at both ends would complicate the pattern and obscure the “footprint.” And, if the protein binds asymmetrically, the pattern becomes even more complicated.

- B.** To determine if a specific known region is bound to the protein, use primers that are specific to those sequences to amplify that sequence and compare the results to necessary controls. Another option is to use a tiling DNA microarray to identify many different sequences.

**Question 12.**

- A.** The western blot detected one band for Protein Z in the embryo and adult lanes. The northern blot indicates that

there are two transcripts for Gene Z in embryonic flies but only one transcript in adult flies. One possible hypothesis is that the antibody used in the western blot does not recognize the form of the protein translated by the faster migrating transcript because it may be missing the coding region for the carboxy-terminal domain as a result of alternative splicing of RNA.

- B.** For the hypothesis given, you could use a new antibody against Protein Z in the western blot. This antibody could be polyclonal to the whole protein or be monoclonal against a central or amino-terminal region of the protein. If you see a second band on the western blot in the embryo lane when using the new antibody, the data would support the hypothesis proposed in part A.

## CHAPTER 8

---

**Question 2.** The chromosomal DNA is located within the nucleoid in prokaryotic cells. The chromosomal DNA is located in the nucleus for eukaryotic cells. The nucleus, unlike the nucleoid, is membrane-bound and typically occupies a small fraction of the cell volume.

**Question 4.** The intergenic sequences may have arisen from transposition events. They may encode miRNAs, may serve as regulatory sequences for transcription, or may simply be non-functional sequences such as pseudogenes.

**Question 6.** Cohesion holds sister chromatids together during S phase and early stages of mitosis. During late mitosis (anaphase), cohesion is eliminated so that the microtubules attached to kinetochores that assemble at the centromere separate sister chromatid pairs into the daughter cells.

**Question 8.** All cells that grow and divide (somatic cells and germ cells) use mitosis. Only cells (germ cells) that produce egg and sperm cells go through meiosis.

**Question 10.** Hydrogen bonds form primarily between the histone proteins and the phosphodiester backbone near the minor groove and additionally between the bases of the minor groove.

These interactions are not sequence-specific. DNA throughout the genome wraps around the histone proteins. Proteins that interact with the minor groove of DNA are much less likely to interact in a sequence-specific manner. In contrast, interactions with the major groove of the DNA generally make sequence-specific interactions (Chapter 4).

**Question 12.** Bromodomain recognizes acetylation. Chromodomains, TUDOR-domains, and PHD fingers recognize methylation. (SANT domains recognize unmodified histone tails.)

**Question 14.**

- A.** The histone deacetylase binds nucleosome bound-DNA (lanes 1, 2, 3, and 4 compared to lane 5). Assuming the histone deacetylase is a monomer, two deacetylases are capable of binding the nucleosome-bound DNA at the same time (two higher migrating bands in lanes 2, 3, and 4). The histone deacetylase seems to recognize nucleosomes that are methylated at lysine 36 of histone H3 better than unmethylated nucleosomes (lanes 1 and 2 vs. 3 and 4).
- B.** Based on this data, the histone deacetylase likely includes a chromodomain to interact with methylated histone H3.

## CHAPTER 9

---

**Question 2.** The basic mechanism of DNA synthesis begins with the hydrogen bond-dependent interaction of the incoming nucleotide to the DNA template. After an appropriate base pair is formed, the 3'-OH of the primer initiates a nucleophilic attack of the  $\alpha$ -phosphate of the incoming nucleotide. Pyrophosphate is released and hydrolyzed to two phosphatases by pyrophosphatase. The incoming nucleotide is now base-paired to the template and covalently linked to the primer DNA strand.

**Question 4.**

- A.** Deoxyguanosine.
- B.** Without the triphosphate group, Acyclovir cannot incorporate into a growing strand of DNA. Kinases phosphorylate

their substrate. The kinase adds the phosphate groups that Acyclovir is missing.

**Question 6.** Some DNA polymerases are only used during special processes like DNA repair. They tend to not be very processive and do not carry out the bulk of DNA synthesis in the cell. Therefore, proofreading is less important for these DNA polymerases that will insert a small number of nucleotides relative to the leading- or lagging-strand DNA polymerases.

**Question 8.**

- A.** Both. Replication will occur in both directions.

- B.** Bottom. The bottom strand serves as the leading-strand template on the right side. Extension of the 3' end of the RNA primer annealed to this strand by DNA polymerase is able to replicate continuously to the end of the template.
- C.** Bottom. DNA ligase is required to create phosphodiester bonds between Okazaki fragments on the lagging strand during DNA synthesis. The bottom strand serves as the template for the lagging strand on the left side, because the DNA synthesis has to be discontinuous.

#### Question 10.

- A.** *E. coli* cells only initiate replication once per cell division, but, when *E. coli* divides rapidly, initiation of the next round of replication starts before the previous round of replication is complete. Under these conditions, the time for cell division can be as low as 20 min.
- B.** The circular *E. coli* genome has no ends like linear chromosomes. Under these conditions, *E. coli* cells do not have the problem of the chromosome length shortening after each round of replication in the absence of telomerase, because the replication machinery can completely replicate the circular genome.

**Question 12.** The 3'-exonuclease activity of each DNA polymerase provides the DNA polymerase with the ability to remove incorrect nucleotides during DNA synthesis. DNA polymerase I

has the additional 5'-exonuclease activity to remove nucleotides ahead of the DNA polymerase. Specifically, this function helps the DNA polymerase remove RNA primers on the lagging strand of DNA.

#### Question 14.

- A.** The  $\alpha$ -phosphate is incorporated into the newly synthesized DNA strand through the nucleophilic attack by the 3'-OH. The  $\beta$ - or  $\gamma$ -phosphates become pyrophosphate, which is later hydrolyzed and never incorporated into the growing strand of DNA.
- B.** Gel electrophoresis separates molecules by size. The  $^{32}\text{P}$ -labeled dNTPs are much smaller than the newly synthesized DNA and migrate much faster than any long strand of DNA.
- C.** One example of a negative control is to run the same DNA synthesis assay but in the absence of DNA polymerase. Without new DNA synthesis, the primer:template junction will not be labeled. If you properly filter the reaction containing the primer:template junction and  $^{32}\text{P}$ -labeled dNTPs over a positively charged membrane, you should find that the radioactivity does not stick to the filter. You can compare this to the same reaction containing the DNA polymerase. If you are worried about any possible effects from the  $^{32}\text{P}$ -labeled dNTPs binding to the DNA polymerase, you could protease-treat the reaction before filtering.

## CHAPTER 10

**Question 2.** Deamination of cytosine produces uracil. A specific glycosylase in base excision repair recognizes uracil as not belonging in DNA. If uracil remains, a mutation could occur after the next round of replication. Deamination of 5-methylcytosine produces thymine, which is not recognized as a mistake by a DNA repair pathway. Following the next round of replication, the thymine produced from the deamination of 5-methylcytosine pairs with adenine for a transition mutation. Therefore, the cell removes uracils to prevent mutations but does not remove the thymines produced from deamination.

#### Question 4.

Correct Order	MMR	BER	NER
Recognition	MutS (MutH determines the strand)	DNA glycosylase	UvrA
Excision	MutH (activated by MutL) and Exo VII, RecJ or Exo I	DNA glycosylase, AP endonuclease, and exonuclease	UvrC and UvrD (help from the UvrB-induced bubble)
DNA synthesis	DNA Pol III	DNA Pol I	DNA Pol I
Ligation	DNA ligase	DNA ligase	DNA ligase

**Question 6.** Without a properly functioning Dam methylase, the parent strand would not be methylated during replication. Without this methylation, MutH has no way to distinguish which strand is parental versus newly synthesized. MutH will nick the incorrect strand at some frequency. Mismatch repair of the parent strand would lead to an increase in spontaneous mutagenesis (not induced by an exogenous agent).

**Question 8.** A cross-link between two guanines distorts the DNA helix similar to a thymine dimer. This allows NER proteins to recognize the distortion to excise the stretch of DNA including the cisplatin-induced cross-link. Base excision repair only excises one nucleotide. Also, a specific DNA glycosylase must recognize the DNA lesion. No glycosylase recognizes cisplatin-induced cross-links.

**Question 10.** Nonhomologous end joining (NHEJ) repairs double-strand breaks (DSBs) at the cost of introducing mutations. The NHEJ enzymes process the free ends of a DSB. Through this processing, DNA sequence is lost or added before the two strands are ligated together.

#### Question 12.

Mutant Pathway	DNA Damage	Percent Survival	Mutagenesis
NER	Increase	Decrease	Increase
Translesion Synthesis	Stays the same	Decrease	Decrease

Relative to wild type, the amount of DNA damage increases for a NER mutant because the thymine dimers are not being repaired as efficiently. DNA damage tolerance through translesion synthesis does not repair the lesions, so the level of DNA damage remains the same, although loss of tolerance does lead to more cell death. This is also true for loss of NER. With more DNA damage in the NER mutant cells, more mutagenesis occurs. Less mutagenesis occurs if the translesion synthesis pathway is disrupted, because translesion synthesis polymerases contribute to mutagenesis normally.

#### Question 14.

A. The medium must be lacking histidine for selection. Only another point mutation in the specific location of original point mutation in the *HisG* gene can lead to a reversion.

This mutation must change the sequence back to the wild-type sequence of the gene, which allows the cells to grow in the absence of histidine.

- B. Free radicals in the cell can damage DNA, which can cause mutations that can lead to a reversion. Other common processes, like replication errors and hydrolytic attack of the bases, also alter the DNA.
- C. Chemical A. There are more revertants indicating a higher frequency of mutations induced by Chemical A (relative to survival) than the control (no chemical added).
- D. Chemical C. There are fewer revertants indicating a lower frequency of mutations induced by Chemical C (relative to survival) compared to the control (no chemical added).

## CHAPTER 11

---

**Question 2.** Alleles differ from each other by minor sequence variation. The majority of the sequence for the gene remains the same, so the alleles are homologous.

**Question 4.** The third step shows invasion of a 5' end. This is a problem because the DNA polymerase needs a 3'-OH at the primer-template junction for extension. To correct the problem, the third step should show invasion of a 3' end and base pairing with the appropriate blue strand.

**Question 6.** RecBCD includes DNA helicase and nuclease activities. Specifically, RecB acts as a 3' to 5' DNA helicase and a nuclease. RecD acts as 5' to 3' DNA helicase. RecC helps improve the efficiency of both RecB and RecD. RecC recognizes and binds to the  $\chi$  site to stop the nuclease activity on the 3' tail. RecBCD plays a critical role in processing the double-stranded DNA at a break to produce single-stranded DNA for invasion.

**Question 8.** The DNA substrate can be analyzed by using an electrophoretic mobility shift assay (EMSA), as follows. RuvA protein is incubated with end-labeled DNA substrate, and the products run on a nondenaturing gel. Two bands should be visible on the gel: the faster-migrating band is the DNA substrate only, and the slower migrating band corresponds to DNA bound to RuvA. One of the DNA strands in the junction (end-labeled with 5'-<sup>32</sup>P) can serve as a negative control in the experiment. Here, because RuvA will not bind to the single strand, only the faster-migrating band will be apparent.

**Question 10.** Spo11 mediates double-stranded cleavage of the DNA. A tyrosine in Spo11 attacks the phosphodiester backbone to cut the DNA. Spo11 stores the energy from breaking the phosphodiester bond by forming a covalent high-energy intermediate with the broken DNA.

**Question 12.** Like DSB-repair homologous recombination, SDSA starts with a DSB at the recombination site, 5' to 3' resection, and invasion of the 3' end to serve as a primer for DNA synthesis. SDSA differs from DSB-repair homologous recombination in that there is no resolution via cleavage of a Holliday junction. Following strand invasion, a complete replication fork forms in SDSA. The 3' end that does not invade is removed in SDSA. The newly synthesized DNA is displaced, and a second DNA synthesis event completes the process resulting in gene conversion.

#### Question 14.

- A. Lane 2 reveal that Protein X cleaves the Holliday junction-like DNA substrate. Because only one DNA strand is labeled, only one band is visible in the autoradiogram. As the full-length DNA substrate is 60 nt and the cleaved product is 31 nt long, the cleavage likely occurs in the center of the strand, near the 90° angle of the “junction.” Lane 3 reveals that RecA (as expected) does not cleave the DNA substrate.
- B. Protein X function resembles that of RuvC. In *E. coli*, RuvC cleaves the Holliday junction in the manner observed for Protein X.

## CHAPTER 12

---

**Question 2.** The recombinases store the energy of the broken phosphodiester bond through a covalent protein–DNA intermediate. The recombinase then reseals the broken DNA strands once a cleaved strand attacks the protein–DNA covalent bond.

**Question 4.** Both recombinases form covalent recombinase–DNA intermediates, use four strands of DNA from two duplexes, catalyze

reversible reactions, and include four subunits that recognize and bind four specific sites in the DNA. Neither recombinase requires external energy to catalyze these processes.

Serine recombinases cleave all four strands in the first step. Tyrosine recombinases cleave and rejoin only two DNA strands first, and then cleave and rejoin the remaining two strands. The first cleavage and rejoining event produces a Holliday junction not formed during the serine recombinase mechanism. The

mechanism for serine recombinases includes a 180° rotation of the dimers in the protein–DNA complex. This type of protein–DNA rotation does not occur in the mechanism for tyrosine recombinases.

**Question 6.**  $\lambda$ Int functions in both integrative and excisive recombination.  $\lambda$ Int is a tyrosine recombinase that catalyzes recombination through the *att* sites. The  $\lambda$ Int also catalyzes the excision with critical assistance from Xis and IHF.

**Question 8.** DNA transposons remain DNA throughout the cycle, whereas retrotransposons include an RNA intermediate in their propagation cycle.

**Question 10.** The insertion of transposons such as Tn5 into the host genome is generally not sequence-specific. Through the cut-and-paste mechanism, Tn5 nearly randomly inserts into the host genome of interest. Depending on the location of insertion, this process may disrupt the coding sequence of a gene or upstream DNA element important for gene expression. A scientist can screen cells for a particular phenotype and identify the interrupted gene because the sequence of the transposon is known. In a screen using chemical mutagenesis, it is much harder to identify the small point mutations that are usually made.

**Question 12.** In the cut-and-paste mechanism, the transpososome excises from its original genomic location. The transposon DNA attacks and inserts through a DNA strand transfer mech-

anism into another location in the DNA. In the replicative mechanism, the transpososome breaks one strand on each side of the transposon. The broken strands attack and become joined to the target DNA site elsewhere in the genome forming a doubly branched structure. After replication, the result of processing of this branched intermediate is a circular co-integrate that includes two transposons.

#### Question 14.

- A. The mother's SstI digest does not show a band corresponding to a 5.5-kb fragment but does include a band corresponding to a 3.2-kb fragment. Only fragments including exon 14 differ. All of the other fragments seem to be the same size.
- B. The patient's KpnI digest does not show a band corresponding to a 7.3-kb fragment but does include bands corresponding to a 5.3- and a 4.3-kb fragment. Only fragments including exon 14 differ. All of the other fragments seem to be the same size.
- C. Because only fragments including exon 14 differ between the patient and mother digests, the transposon likely inserted into exon 14. Based on the SstI digest, the patient had one fragment 2.3 kb larger than the mother's 3.2-kb fragment. Based on the KpnI digest, the patient has two smaller fragments instead of the mother's 7.3-kb fragment, but the sum of the smaller fragments is 9.6 kb. These observations could be explained by a transposon insertion (exon 14), and the transposon DNA itself includes a KpnI recognition site.

## CHAPTER 13

---

**Question 2.** The RNA polymerase initially binds the promoter sequence. Once bound, structural changes occur in the promoter–RNA polymerase complex to initiate transcription. To prevent transcription or enhance transcription of a specific gene, it is most direct to inhibit or enhance initiation through the promoter.

**Question 4.** A DNA footprinting assay is the best choice. This assay will show a footprint around those positions if RNA polymerase binds the promoter. An EMSA will also test protein binding to DNA. You could see a result that RNA polymerase binds the promoter with an EMSA, but you will not know that the binding is centered on the –35 and –10 sites. ChIP is also possible, but DNA footprinting is more relevant when considering only a specific DNA sequence.

**Question 6.** Experimental results support the scrunching model. These experiments indicate that RNA polymerase remains bound to the promoter, while DNA is unwound and pulled in to the RNA polymerase during initial transcription.

**Question 8.** The bolded nucleotides are within regions that base-pair with each other to form the termination hairpin. If one of these nucleotides is mutated, then the base pairing is disrupted, preventing the formation of the hairpin and disrupting termination. To test this model, you could mutate one of the nucleotides, as just described, to disrupt termination. Then you could mutate the nucleotide in the other arm of the stem loop in a manner that

would re-establish base-pairing with the initial mutation. The double mutant will restore termination.

**Question 10.** Phosphorylation of serine residues in the CTD tail of Pol II is required for promoter escape and for efficient elongation. In addition, different patterns of phosphorylation allow the tail to recruit factors required for RNA processing as well. Thus, regulation of tail phosphorylation ensures these events are coordinated appropriately.

**Question 12.** Poly-A polymerase does not require a DNA template and adds up to 200 As to the 3' end of mRNAs. RNA polymerase requires a DNA template and incorporates all four NTPs for RNA.

#### Question 14.

- A. The input reaction includes all of the DNA (chromatin) in the cells. All regions of the DNA should be present at equal levels in the input. PCR amplification using any primer set worked.
- B. Prominent bond from the PCR reaction using primers specific for amplification of the DNA just 3' of the sequence encoding the poly-A signal sequence (reaction in lane 3, upper band). This shows that the Rat1 must localize at this region of the *ADH1* gene.
- C. In the torpedo model, Rat1 degrades (in the 5'-to-3' direction) the RNA transcribed downstream from the poly-A site. This eventually displaces RNA polymerase from the DNA. The

results from this experiment are consistent with the model in that they show Rat1 is associated with the transcription machinery predominantly at the 3' end of the gene, at the

location one would predict if it joins the cleaved transcript immediately after polyadenylation, as predicted in the model.

## CHAPTER 14

**Question 2.** The 5' and 3' splice sites are named with respect to the intron. The 5' splice site is located at the 5' end of the intron where it meets the 3' end of the upstream exon. The 3' splice site is found at the 3' end of the intron where it meets the 5' end of the other exon. The branchpoint A, located within the intron, is also required for the splicing reaction. A polypyrimidine tract follows the branchpoint site.

**Question 4.** In vivo, Prp22 rapidly removes the spliced mRNA from the spliceosome and disassembles the spliceosome. Also, the lariat RNA is quickly degraded.

**Question 6.** The U1 snRNP binds the 5'-GUAAGU-3' sequence with perfect complementarity to the U1 snRNA sequence 5'-ACUUAC-3'. The mutation will likely cause a decrease in binding of U1 to the 5' splice site but not completely disrupt that binding as five potential base pairs remain. The U6 snRNA (relevant sequence 5'-ACAGAG-3') can form three base pairs with the 5'-GUAAGU-3' sequence and can form four base pairs with the sequence 5'-GUAUGU-3'. This will likely enhance binding of the U6 snRNP with the 5' splice site.

**Question 8.** The final product will include a piece of the intron sequence retained between the two exons. During translation, this RNA will code for unintended amino acids or introduce a premature stop codon. This insertion of extra sequence in the mRNA could also cause a reading frameshift for codons downstream of

the inserted sequence. These changes will almost inevitably be detrimental to the protein.

**Question 10.** The nonsense-mediated decay pathway degrades mRNAs that contain a premature stop codon. This affords the cell one mechanism to ensure alternative splicing in one class of gene. Thus, in a given gene, of two alternative exons, only one or the other—never both—is included in the final spliced mRNA. This works because although each of the alternative exons has an open reading frame that fits into the rest of the message, the sequence of the two exons results in a premature stop codon if both are included in the mature mRNA.

**Question 12.**

A: Spliceosome-requiring intron. In the presence of nuclear abstract, you see a band for the lariat and the spliced product. In the absence of nuclear abstract, you see only the pre-mRNA.

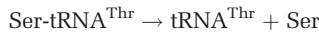
B: Group II intron. Splicing takes place in both the absence and presence of nuclear abstract, and so the reaction is self-splicing. The products migrate the same as in the spliceosome-requiring reaction, so it must be a group II intron.

C: Group I intron. Splicing takes place in the absence and presence of nuclear abstract, and so again the reaction is self-splicing. One product migrates faster than the lariat product in reactions A and B, so it must be linear, a key feature of a group I intron.

## CHAPTER 15

**Question 2.** The tRNA is coupled to the cognate amino acid at this 3' end. The high-energy acyl bond forms between the amino acid and the 3'-OH or 2'-OH of 5'-CCA-3'. All tRNAs end with this sequence.

**Question 4.**



**Question 6.** The alanyl-tRNA synthetase has an editing pocket that hydrolyzes Gly-tRNA<sup>Ala</sup>. The side chain of alanine (methyl group) is slightly larger than the hydrogen found on glycine. So the active site of the alanyl-tRNA synthetase can accommodate glycine and mischarge tRNA<sup>Ala</sup> with glycine. Alanyl-tRNA synthetase has an editing pocket that can fit (and therefore remove) glycine—but not alanine-coupled tRNA<sup>Ala</sup>.

**Question 8.** The simplest experiment is to treat the ribosome with a protease and ask if the resulting ribosome can still synthesize new protein. After such a treatment, it was found that, even after most of the protein was removed, peptide bond formation could

still occur. Structural studies further supported that the RNA could catalyze peptide bond formation because no amino acid is present within 18 Å of the active site.

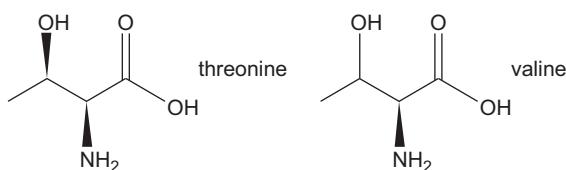
**Question 10.** The structures of each complex are very similar. A portion of the EF-G protein adopts a shape similar to the tRNA in the EF-Tu-GTP-tRNA complex. This helps explain how both bind to the A site of the ribosome.

**Question 12.** Antibiotics target and inhibit one of the steps of translation by binding to a specific position in the ribosome or EF-Tu or EF-G. Inhibition of one step of translation will stop all of the steps. Translation must function for the cell to survive. The exact protein and rRNA components of the ribosome differ in prokaryotic versus eukaryotic cells. Therefore, antibiotics specifically bind to a component found in prokaryotic ribosomes versus eukaryotic ribosomes.

**Question 14.**

A. The structure of threonine is very similar in size to valine and could fit in the active site of the ValRS. The threonine

has a hydroxyl group where valine only has a methyl group.



- B. The most Thr-tRNA<sup>Val</sup> is produced in the presence of the K270A and D279A mutants (changing the lysine at position

270 to alanine, changing the aspartic acid at position 279 to alanine). This information indicates a potential problem with editing.

- C. Each round of misincorporation and editing consumes one molecule of ATP. The ATP is consumed because one amino acid is hydrolyzed from the tRNA, and the new amino acid must be adenylated and transferred. If the ValRS mutant does not edit, the amount of Thr-tRNA<sup>Val</sup> would increase and the amount of ATP consumption would decrease. This is what we see for the K270A and D279A ValRS mutants.

## CHAPTER 16

**Question 2.** The most common mutation is a transition A:T to G:C or G:C to A:T. If the DNA encoding the middle nucleotide of the codon undergoes a transition mutation, lysine would replace arginine or vice versa. These two amino acids are positively charged. This amino acid substitution in a protein is more conservative than other options and gives the cell the best chance of not altering protein structure or function.

**Question 4.**

- |     |        |
|-----|--------|
| No  | A. UGC |
| Yes | B. CGA |
| No  | C. UGA |
| Yes | D. CGU |
| No  | E. GCG |

**Question 6.** Use the repeated dinucleotide sequence of GU. 5'-GUGUGUGUGUGU...-3' codes for a polypeptide with alternating valine (5'- GUG-3') and cysteine (5'- UGU-3').

**Question 8.** The coding strand has the same sequence as the mRNA except that mRNA has U instead of T.

Using the first frame (starting with the first nucleotide on the 5'-end), NH<sub>2</sub> – threonine – valine – serine – alanine – arginine – COOH.

Using the second frame (starting with the second nucleotide on the 5'-end), NH<sub>2</sub> – proline – phenylalanine – arginine – leucine – COOH.

Using the third frame (starting with the third nucleotide on the 5'-end), NH<sub>2</sub> – arginine – phenylalanine – glycine – (stop) COOH. Because this sequence is in the middle of a gene, it is unlikely to be the frame used.

**Question 10.**

- A. Insertion of 2 base pairs—frameshift mutation. The frameshift causes downstream codons to code for a different sequence of amino acids.
- B. Insertion of two base pairs from altered sequence 1, deletion of one base pair immediately before the inserted base pairs—remains a frameshift—shift the reading frame by one. The

frameshift causes downstream codons to code for a different sequence of amino acids.

- C. Insertion of two base pairs from altered sequence 1, deletion of two base pairs immediately before the inserted base pairs—eliminates the frameshift. So this is an example of an intragenic suppressor (within the gene itself) mutation that returns the amino acid sequence back to wild type even though the DNA sequence is not identical to wild type.

**Question 12.** Universality refers to the conservation of the genetic code between all organisms. For the most part, organisms use the same genetic code. There are variations from the standard genetic code for a few specific codons or amino acids. The mammalian mitochondria, *Candida albicans*, and *Mycoplasma capricolum* use a genetic code with exceptions.

**Question 14.**

- A. The suppressor mutation is an intergenic nonsense suppressor—intergenic because the mutation is in a different gene than the gene of interest.
- B. If a very commonly used tRNA<sup>Leu</sup> carries a suppressor mutation, then many genes in the cell will have trouble encoding leucines requiring that specific tRNA.
- C. The stop codon 5'-UAG-3' is recognized by the anticodon 5'-CUA-3'. 5'-UAG-3' is only one nucleotide different from the leucine codon 5'-UUG-3'. Therefore the sequence of the mutated anticodon is 5'-CUA-3', and the sequence of the wild-type anticodon is 5'-CAA-3'.
- D. The 5'-UAG-3' codon is infrequently used as a stop codon in the *E. coli* genome. Therefore, insertion of an amino acid instead of ending translation of a protein does not cause many problems for the other proteins in *E. coli*, which normally would cause cells to die or grow slowly. The insertion of an amino acid at each 5'-UAG-3' codon as a result of a suppressor mutation would mostly affect the translation of the mutant protein of interest.

For more information, see Thorbjarnardóttir et al. (1985. *J. Bacteriol.* **161**: 219–222).

## CHAPTER 17

**Question 2.** Despite the small genome in *Mycoplasma genitalium*, they do have cellular structure, undergo cell division, and do not rely on a host like viruses. Similarly, the symbiont *Hodgkinia cicadicola* depends on the host cells for survival and is not considered alive.

**Question 4.** The most prominent example is the RNA component in the large subunit of the ribosome that catalyzes the peptide bond. Other examples are possible. Because the primary reaction that takes place in protein synthesis requires RNA, the idea that RNA preceded protein in the RNA World hypothesis seemed more plausible. The catalytic RNA in the ribosome may be a molecular fossil to the RNA World.

**Question 6.**

- i. RNA polymerase.
- ii. Reverse transcriptase.
- iii. DNA polymerase.
- iv. RNA replicase, a ribozyme. (Note that some RNA-dependent RNA polymerases exist, too.)

**Question 8.** When several RNA replicases become membrane-bound, the chance of a less efficient replicase copying the mutant replicase increases. When that protocell eventually divides, there is a reasonable chance that two or more mutant replicases will be binned in one protocell. Over time, the propagation of protocells containing the more efficient, mutant

replicase will outcompete the protocells with the less efficient replicases.

**Question 10.** Scientists did generate pyrimidines using organic molecules possibly present on primitive Earth as the starting materials. The experiments attempting to create a nucleotide from a phosphate, ribose, and nucleobase under prebiotic conditions did not work. So scientists do not favor this hypothesis.

**Question 12.** The RNA replicase ribozyme must be able to catalyze phosphodiester bond formation more than just one time. The replicase ribozyme also requires free ribonucleotides to use in catalysis. The replicase ribozyme sequence and its complement must also serve as the template. Replication of the replicase sequence produces a complement sequence. Replication of the complement sequence produces another copy of the RNA replicase.

**Question 14.**

- A. The ribozyme is able to synthesize longer RNA products in reactions 1 and 3 than in reaction 2.
- B. The change in sequence for the template in reaction 2 had a deleterious effect. The change in sequence for the template and ribozyme restored the catalytic ability of the ribozyme to wild-type levels. The ribozyme sequence 5'-UCAUUG-3' is complementary to the template sequence 5'-CAAUGA-3'. In reaction 3, the change in ribozyme and template sequence restores complementarity between the sequences. The ribozyme likely base-pairs with the template RNA during synthesis.

## CHAPTER 18

**Question 2.** Allolactose binds the Lac repressor in a region separate from its DNA-binding domain. Once bound, allolactose causes the Lac repressor to change shape and release from the DNA, stopping repression. Similarly, allosteric effectors exist to regulate the *araBAD* and *gal* operons. In the presence of low glucose levels, cAMP binds CAP to induce a change in shape of CAP that allows CAP to bind the DNA for activation. Through allostery, NtrC and MerR activate transcription of the *glnA* and *merT* genes, respectively. Additional answers are possible.

**Question 4.**

- A. Constitutive expression means that the genes in the *araBAD* operon are expressed in the presence or absence of arabinose. That is, regulation is lost, and the genes are expressed.
- B. A mutation in the gene encoding AraC could lead to constitutive expression of the *araBAD* operon. The mutation must prevent the AraC-induced DNA loop formed in the absence of arabinose. Also, elimination by mutation of the *araO<sub>2</sub>* site could lead to constitutive expression.

For more information, see Englesberg et al. (1965. *J. Bacteriol.* **90:** 946–957).

**Question 6.**

**Aa.** Basal level of expression. In the presence of glucose, CAP is not bound. The mutation prevents lac repressor from binding, and so there is basal level of lacZ expression.

**Ab.** Basal level of expression. In the presence of glucose, CAP is not bound. The repressor is not bound in the presence or absence of lactose in this mutant, allowing basal level of lacZ expression.

**Ac.** Activated level of expression. In the absence of glucose, CAP is bound. The repressor is not bound in the presence or absence of lactose in this mutant allowing an activated level of lacZ expression.

**Ad.** Activated level of expression. In the absence of glucose, CAP is bound. The repressor is not bound in the presence or absence of lactose for this mutant allowing an activated level of lacZ expression.

**Ba.** No expression. With a mutation in the promoter that prevents RNA polymerase from binding, the lacZ is never expressed even if the repressor is not bound to the operator and CAP is bound (applies to **a–d**).

**Bb.** No expression.

**Bc.** No expression.

**Bd.** No expression.

**Question 8.** Does ZntR bind Zn(II)? Is the spacing between the –10 and –35 regions in the promoter not consensus? Does ZntR bind the promoter region of *zntA*? Does the addition of Zn(II) cause a different pattern of ZntR binding to the promoter? Does the addition of Zn(II) cause a distortion of ZntR-bound DNA? Additional questions are possible as appropriate answers.

For more information, see Outten et al. (1999. *J. Biol. Chem.* 274: 37517–37524).

#### Question 10.

- A. If repressor levels drop too low, the cells might induce the lytic cycle without the bacteriophage being ready for release. If repressor levels increase to a level too high, induction would be inefficient, as more repressor would need to be inactivated before repressor vacates  $O_{R1}$  and  $O_{R2}$  and lytic growth is induced.
- B. When the concentration of  $\lambda$  repressor is too high,  $\lambda$  repressor prevents transcription of itself by binding  $O_{R3}$ . This inhibits RNA polymerase from binding at  $P_{RM}$ .

**Question 12.** To find the specific region where the repressor binds, you should perform a DNA footprinting assay (see Chapter 7 for more information). If the presence of an inducer stops repression, then the reactions should not include inducer in the experimental reaction. The presence of inducer could be used in a control. For the DNA, use a segment of DNA upstream from the structural genes of the operon. If the promoter is known, be sure to include that region. Radiolabel the DNA on one end only. For the experimental reaction, incubate the DNA with repressor, and then briefly treat with DNase I. Run the products on a denaturing gel to find the region of bound repressor during the DNase I treatment. As one potential negative control, include inducer with repressor during the first incubation. Another neg-

ative control is to incubate the DNA with DNase I first, and then incubate with repressor. Instead of DNase I, you may want to try specific chemicals that cleave unprotected DNA.

#### Question 14.

- A.  $\lambda$  repressor bound at  $O_{R1}$  helps repressor bind to  $O_{R2}$  through cooperativity. This allows repressor at  $O_{R2}$  to bind at a lower concentration than would otherwise be necessary due to the lower affinity of  $O_{R2}$ .
- B. From Chapter 18,  $\lambda$  repressor binds  $O_{R1}$  and  $O_{R2}$  cooperatively at low concentration. When these two sites are bound cooperatively, repressor cannot bind cooperatively at  $O_{R3}$ . If  $\lambda$  repressor concentrations get high enough,  $\lambda$  repressor does bind  $O_{R3}$ .
- C. Mutant X is a DNA with a mutation in  $O_{R1}$ . From the data in the table, the researchers did not detect binding to  $O_{R1}$  with Mutant X. The mutation likely disrupted the ability of repressor to bind that sequence. Mutant Y is a DNA with a mutation in  $O_{R2}$ . From the data in the table, the researchers did not detect binding to  $O_{R2}$  with Mutant Y. The mutation likely disrupted the ability of repressor to bind that sequence.
- D. The  $\lambda$  repressor binds DNA cooperatively. In the absence of  $O_{R1}$ ,  $\lambda$  repressor binds  $O_{R2}$  and  $O_{R3}$  cooperatively. This decreases the relative concentration required to fill  $O_{R3}$ .

## CHAPTER 19

---

**Question 2.** Bacterial cells and eukaryotic cells include promoter sequences within the DNA upstream of the coding sequence of a gene. Bacterial and eukaryotic cells also include DNA-binding sites for regulatory proteins like repressors or activators. One activator and/or one repressor protein typically control bacterial genes, whereas the regulatory elements of eukaryotic genes can be more elaborate. In eukaryotic cells, more regulatory elements may be present, the regulatory elements may be present upstream or downstream of the promoter, and regulatory elements may include binding sites for multiple activators and/or repressors. Multiple regulatory elements are grouped as enhancers in multicellular organisms and insulators or boundary elements may be present. The regulatory elements of eukaryotic genes may also be located at much greater distances from the gene they regulate than is the case in bacteria.

**Question 4.** Genomic DNA is wrapped in nucleosomes. Initiation of transcription involves remodeling or removing nucleosomes in a specific area of the genome. This process requires additional proteins. Template DNA like that from a PCR does not include nucleosomes.

#### Question 6.

- A. Correct order: d, c, a, e, b. Review the order in Figure 7-35 in Chapter 7.
- B. The researcher must suspect or know that protein X binds DNA and is asking if protein X binds to *gene Y* or the promoter region of *gene Y*. The researcher may also be testing the

interaction under certain conditions (e.g., in the presence of DNA damage, in the presence of a specific sugar).

#### Question 8.

- A. Methylation, acetylation, and phosphorylation.
- B. Modifications of residues in histone tails are often associated with particular expression profiles of a given gene. Thus, acetylation is in general associated with actively transcribed genes. Other modifications (e.g., methylation) can be associated with either activation or repression of gene expression. Thus, methylation of different residues within histone tails can have different effects, or the context of any given modification (i.e., what other modifications are also present) can affect the outcome of any given modification on expression levels.

**Question 10.** Cytokine—signal. Cytokine receptor—receptor. JAK—relay molecule.

STAT—relay molecule. Transcriptional expression of specific genes—output.

#### Question 12.

- A. Protein A does bind DNA, specifically the DNA fragment included in this EMSA assay. Lane 4 shows the result for the reaction including Protein A and the DNA fragment containing the DNA-binding site for Protein A. The slower migrating band represents the DNA fragment bound to Protein A. The excess unbound DNA is on the bottom of the gel, as in lane 1.

- B.** Proteins A and B bind the DNA fragment. Based on the data in lanes 2 and 5, Protein B does not bind the DNA fragment alone (lane 2). Protein B likely binds to Protein A when Protein A is bound to DNA. The highest slowest migrating band in lane 5 represents the complex of Protein A, Protein B, and the DNA fragment. The next band represents Protein A bound to the DNA fragment. The fastest migrating band represents unbound DNA.

- C.** Protein B does not bind the DNA fragment alone but does bind the DNA fragment if Protein A is bound. Protein A and Protein B could serve as an activator complex to recruit the transcriptional machinery upstream of a specific gene or Protein B alone is the activator but needs Protein A to bring it to the DNA or to bind cooperatively with it.

## CHAPTER 20

**Question 2.** Low transcription. The ribosome will read and translate the mRNA sequence that codes for the leader peptide of the *trp* operon without pausing in the presence of low tryptophan. Without the two *trp* codons, the ribosome easily translates the leader peptide. The 3:4 attenuator forms to prevent transcription of the *trp* operon genes.

**Question 4.** Clustered regularly interspaced short palindromic repeats (CRISPRs) in the genome of prokaryotes and Archaea protect the organism against viral infections. Spacer sequences are added to the arrays and increase resistance to future infection by that same virus.

**Question 6.** The genome encodes pri-miRNAs that form a secondary structure (stem-loop) after being transcribed. pri-miRNAs are processed by Drosha to cleave the lower stem from the upper stem, and these pre-miRNAs can then be further processed by Dicer to liberate mature miRNAs. siRNAs arise from double-

stranded RNAs that form when two complementary RNAs base-pair. After processing by Dicer, they look similar to the processed miRNAs and inhibit gene expression in a similar manner.

**Question 8.** False. Pre-miRNAs sequences can be found in introns, exons, or noncoding regions of transcripts.

**Question 10.** The efficiency of transfection of the long dsRNA into mammalian cells is low. Also, those mammalian cells may shut down translation nonspecifically when dsRNA enters the cells because the dsRNA triggers the same response as viral infection. To overcome this, researchers use short hairpin RNA genes (shRNAs) to express a transcript that folds into a stem-loop processed by Dicer resulting in an siRNA.

**Question 12.** *Xist* recruits proteins to the X chromosome for chromatin remodeling such as methylases, deacetylases, and enzymes that condense the genome.

## CHAPTER 21

**Question 2.** The ability to transform a differentiated cell into an iPS cell demonstrates that each cell type is genetically equivalent. The iPS cell can then become any differentiated cell type. This concept has great potential for medical applications.

### Question 4.

1. A cell or cells synthesize and release the morphogen or signaling molecule.
2. The distribution of released morphogen establishes an extracellular concentration gradient.
3. The morphogen binds receptors on the surface of other cells. The percent occupancy of morphogen decreases as the distance increases between the source cell and the receptor cell.
4. Through a signaling pathway, activated receptor leads to an increase in expression of a transcriptional regulator that controls expression of many genes.

**Question 6.** *B. subtilis* cells use cell-to-cell contact to have the forespore influence mother cell gene expression.  $\sigma^F$  in the forespore activates expression of SpoIIR. Local secretion of the SpoIIR triggers pro- $\sigma^E$  to be cleaved into active  $\sigma^E$  in the abutting mother cell.

**Question 8.** Researchers engineered an mRNA that had the 3' UTR of *oskar* mRNA replaced with the 3' UTR from *bicoid* mRNA. They observed *oskar* mRNA localization in the anterior pole as normally observed for *bicoid* mRNA normally. This is enough to induce pole cell formation at the wrong locations.

**Question 10.** You can fuse the stripe #2 enhancer to the *lacZ* gene from *E. coli*. If LacZ expression is observed in the same position, then the stripe #2 enhancer is sufficient for stripe #2 expression. You can fuse the *lacZ* gene to the *eve* gene in the endogenous location and delete the stripe #2 enhancer in the 5' regulatory region of the *eve* gene to test if it is necessary. If it is necessary, then the stripe #2 should not be observed in the absence of the stripe #2 enhancer. The stripe #2 enhancer is necessary and sufficient for stripe #2 expression.

**Question 12.** Cooperative binding allows sharp changes in protein concentrations from the anterior to posterior regions of the embryo.

**Question 14.** Toolkit genes are conserved across many organisms that are important for the development of all animals.

## CHAPTER 22

---

**Question 2.** A regulatory circuit that follows the logic of an “AND gate” requires that two input conditions are met for output rather than operating as a simple on and off switch.

**Question 4.** Stochasticity means that a system is subject to some degree of randomness. This randomness in regulation of gene expression produces noise or variation in gene expression under apparently identical conditions.

**Question 6.** Negative autoregulation allows the output of the regulatory circuit to be insensitive to a parameter causing noise while maintaining homeostasis.

**Question 8.** ComK cooperatively binds the promoter for *comK*. The response in the positive autoregulation is highly sensitive to slight changes in ComK protein levels, and therefore the output is not linear when threshold is met. The switch is said to be at a “knife edge” between the ON and OFF states.

**Question 10.** The regulation of the gene is under an incoherent feed-forward loop because there is a pulse of gene expression or output. The input of light leads to the production of an activator that directly activates output. Also, the activator leads to the production of a repressor that blocks output. Expression of the gene occurs in the brief time before the repressor is made but after the activator is made.

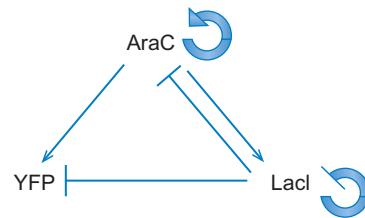
**Question 12.**



Negative autoregulation.

**Question 14.**

- A. AraC in the presence of arabinose will bind the AraC binding site and turns on transcription of the *araC*, *lacI*, and YFP genes and therefore production of their protein products.
- B. LacI is also made in the presence of arabinose; thus, it now binds to all of the *lac* operator sites to shut off transcription of all three genes, so YFP is not made anymore.
- C. LacI turns off its own synthesis but eventually its levels drop below that required for repression as a result of dilution following rounds of cell division and/or degradation. In the presence of arabinose and the small amount of AraC remaining, more AraC will be made to turn on YFP production again.
- D. The YFP signal will persist longer after the initial addition of arabinose. Because IPTG prevents LacI binding, more LacI or a longer time will be required before the system shuts off YFP expression.



# Index

## A

AAA<sup>+</sup> protein, 282b, 289, 295, 298  
*Abdominal-A* (*abd-A*) gene (*Drosophila*), 764b, 766  
*Abdominal-B* (*Abd-B*) gene (*Drosophila*), 764b, 765b  
Absorbance, of DNA, 89–90  
Ac (activator element), 408b  
Acceptor sites, 469  
Accommodation, 538, 539f  
Acetylation, histone, 242, 242f–244f, 244–245, 248–249, 248t  
Acetyl-CoA, high-energy sulfur bond in, 66  
A complex, 476  
Acridine, 324, 324f  
Actin filaments, 735, 735f, 741b  
Activated molecule, 70  
Activated state, 64–65, 64f  
Activating region, of CAP, 622–623  
Activation energy, 64–65, 64f, 65f  
Activator bypass experiments, 624b  
Activators, 453–454, 453f  
allostery and, 618, 618f, 630–633, 632f  
bacteriophage λ CII protein, 645, 647, 648, 651  
cooperative binding and, 619–620  
description, 616  
DNA looping and, 618–619, 619f  
promoter regulation by, 616–618, 617f, 618f  
synergy in development, 752b  
Activators, eukaryotic  
activating regions, 660–661, 660f, 661f, 663–665, 686  
activation at a distance, 672–673  
combinatorial control and, 678–681, 679f, 680f  
cooperative binding, 674–677, 676f, 677f, 678f  
DNA-binding regions, 660–663  
Gal4, 660, 660f, 661f  
helix-loop-helix, 663, 663f  
heterodimers, 662  
homeodomain, 662, 662f  
leucine zipper motif, 662, 663f

zinc-containing domains, 662, 662f  
recruitment of elongation factors, 669–672  
recruitment of nucleosome modifiers, 667–669, 668f  
recruitment of transcriptional machinery to gene, 665–666, 665f  
synergism, 675  
Active site (catalytic site), 130b, 141–142  
Acyclovir, 268b  
Acyl bond, 66, 515  
Acyl-homoserine lactones (AHLs), 635b  
Adaptor hypothesis of Crick, 34  
Adaptor proteins, 735, 735f  
ADAR (adenosine deaminase acting on RNA), 500–501, 501f  
Adenine  
base pairing, 81–82, 81f  
binding to thymine, 24  
Chargaff's rules, 26  
structure, 25f, 27f, 79t, 80, 80f  
Adenosine, deamination, 321, 500–501, 501f  
Adenovirus, 471b–472b  
Adenylylation, of amino acid, 515, 516f  
Adult muscular (myotonic) dystrophy, 316b  
Affinity chromatography, 175–176  
A form of DNA, 86, 87f, 90t  
Agarose gel electrophoresis, 148–149, 148f, 149f  
Agassiz, Jean L., 5  
Aging, telomere hypothesis and, 307b  
*Agrobacterium tumefaciens*  
chromosome makeup, 201t  
gene density, 203t  
genome size, 203t  
Ti plasmid, 813  
AIR, 728  
Alanine, stereoisomers of, 61  
Alanine-tRNA<sup>Lys</sup>, 519, 519f  
Alignment tools, 171–172  
Alkylation, of DNA, 322, 322f  
Alleles  
defined, 7, 343  
dominant, 6–7  
recessive, 6–7  
Allolactose, 626–627, 626f  
Allosteric model of termination, 461f, 462  
Allosteric regulation, 130b, 142f, 143  
Allostery  
activators and, 618, 618f, 630–633, 632f  
cooperativity and, 642b  
*lac* operon control and, 626–627, 627f  
RNA polymerase activation, 618, 618f  
roles in gene regulation, 619–620  
α-carbon, of amino acids, 121, 122f  
αCTD, 438, 438f  
α-helix, 26, 128f, 437  
Alternative spliceosome, 483, 483f, 486, 487f  
Alternative splicing, 483–496  
description, 469, 482  
*Dscam* gene, 487–489, 487f, 489f  
isoforms, 469, 487–488  
mutually exclusive splicing, 486–490, 486f, 487f, 489f  
overview, 484–486, 484f  
pluripotency and, 495–496, 496f  
regulation of, 491–496  
sex determination and, 493–494, 494f, 495f  
SV40 T-antigen, 485f  
troponin T gene, 484, 484f  
*Alu* sequence, 415  
Ambros, Victor, 722b  
Ames, Bruce, 321b  
Ames test, 321b  
Amide bond, 122  
Amino acid residues, of polypeptide chain, 122  
Amino acids  
activation by attachment of AMP, 70–71  
attachment to tRNA, 515–519, 516f  
genetic code for, 37–38, 38t  
hydrophobic and hydrophilic side chains, 125  
hydrophobic bonds, 61–62, 61f  
incorporation by synthetic mRNAs, 578  
in Murchison meteorite, 598

- Amino acids (*Continued*)  
 predicting protein structure from amino acid sequence, 135–136  
 with special conformation properties, 124–125, 125f  
 stereoisomers, 61  
 structure of, 121–122, 122f  
 unusual, 589b–590b
- Aminoacyl synthetases, 70
- Aminoacyl-tRNAs  
 binding to A site of ribosome, 536–537, 537f  
 delivery to A site by elongation factor EF-Tu, 537, 538f  
 formation, 518  
 peptidyl transferase reaction, 524, 524f
- Aminoacyl-tRNA synthetases, 515–519  
 accuracy of, 518–519  
 classes of, 515, 516t  
 description, 510  
 editing pocket, 518–519  
 recognition of correct tRNA, 517–518  
 structure, 518f  
 tRNA charging, 515–519, 516f
- Aminopeptidases, 529
- AMP, amino acid activation by attachment of, 70–71
- Anaphase, 216f, 217
- Anaphase II, 218f, 219
- AND gate, 776, 776f, 785
- Anfinsen Experiment, 134, 135b
- Annotation, 169
- Antennapedia complex, 763, 764b
- Antiactivation, 634
- Antibiotics, translation as target for, 552b–553b
- Antibodies, 133b, 416–418
- Anticancer agents, 268b
- Anticodon loop, of tRNA, 514, 514f, 515f, 517–518, 517f, 525
- Anticodons  
 codon pairing, 574f  
 description, 576–577, 577f  
 wobble concept and, 575–577, 575t, 576f
- anti* conformation, glycosidic bond, 87, 89, 89f
- Antigen-antibody complexes, binding in, 57, 57f
- Antigen-binding site, 416
- Antiparallel orientation of DNA, 81
- Antisense RNA, 171, 410, 410f, 702, 722b
- Antitermination, 620, 648–651, 649f
- Antiviral agents, 268b
- Antp* (*Antennapedia*) (*Drosophila*), 762, 762f, 764b
- APOBEC1 (apolipoprotein-B editing enzyme, catalytic polypeptide-like 1), 503b
- APOBEC3G (A3G), 503b
- Apolipoprotein-B gene, 500, 501f, 503b
- Apoptosis, in *C. elegans*, 818
- Aptamer, 114, 115  
 riboswitch, 703, 703f, 704
- Aqueous solutions, weak bonds between molecules in, 59–60
- araBAD* operon, 634, 634f
- Arabidopsis thaliana*  
 chromosome makeup, 201t  
 development, 815–816  
 environmental response, 815  
 epigenetics, 815–816  
 gene density, 203t, 205t  
 genome, 813–814  
 genome size, 203t  
 haploid and diploid phases, 812–813, 812f  
 life cycle, 812–813, 812f  
 as model organism, 811–816  
 mutagenesis, 814  
 pattern formation, 815  
 repetitive DNA, 205t  
 reverse genetics, 813  
 RNA interference, 814  
 transformation, 813
- Arabinose, 634, 634f
- AraC, 268b, 634, 634f
- Archaeal eon, 595, 596f
- Archaea, RNA polymerases, 431t, 432f, 435f
- Archaeoglobus fulgidus*, 360f
- Architectural proteins, 386
- Arc repressor, 626
- Argonaute family, 713, 718, 719f, 720
- Artemia*, 766
- Artemis, 332, 333f
- Ash1 repressor, 738–741, 739f, 740f
- A site, 525–527, 525f–528f, 528, 529, 531, 531f, 535, 536, 537f–540f, 538, 540–543
- Astbury, William, 24
- AT-AC spliceosome, 483, 483f
- AT hooks, 663
- ATP (adenosine triphosphate)  
 DNA helicase and, 273  
 DNA movement in nucleosomes, 237, 238, 239f, 240t  
 as energy donor, 69  
 in group-transfer reactions, 70–74  
 high-energy bonds, 66  
 as poly-A precursor, 459  
 sliding DNA clamp loading, control of, 281, 282b–283b, 283  
 used in translation, 543–544
- ATPase  
 FtsK, 392b  
 MuB, 413b  
 NtrC, 631  
 TFIID, 453
- ATP-binding motif, of initiator proteins, 289
- Attenuation, 702, 704–705, 707b–708b
- attP* and *attB* sites, 387–389, 388f
- AUG codon, 40, 528, 530, 533b, 535, 536f, 559. *See also* Start codon
- Aurora B kinase, 691b
- Autoinducers, 635b–636b
- Autonomously replicating sequences (ARSs), 290b
- Autonomous transposons, 396
- Autoradiogram, 152
- Autoregulation  
 description, 776–777  
 negative, 643–644, 777, 777f  
 noise, 777–779, 778f  
 positive, 643, 777f, 779–780, 780f, 782b
- Avery, Oswald T., 23
- Azidothymidine (AZT), 268b

**B**

- BAC (bacterial artificial chromosome), 154, 167
- Bacillus subtilis*  
 bacteriophage  $\phi$ 29, 633  
 bacteriophage SPO1 infection, 630, 631f  
*bmr* gene, 633f  
 competence, 780, 783  
 nonhomologous end joining (NHEJ), 332–333  
 riboswitches, 703  
 sporulation in, 743, 743f, 780
- Bacteria. *See also specific species*  
 circular chromosomes of, 92  
 DNA compaction, 200  
 homologous recombination, 342, 349–361, 351t  
 as model organism, 802–808  
 biochemical analysis, 806–807  
 cytological analysis, 807  
 genetic exchange, 803–805  
 conjugation, 803–804, 804f  
 transduction, 804, 805f  
 genetic studies, 807–808  
 growth assays, 803, 803f  
 molecular biology studies, 806  
 plasmid cloning, 805  
 synthetic circuits and regulatory noise, 808  
 transposon-mediated insertional mutagenesis, 805–806, 806f  
 nonhomologous end joining (NHEJ), 332–333  
 regulation by RNAs in, 701–711  
 regulatory elements of, 658f  
 RNA polymerases, 431, 431t  
 transcription in, 434–447
- Bacteriophage. *See also specific phages*  
 complementation tests, 801  
 crosses, 801  
 genome, 798  
 growth assays, 800  
 growth curve, single-step, 800–801, 800f

- Hershey–Chase experiments, 23, 24f  
 induction, 800  
 lysogeny, 798, 799f, 800  
 lytic, 798, 799f  
 as model organism, 798–802  
 prophage, 799f, 800  
 recombinant DNA, 801–802  
 temperate, 798  
 transduction, 802, 804, 805f
- Bacteriophage  $\phi$ 29, 633
- Bacteriophage  $\lambda$   
 antitermination, 648–651, 649f  
 bistable switch, 781  
 $cI$  gene, 777, 779  
 $CII$  protein, 647, 648, 651  
 Cro repressor, 639, 640, 643, 647  
 $E. coli$  growth conditions, effect of, 648  
 excision from host chromosome, 388f, 389  
 integration into host chromosome, 378, 379f, 386–389, 388f  
 Jacob and Monod experiments with, 628b–629b  
 linear and circular forms of DNA, 92  
 $\lambda$  switch, evolution of, 645b–646b  
 lysogeny, 386–387, 636–649, 637f, 694, 695f, 804  
 lytic growth, 386, 636–649, 637f  
 map of, 637f  
 multiplicity of infection (moi), 647–648  
 operators, 639–640, 639f, 640f, 642–644, 647, 647f  
 plaques, 649b  
 promoters, 638, 638f, 639f, 640, 642–643, 645, 645f, 647, 647f, 650–651  
 regulation of gene expression, 636–652  
 repressor, 137, 138f, 139  
   binding sites, 639, 639f, 640  
   cleavage of, 642–643  
   cooperative binding, 639–640, 640f, 641b–642b, 643–644, 644f  
 DNA binding by, 625–626  
 function, 638, 640, 640f, 642  
 negative autoregulation, 643–644  
 structure, 638, 639f, 644  
 retroregulation, 651, 652f  
 vectors, 802
- Bacteriophage Mu, 411, 805
- Bacteriophage P22, 626
- Bacteriophage PM2, 96f
- Bacteriophage SPO1, 630, 631f
- Bacteriophage T4  
 rII locus, 808  
 sliding DNA clamp, 281f
- Bacteriophage T7, RNA polymerase, 443b
- Bait, 664b
- Balancer chromosomes, 822, 822f
- bam* gene, 755b–756b  
 Band shift assay, 183–184, 183f  
 Basal level, of transcription, 617, 617f  
 Base  
   glycosidic bond to sugar, 78  
   hydrolytic damage, 320–322, 320f  
   modifications, 320–322, 320f, 322f  
   purines, 80, 80f  
   pyrimidines, 80, 80f  
   RNA, 107  
   tautomers, 80–81, 81f
- Base analogs, 323, 324f
- Base excision repair, 325t, 326–328, 327f, 328f
- Base flipping, 83, 83f, 327
- Base pairing  
   antiparallel orientation and, 81  
   complementary nature of, 81–82, 81f  
   G:U, 109, 110f  
   hydrogen bonding in, 81f, 82–83, 83f  
   specificity of, 83, 83f  
   Watson–Crick, 81–82, 81f
- Base stacking, 82, 108, 109, 109f
- Base triples, 110, 111f
- Basic HLH proteins, 663
- Basic Local Alignment Search Tool (BLAST), 171
- Basic zipper, 663
- BBP (branch-point-binding protein), 474
- BCNU (bischloroethylnitrosourea), 268b
- B complex, 476
- BDNF (brain-derived neurotrophic factor), 696b
- Beadle, George W., 16, 21
- Beckwith–Wiedemann syndrome (BWS), 696b
- Benzer, Seymour, 808
- $\beta$ -galactosidase, 143, 621, 626, 628b
- $\beta$ -globulin gene, 673–674, 674f
- $\beta$ -interferon gene, 676–677, 678f
- $\beta$  sandwich, 132
- $\beta$ -sheet  
   in DNA polymerase, 263  
   in TATA-binding protein (TBP), 451
- $\beta$ -strand, 126–127, 128f
- $\beta$ -thalassemia, 497b
- B form of DNA, 86–87, 87f, 88, 89f, 90t
- bicoid* mRNA, 751, 753, 753f
- Bicoid protein, 753–754, 754f, 757b
- Biofilm formation, 635b
- Biosynthetic pathway, free energy and, 67–69, 68f
- Biosynthetic reactions, pyrophosphate split and, 73–74
- Bischloroethylnitrosourea (BCNU), 268b
- Bistability, 780–784, 780f, 782b, 784f
- Bithorax complex, 763
- Bivalent attachment, 214, 216f, 217
- BLAST (Basic Local Alignment Search Tool), 171
- Blastocyst, 826, 826f
- Bleomycin, 323
- BmrR, 633f
- Bond angle, 52
- Bonds. *See* Chemical bonds
- Boundary elements, 449, 659
- Bovine papillomavirus E1 hexameric helicase, 274f
- Brachet, Jean, 32
- Brain-derived neurotrophic factor (BDNF), 696b
- Branchipods, 766, 766f
- Branch migration, 344, 345f
- Branch-point-binding protein (BBP), 474
- Branch point site, 469, 474, 486
- BRCA2* gene, 367, 367b
- Brenner, Sydney, 37–38, 288, 583, 807, 808, 816
- Bridges, Calvin B., 10, 13, 821
- Bromodomains, 244, 245, 248, 668
- Bruno, 558f
- Bubble-to-Y transition, 292b
- Budding, in yeast, 738–740, 739f, 740f
- Bulge, RNA structure, 108, 109f
- Buried amino acid side chains, 130
- Burst size, 801
- BWS (Beckwith–Wiedemann syndrome), 696b
- C**
- C9 complement gene, 499, 499f
- Cactus protein, 686, 747f, 750
- Caenorhabditis elegans*  
   bistable switch, 781  
   body plan, 817–818, 817f  
   cell death pathway, 818  
   chromosome makeup, 201t  
   dauer, 816, 817f  
   gene density, 203t, 205t  
   genome size, 203t  
   life cycle, 816–817, 817f  
   miRNAs, 714, 722b  
   as model organism, 816–819  
   repetitive, 205t  
   RNA interference, 722b, 725, 726f, 818–819  
 RNA polymerase II, 450
- Tc elements, 411
- trans*-splicing in, 482–483
- vulva mutations, 818
- CAF-I, 253, 254f
- CAGE (conjugative assembly genome engineering), 590b
- CAG triplet repeat, 316b
- Calico cat, 729, 729f
- Calorie, 53
- Cancer  
   chemotherapeutic agents, 268b  
   leukemia, 670b  
   RNAi and, 727b  
   telomerase activity and, 307b
- Candida albicans*, 588, 683b
- CAP (catabolite activator protein)  
   activating region, 622–623  
   in activator bypass experiments, 624b

- CAP (catabolite activator protein)  
*(Continued)*
- allostery and, 627, 627f
  - araBAD* operon and, 634
  - binding site, 621f, 622, 622f
  - combinatorial control, 627
  - description, 776
  - DNA binding, 621, 622–626
  - effect on RNA polymerase binding, 621f, 622
  - gene location, 621
  - helix-turn-helix motif, 624–626
  - positive control (*pc*) mutants, 622
  - response to glucose, 621, 627
- Cap cells (*Drosophila*), 755, –756b
- Capping, 458, 459f
- Carbonyl bonds, 64
- CA repeats, 315
- Cascade complex, 710, 711f
- Cas* genes/proteins, 709, 710
- Caspersson, Torbjörn, 24, 32
- Cassette exons, 485–486
- Catalysis, substrate-assisted, 541
- Catalyst, 64
- Catenane, 303
- Catenation, 98–99, 99f
- CATH database, 132
- Caulobacter crescentus*, 786, 786f
- CBP (CREB-binding protein), 677
- Cbx* (*Contrabithorax*) (*Drosophila*), 763
- CCA-adding enzymes, 513b
- cccDNA. *See* Covalently closed, circular DNA
- C complex, 476
- CD4 protein, 131f
- Cdc6, 298, 299f
- Cdc13, 308
- Cdc45, 299, 300f
- CDK (cyclin-dependent kinase), 299, 301, 302f
- cDNA (copy DNA)
- library, 156, 157f
  - retrotransposon, 403, 404f
  - retroviral, 403, 404f
  - reverse transcriptase and, 206
- Cdt1, 298, 299f
- Cell cycle
- chromosome replication, 297, 297f, 300–301
  - defined, 210–211
  - gap phases, 212f, 217
  - meiotic, 217–219, 218f
  - mitotic, 211–217, 212f, 213f, 215f, 216f
  - in yeast, 810–811
- Cell cycle checkpoints, 217
- Cell death pathway, in *C. elegans*, 818
- Cell defective (*ced*) mutants, 818
- Cell division, mitotic, 211–217, 212f, 213f, 215f, 216f
- Cell extracts, 173–174
- Cell surface receptor, 684, 736, 736f, 738
- Cell-to-cell contact, 735f, 736, 743, 743f
- Cellulose synthase, 769–770, 770f
- CENP-A (histone variant), 235–236, 236f
- Central dogma, 33–34, 573
- Centromeres, 687
- in cell division, 209, 210f
  - location, 208f
  - repetitive DNA in, 209, 211f
  - silencing in *Schizosaccharomyces pombe*, 713, 719–720, 721f
  - size, 209, 211f
- Centrosomes, 211
- Cernunnos-XLF, 332, 333f
- CGG triplet repeat, 316b
- Chain-terminating nucleotides, 160–162, 163b
- Chain termination, 577, 581, 587–588
- Chalcone synthase, 722b
- Chaperone, 130b, 134, 253, 253t, 254f
- Chargaff, Erwin, 26
- Chargaff's rules, 25, 26
- Chase, Martha, 23, 807
- Chemical bonds, 51–75
- characteristics of, 51–53
  - bond angle, 52
  - energy change, 53, 54
  - equilibrium constant, 53, 54, 54t
  - freedom of rotation, 52, 52f
  - quantum mechanics, 52–53
  - strength, 52
  - valence, 52
- description, 51
- strength, 51, 52, 54
- weak, 51, 52, 55–63
- energies of, 55
  - in enzyme–substrate interaction, 62
  - hydrogen bonds, 57–60, 58f, 58t, 59f
  - hydrophobic bonds, 60–62, 61f
  - between molecules in aqueous solutions, 59–60
  - in protein–DNA interaction, 62–63
  - in protein–protein interaction, 62–63
  - strength, 55
  - van der Waals, 56–57, 56f, 57f, 57t
- Chemical interference footprinting, 184–185
- Chemical protection footprinting, 184
- Chemical reaction, rate of, 65
- Chemotherapy, telomerase inhibitors as, 307b
- ChIP. *See* Chromatin immunoprecipitation
- ChIP-chip assay, 666b–667b
- Chip protein, 672
- ChIP-Seq assay, 666b–667b
- Chi sites, 351–355, 354f
- Chloroflexus*, 709
- Chromatid, 211
- Chromatin
- accessibility, 200
  - activators and, 666–669, 668f
  - defined, 199
  - enhancers and, 672
  - euchromatin, 229, 232, 687, 688
  - gene silencing by modification of, 719–720, 721f
  - heterochromatin, 229, 232, 673, 687, 687f
  - higher-order structure, 229–236, 233f, 235f
  - microscopy, 219–220, 220f
  - regulation of structure, 236–249
  - structure changes during cell cycles, 213f
  - transcription impeded by, 457
- Chromatin immunoprecipitation (ChIP)
- ChIP-chip assay, 666b–667b
  - ChIP-Seq assay, 666b–667b
  - description, 185–187, 186f
  - recruitment by activators, 186f
  - visualization of, 666
- Chromatin remodeling complexes, 250f
- Chromatography
- affinity chromatography, 175–176
  - column, 174–176, 174f
  - gel-filtration chromatography, 174–175, 174f
  - high-performance liquid (HPLC), 177–178
  - immobilized metal affinity chromatography (IMAC), 182
  - ion-exchange chromatography, 174, 174f
  - polyacrylamide gels, 176, 176f
- Chromobacterium violaceum*, 636b
- Chromodomains, 244–245, 248, 668
- Chromosomal theory of heredity, 8
- Chromosome conformation capture assay, 187–189, 188f, 667b
- Chromosome mapping, 11–13
- Chromosomes
- balancer, 822, 822f
  - breakage, 209, 210f, 297f
  - centromeres, 208–209, 208f
  - circular, 92, 200–201
  - condensation, 213
  - definition, 199
  - diversity, 200–202, 201t
  - duplication and segregation, 208–219
  - function, 199
  - gene density, 203t, 204–206, 204f, 205t
  - hot spots of mutation, 315
  - intergenic sequences, 205–208
  - microscopy, 219–220, 220f
  - origins of replication, 204, 208f, 209
  - ploidy, 201–202
  - polytene, 821–822, 821f
  - telomeres, 208f, 209, 211f
- Chromosome segregation, 211, 214, 217, 219, 362–363, 362f

- Chronic lymphocytic leukemia (CLL), 727b  
*cI* gene, 638, 638f, 694, 777, 779, 802  
*CII* activator, 694  
*CIII* protein, 648  
*CII* protein, 645, 647, 648, 651  
*Ciona*, 203t, 770f  
 Circular DNA, multimeric, 391, 391f  
 Cisplatin, 268b  
*cis* splicing, 482  
 Cistron, 808  
 Clastogenic, 323  
 Clay, 598, 598f  
*CLL* (chronic lymphocytic leukemia), 727b  
 Clock protein, 787  
 Cloning. *See* DNA cloning  
 Closed complex, 433f, 434, 438  
 Coactivator, 677  
 Cochran, William, 24  
 Codon–anticodon pairing, 537, 538f, 574f  
 Codons. *See also* Genetic code  
   assignments, 579–582, 580t–581t  
   AUG (start), 40, 528, 530, 533b, 535, 536f, 559  
   description, 38, 510  
   genetic code, 37–38, 38t  
   nonsense, 40  
   stop (chain termination), 40, 510, 554, 577, 583, 584–585, 585f, 587–588  
   synonyms, 573, 586–587  
   wobble concept, 575–577, 575t, 576f  
 Cohesin, 211, 214, 215f, 672  
 Coiled-coils, 129, 129f, 214  
 Cold Spring Harbor Laboratories, 797  
 Colinearity, 37–38  
 Colony hybridization, 156–157  
 Column chromatography, 174–176, 174f  
 Combinatorial control, 627, 630, 678–681, 679f, 680f, 752b  
 ComK, 780–781, 780f, 782b, 783  
 Comparative genome analysis, 769–773  
 Comparative genomics, 40–41  
 Competence, 155, 780–781, 783, 804  
 Complementary structure, weak interactions and, 58  
 Complementation  
   *C. elegans*, 816  
   phage, 801  
   in *Saccharomyces cerevisiae*, 809  
 Composite transposons, 409–410  
 Condensation, chromosome, 213  
 Condensin, 214, 215f  
 Conformational changes in proteins, 136–137  
 Conformation capture assay, 187–189, 188f  
 Conformation, polypeptide chain, 123  
 Consensus sequence  
   intron–exon boundaries, 469f  
   promoter, 435–436, 436b, 449  
 Conservative site-specific recombination (CSSR), 377–393  
   architectural proteins, role of, 386  
   bacteriophage  $\lambda$  excision, 388f, 389  
   bacteriophage  $\lambda$  integration, 378, 379f, 386–389, 388f  
   biological roles of, 377, 386–393  
   by Cre recombinase, 384–385, 386b, 387f  
   enhancers, 390–391, 390f  
   genetic engineering applications, 386b  
   mechanism, 380–381, 381f, 384–385, 387f  
   multimeric, circular DNA conversion to monomers, 391, 391f  
   programmed rearrangements, 389  
   recombination sites, 378–379, 378f, 379f, 380f  
   by *Salmonella* Hin recombinase, 389–391, 390f  
   by serine recombinases, 380, 381f, 381t, 382–383, 382f, 384f  
   structures involved in, 379, 380f  
   types of DNA rearrangements, 379, 379f  
     deletion, 379, 379f  
     insertion, 378f, 379, 379f  
     inversion, 379, 379f, 389–391, 390f  
   by tyrosine recombinases, 380, 381f, 381t, 383–385, 385f  
   Xer recombinase, 391f, 392b  
 Constitutive expression, 617, 628  
 Contigs, 165–167, 165f, 166f  
 Coomassie brilliant blue, 176  
 Cooperative binding  
   of activators, 674–677, 677f, 678f  
   description of, 641b–642b  
    $\lambda$  repressor, 639–640, 640f, 641b–642b, 643–644, 644f  
   recruitment of RNA polymerase, 617, 619f  
   regulation by, 619–620  
   single-strand DNA-binding proteins, 274  
 Copy DNA. *See* cDNA  
 Copy number control, of transposons, 408, 409–410  
 Core DNA, 220, 221f  
 Core enzyme, RNA polymerase, 431, 431f, 434, 438f  
 Core histones, 221–222, 222f, 222t  
 Corn (maize). *See* *Zea mays*  
 Correns, Karl, 6  
 Cosuppression, 722b  
 Coupled reaction, 69, 71  
 Covalent bonds  
   bond angle, 52  
   bond energy, 63–64  
   freedom of rotation, 52, 52f  
   quantum mechanics, 53  
 strength, 51, 54  
 Covalently closed, circular DNA (cccDNA)  
   decatenation of, 98, 99f  
   migration during gel electrophoresis, 102, 102f, 103  
   nucleosome assembly, 230b–231b  
   topological properties of, 93–96, 94f, 95f  
*coxII* gene (trypanosome), 501, 502f  
 CPEB protein, 557  
 CPSF (cleavage and polyadenylation specificity factor), 459, 460f  
 Creighton, Harriet B., 11, 11f  
 Cre recombinase, 384–385, 386b, 387f  
 Crick, Francis H.  
   adaptor hypothesis, 34, 509  
   central dogma, 33  
   DNA structure, 24, 88, 596  
   genetic code and, 38, 807, 808  
   seeding of life on Earth, 606  
   wobble concept, 575  
 CRISPRs (Clustered Regularly Interspaced Short Palindromic Repeats), 706, 709–710, 709f  
 Cro (control of repressor and other things), 639, 640, 643, 647  
 Crossing over  
   chromosome mapping, 11–13  
   consequences of, 363  
   cytological view, 363f  
   description, 9, 11, 341  
 Crossover product, 346, 347f  
 Crossover region, 379, 380f  
 CRP (cAMP receptor protein), 621, 627  
 CrRNAs, 710  
 Crustaceans  
   appendages of, 766, 766f  
   *Ubx* gene in, 763, 766, 766f, 767  
 CSSR. *See* Conservative site-specific recombination  
 CSTF (cleavage stimulation factor), 459, 460f  
 CtBP, 744, 745f  
 CTCF protein, 673, 692  
 CTF1 activator, 664  
 CtrA, 787, 788f  
 Cup, 557, 558f  
 Cut-and-paste transposition, 397–401, 398f, 399f, 400f, 409, 412  
*cut* gene (*Drosophila*), 672  
 Cuzin, Jacques, 288  
 Cycle protein, 787  
 Cyclin-dependent kinase. *See* Cdk  
 Cyclobutane, 322, 322f  
 Cysteine, 124–125, 125f, 519, 519f, 520b  
 Cysteine-tRNA<sup>Cys</sup>, 519, 519f  
 Cystine, 124, 125f  
 Cytidine deaminase, 500, 501f  
 Cytokinesis, 216f, 217, 594  
 Cytological analysis, of bacteria, 807

- Cytosine  
 base pairing, 81–82, 81f  
 binding to guanine, 24  
 Chargaff's rules, 26  
 deamination of, 107–108, 320, 320f, 500–501, 501f, 503b  
 methylation in *Arabidopsis*, 814  
 structure, 25f, 27f, 80–81, 80f, 81f
- Cytosine arabinoside, 268b
- Cytoskeleton, 735, 741b–742b
- D**
- Dam methylase, 318–319, 319f  
 Dam methyltransferase, 294b  
 Darwin, Charles, 5  
 Darwinian evolution  
   described, 594  
   self-replicating protocells, 603–606  
 Darwinian selection,  
   compartmentalization  
   and, 604f  
 Dauer, 816, 817f  
 DCE (downstream core element), 448, 448f, 449  
 DDE transposase superfamily, 404, 412  
 DDK (Dbf4-dependent kinase), 294b  
 Deacetylation of histone, 688, 688f  
 DEAD-box helicase proteins, 474, 477  
 Deamination, 107–108, 320–322, 320f  
   HIV infection and, 503b  
   RNA editing by, 500–501, 501f, 503b  
 Decatenation, 98–99, 99f, 303, 303f  
 Decoding center, ribosome, 521, 525, 527  
*Deformed (Dfd) gene (Drosophila)*, 764b  
 Deformylase, 529  
 Degeneracy, of genetic code, 573  
 Degradative pathways, free energy change, 66  
 Delbrück, Max, 797, 803  
 Deletions  
   by site-specific recombination, 379, 379f  
   from transposition, 403  
 Delta protein, 744, 745f, 789  
 Denaturants, 134, 135b  
 Denaturation, 134, 135b  
   defined, 130b  
   DNA, 89–91, 91f, 92f  
 Deoxynucleoside triphosphates, 258, 258f, 260, 261b–262b  
 Deoxyribonuclease, 23  
 Deoxyribose, 33, 33f, 78, 79f  
 Depurination, 320f, 321  
 Determined nuclei, 747  
 Deuterostome phylogeny, 771f  
 Development  
   *Arabidopsis*, 815–816  
   cis-regulatory sequences, 759b–760b  
   *Drosophila*, 746–769, 819–820, 820f  
     abdominal limb loss, 766–767, 767f  
     activator synergy, 752b, 756–757  
     cis-regulatory sequences, 759b–760b  
   dorsal-ventral patterning, 747, 747f, 750–751, 750f  
   eve gene, 758, 758f, 759f, 761–762, 761f–762f  
   flight limbs, 767–769, 768f  
   gap gene expression, 754  
   homeotic genes, 762–769, 762f–768f, 764b–765b  
    gene, 753–754, 754f  
   overview of embryogenesis, 746–747, 748b–749b  
   segmentation, 751–753  
   stripes of gene expression, 758, 758f, 759f, 761, 761f  
 examples  
   Ash1 repressor control of  
     mating type, 738–740, 739f, 740f  
   muscle differentiation in sea squirt, 740–741, 742f  
   Notch signaling, 743–744, 744f, 745f  
   sporulation in bacteria, 743, 743f  
   vertebrate neural tube, 744–745, 745f  
 feed-forward loops used in, 786  
 gene regulation in, 733–773  
 mouse, 826, 826f  
 strategies for initiating differential gene activity  
   cell-to-cell contact, 735f, 736, 743, 743f  
   mRNA localization, 735–736, 735f, 740–741  
   overview, 735f  
   secretion of signaling molecule, 735f, 737–738, 737f  
 De Vries, Hugo, 6  
 DGCR8, 716  
 Dicer, 715, 717f, 718, 718f, 720, 726  
 Dideoxynucleotides, used in DNA sequencing, 160–161, 160f  
 Differential gene expression, 733, 738–745  
 dif sites, 392b  
 Dihydrofolate reductase gene, 467  
 Dihydrouridine, 514, 514f  
 Dintzis, Howard M., 524  
 Diploid, 201  
 Dipole moment, 55  
 Directed evolution  
   of RNAs, 114, 114f, 115  
   self-replicating ribozymes, 599–603  
 Direct repeat, site-specific recombination and, 380  
 Discriminator, 435f, 437f, 467, 518  
 Disorder, 54  
 Distal-less (*Dll*), 766–767, 767f  
 Disulfide bond, 124–125, 125f  
 D loop, of tRNA, 514–515, 514f, 515f  
 Dmc1, 365f, 366, 366f  
 DNA  
   absorbance, 89–90  
   accessibility, 200, 236–237, 237f, 249, 250f  
   compaction, 200, 221, 232–234  
   complementary base pairing, 24  
   damage to (see DNA damage)  
   as genetic information carrier  
     Avery, MacLeod, and McCarty experiments, 23, 23f  
     Griffith experiment, 22–23, 22f  
     Hershey–Chase experiments, 23, 24f  
     nucleotide sequence and, 30–31  
   hybridization, 151–153  
   isolation of specific segments, 153–154  
   junk, 208  
   labeling, 152  
   looping, 618–619, 619f, 643, 644f, 672  
   melting point, 90, 92f  
   microsatellite, 207  
   precursors of, 71–72  
   protein recognition of DNA sequence, 137, 139–140  
     bacteriophage λ repressor, 137, 138f, 139  
     GCN4, 137  
     lymphocyte enhancer factor-1 (LEF-1), 140, 140f  
     zinc-finger proteins, 139–140, 139f  
   repair (see DNA repair)  
   repetitive, 207–208, 209, 211f  
   replication (see DNA replication)  
   restriction enzyme cleavage of, 149–151, 150f, 150t, 151f  
   separation by gel electrophoresis, 148–149, 148f, 149f  
   sequencing (see DNA sequencing)  
   structure (see DNA structure)  
   synthesis (see DNA synthesis)  
   topology (see DNA topology)  
   transposition (see DNA transposition)  
   X-ray diffraction pattern, 24, 24f, 86, 88
- DnaA, 289, 293, 294b, 295b, 296f  
 DnaB, 293, 295, 296f  
 DNA-binding motif, of σ factor, 437  
 DNA-binding proteins. *See also specific proteins*  
   consensus sequence of binding site, 436b  
   cooperative binding and, 617, 619–620, 619f  
 DNA looping for interaction between, 618–619, 619f  
 initiator proteins, 288  
 nucleosome positioning and, 240–241  
 recognition helix, 624–625, 625f  
 sequence-specific, 274  
 ssDNA-binding proteins, 273–274, 275f
- DnaC, 293, 296f  
 DNA cloning

- clone identification by hybridization, 156–157  
 description, 154  
 library construction, 156–157, 157f  
 in plasmid vectors, 154–155, 155f  
 transformation, 155
- DNA damage, 320–324  
 alkylation, 322, 322f  
 base analogs, 323, 324f  
 deamination, 320–322, 320f  
 depurination, 320f, 321  
 hydrolysis, 320–322, 320f  
 intercalating agents, 323–324, 324f  
 oxidation, 322  
 quantitation of, 323b  
 radiation, 322–323
- DNA fingerprinting, 160b
- DNA footprinting, 184–185, 184f
- DNA gyrase, 97
- DNA hairpin, 400–401, 400f, 409, 419, 419f
- DNA helicase  
 action of, 271, 272–273, 272f  
*DnaB*, 293, 295, 296f  
 DNA polymerase interaction, 284, 285f, 286, 287f  
 in mismatch repair, 316  
 polarity, 272f, 273  
 primase interaction, 271, 284, 285f, 286, 287  
 processivity, 272  
 RecBCD helicase/nuclease, 351–355, 352f–354f  
*RecQ*, 367, 368b  
 structure, 272, 273, 274f  
 supercoiling produced by, 275, 276f  
*UvrD*, 316, 328–329, 329f
- DNA helicase loader, 293, 295, 296f
- DNA ligase  
 in DNA cloning, 155, 155f  
 in DNA replication process, 271, 271f  
 in DNA transposition process, 399  
*Ligase IV*, 332, 333f  
 in mismatch repair, 316  
 in nucleotide excision repair, 329, 329f
- DNA looping, 618–619, 619f, 643, 644f, 672
- DNA methylases, 687, 693f, 694, 695f
- DNA methylation  
 epigenetic regulation, 694, 695f, 696–697  
 imprinting, 692, 693f  
 silencing by, 692, 693f
- DNA microarrays, ChIP and, 186–187
- DNA microsatellites, CA repeats in, 315
- DNA photolyase, 325, 326f
- DNA polymerase  
 accuracy, 267–269, 277  
 active site, 260, 262, 265f  
 activity measurement by incorporation assay, 261b–262b  
 catalysis, 260, 261f, 263–266, 265f, 266f  
 DNA helicase interaction, 284, 285f, 286, 287f  
 DNA Pol  $\alpha$ /primase, 278, 278t, 279f, 286, 299, 300f  
 DNA Pol  $\delta$ , 278, 278t, 279f, 286, 299  
 DNA Pol  $\epsilon$ , 278, 278t, 279f, 286, 299  
 DNA Pol  $\eta$ , 330b, 335, 337f  
 DNA Pol I, 277, 278t  
 identification of, 26–27  
 in nucleotide excision repair, 329, 329f  
 role of, 26–27, 28f  
 DNA Pol II, 670, 671f  
 DNA Pol III, 277, 278t, 284, 286  
 holoenzyme, 277, 284, 284f, 285f, 286, 287f, 296f  
 in mismatch repair, 316, 319, 320f  
 in translesion DNA synthesis, 330b, 334f  
 DNA Pol  $\kappa$ , 335, 337f  
 gap repair, 271, 271f  
 inhibitors of, 268b  
 kinetic proofreading, 260  
 mechanism of, 260–269  
 pausing and release of, 670, 671f  
 primer requirement, 270–272  
 processivity, 265–267, 267f, 277–281, 281f  
 proofreading, 260, 267–269, 269f  
 at replication fork, 270, 270f, 275–277  
 reverse transcriptase, 396  
 ribonucleoside triphosphates (rNTPs), steric exclusion of, 260, 262, 263f  
 sliding clamp association, 278–280, 279f, 280f  
 specialization, 277–283  
 structure, 263–265, 264f, 265f  
 switching during DNA replication, 278, 279f  
 telomerase, 305–309, 306f, 307f  
 translesion, 334, 334f, 335–336, 335f, 337f, 338  
 trombone model for coordination of replication, 285f, 286  
 $\gamma$  family, 334, 335, 336b
- DNA profiling, 160b
- DNA-protein interactions  
 cooperative binding in, 617, 619f  
 in replication initiation, 293, 295  
 weak bonds in, 62–63
- DNA repair  
 of DNA damage, 324–338  
 base excision repair, 325t, 326–328, 327f, 328f  
 double-strand break (DSB) repair, 325, 325t, 330–333  
 nonhomologous end joining (NHEJ), 331–333, 332b, 333f  
 nucleotide excision repair, 325t, 328–329, 329f, 330b, 330f  
 overview, 324–325, 325t  
 photoreactivation, 325, 325t, 326f  
 recombinational repair, 325, 330–331  
 transcription-coupled DNA repair, 329, 330f  
 translesion DNA synthesis, 325t, 333–338, 334f, 335f, 337f  
 gene conversion, 373–374, 374f  
 mismatch repair system, 316–321, 317f–320f, 325t  
 proofreading exonuclease, 315
- DNA replication, 257–311. *See also DNA polymerase*  
 accuracy, 315  
 DNA disentanglement by topoisomerase, 98  
 DNA polymerase, role of (*see DNA polymerase*)  
 DNA polymerase switching during, 278, 279f  
 DNA synthesis, chemistry of, 258–260, 258f, 259f  
 double helix and, 25f  
 double-strand break formation during, 343f  
 end replication problem, 303–305, 304f  
 errors, 314–320, 315f, 317f, 331  
 finishing, 302–310  
 circular chromosomes, 302, 303, 303f  
 linear chromosomes, 303–309, 304f, 306f–310f  
 incomplete, 297f  
 initiation, 288–302  
 in *Escherichia coli*, 293–295, 296f  
 in eukaryotes, 297–302, 297f–301f  
 initiator proteins, 288–289, 288f  
 replicators, 288, 289, 289f, 290b, 292, 297, 297f  
 replicon model, 288, 288f  
 nucleosome assembly and, 249–253, 251f, 252f, 253t, 254f  
 proofreading, 315  
 at replication fork, 269–277  
*E. coli* replisome, 287–288  
 enzymes active at, 275–277, 277t  
 general description, 269–270, 270f  
 lagging strand, 270, 270f, 284, 285f, 286  
 leading strand, 270, 270f, 284, 285f, 286  
 primer removal, 271–272, 271f  
 primer synthesis, 270–271  
 ssDNA stabilization, 273–274, 275f  
 strand initiation, 270–271  
 trombone model for coordination of polymerases, 285f, 286  
 unwinding of DNA, 272–273, 272f, 275, 276f, 287f  
 as semiconservative process, 29–30, 30f

- DNA replication (*Continued*)  
 speed of, 520  
 in S phase of cell cycle, 211, 212f  
 strand separation, evidence for, 27–30, 29f  
 summary, 310–311  
 as target for chemotherapeutic drugs, 268b  
 transcription compared, 429–430
- DNase  
 in mica experiment, 84  
 nuclease protection footprinting, 184, 184f  
 use to relax DNA, 95, 95f
- DNA sequencing  
 chain-termination method, 160–162, 161f, 163b  
 contigs, 165–167, 165f, 166f  
 454 sequencing machine, 167–168, 167f  
 gel, 162f  
 high-throughput, 163b  
 human genome, 164–168  
 nested sets of fragments, 159–162  
 paired-end strategy, 165–167, 166f  
 readout of sequence, 163f  
 scaffolds, 165–167  
 Sequenators, 163b, 168  
 sequence coverage (10x), 162  
 shotgun, 162–167, 164f–166f
- DNA strand transfer, 397
- DNA structure, 77–104  
 antiparallel orientation, 81  
 base flipping, 83, 83f  
 base pairing, 81–83, 81f, 83f  
 bases, 80–81, 80f  
 Chargaff's rules, 25, 26  
 circular, 92  
 complexity in, 77  
 conformations, 86–87, 87f, 89, 89f  
 denaturation, 89–91, 91f, 92f  
 double helix, 77, 78, 78f  
   base pairs per turn, 84  
   B form, 86–87, 87f, 89f, 90t  
   discovery of, 24–25  
   A form, 86, 87f, 90t  
   left-handed, 87, 89  
   major groove, 84–86, 85f, 138f  
   minor groove, 84–86, 85f, 138f  
   periodicity, 84, 89, 103  
   propeller twist, 86, 87f  
   right-handed, 83, 83f  
   strand separation and  
     reassociation, 89–91, 91f  
     Z form, 87f, 89, 89f, 90t  
   phosphodiester bonds, 24, 25f  
   polarity of chains, 79  
   polynucleotide chains, 78–79, 78f, 80f, 81  
 RNA compared, 32–33, 33f, 107–108  
 schematic model, 78f  
 space-filling model, 78f  
 topology, 93–103
- X-ray diffraction pattern, 24, 24f, 88
- DNA synthesis. *See also* DNA replication  
 accuracy, 267–269  
 cell-free, 26–27  
 chemical synthesis of  
   oligonucleotides, 157–158  
 hydrolysis of pyrophosphate, 259f, 260  
 mechanism, 259–260, 259f  
 rate of, 265–266  
 at replication fork, 283–288  
 strand initiation, 270–271  
 substrates, 258–259, 258f  
 translesion, 325t, 333–338, 334f, 335f, 337f
- DNA topoisomers  
 defined, 102  
 separation by electrophoresis, 102, 102f
- DNA topology, 93–103  
 of circular DNA, 93–96, 94f–95f  
 knotted, 98–99  
 linking number, 93–96, 94f, 97, 98f, 101–102  
 nucleosomes and, 96–97  
 relaxed DNA, 95–96, 95f  
 supercoiling, 94–97, 94f, 95f  
 topoisomerase action and, 97–102, 97f–101ff, 98f, 100f
- DNA translocase, 237
- DNA transposition  
 cleavage of nontransferred strand, 399–401, 400f  
 cut-and-paste mechanism, 397–401, 398f, 399f, 400f, 409, 412  
 footprints, 414  
 gap repair, 398–399  
 by replicative mechanism, 401–403, 402f
- DNA transposons  
 autonomous and nonautonomous transposons, 396  
 cut and paste, mechanism, 397–401, 398f, 399f, 400f, 409, 412  
 Ds, 408b  
 genetic organization, 395–396, 395f  
 replicative transposition mechanism, 401–403, 402f  
 target-site duplications, 395f, 396  
 Tc1/*mariner* elements, 411–414
- Dobzhansky, Theodosius, 15
- Docking site, 488, 489f, 490, 490b
- Domains, 130–134, 131f  
 classes of, 132, 132f  
 closure of, 136f  
 defined, 130b  
 exon shuffling and, 498, 498f  
 immunoglobulin G example, 133–134  
 linkers and hinges, 133  
 post-translational modifications, 133–134  
 size of, 130
- Domain swap experiment, 661f
- Dominant, defined, 6–7
- Donor site, 469
- Dormancy, in *Arabidopsis*, 812
- Dorsal protein (*Drosophila*), 686, 747, 747f, 750–751, 750f, 786
- Dorsoventral patterning, of *Drosophila* embryo, 747, 747f, 750–751, 750f
- Dosage compensation, 729
- Double helix  
 antiparallel orientation, 81  
 discovery of, 24–25, 77  
 hydrogen bonding and, 82–83  
 periodicity, 84, 89, 103  
 RNA, 108–110, 109f, 110f
- double-sex* gene (*Drosophila*), 493, 494, 495f
- Double-strand break (DSB) repair  
 description, 325, 325t, 330–333  
 homologous recombination, 342, 346, 348f, 349
- Double-strand breaks (DSBs), 173  
 from DNA transposition, 399  
 formation during DNA replication, 343f  
 by serine recombinases, 382, 382f
- Double-stranded DNA, during meiosis, 363–365, 364f, 365f
- Double-stranded RNA (dsRNA), 712, 712f, 713, 714f, 716, 716f, 717, 720, 720b, 722b, 725–726, 726f, 818–819
- DPE (downstream promoter element), 448–449, 448f
- Dpn (Deadpan), 493, 494f
- Dpp protein, 755b–756b
- Drosha, 715–717, 717f
- Drosophila*. *See also* specific genes  
 alternative splicing, 483  
 centromere size and composition, 211f  
 chromosome makeup, 201t  
*cis*-regulatory sequences, 759b–760b  
 Cup protein, 557, 558f  
 deamination in, 501  
 development, 746–769  
   abdominal limb loss, 766–767, 767f  
   activator synergy, 752b  
   *cis*-regulatory sequences, 759b–760b  
   dorsoventral patterning, 750–751, 750f  
   eve gene, 758, 758f, 759f, 761–762, 761f, 762f  
   flight limbs, 767–769, 768f  
   gap gene expression, 754  
   homeotic genes, 762–769, 762f–768f, 764b–765b  
    gene, 753–754, 754f  
 overview of embryogenesis, 746–747, 748b–749b

- segmentation, 751–753  
 stripes of gene expression, 758, 758f, 759f, 761, 761f  
 docking sites and selector sequences, 490b–491b  
 embryo development, 686  
 embryogenesis, 786  
*engrailed*, 822  
*eve (even-skipped)* gene, 758, 758f, 759f, 761–762, 761f, 762f, 822  
 gene density, 203t, 204f, 205t  
 genetic map of chromosome 2 of, 15f  
 genome size, 203t  
*HP1*, 689, 690, 691b  
*HSP70* gene, 669–670  
 life cycle, 819–820, 820f  
 linked genes in, 9, 10  
*Mariner* element, 411  
 as model organism, 819–825  
     balancer chromosomes, 822, 822f  
     FLP recombinase studies, 823–824, 823f  
     genetic mosaics, 822–823f  
     genome maps, 820–822, 821f  
     hybrid dysgenesis, 824, 824f  
     imaginal disk studies, 820, 820f  
     transgenic flies, 824  
 mutant genes reported in 1915, 14t  
 Nanos protein, 557  
 Oskar mRNA translation regulation, 556–557, 558f  
*patched* gene, 825  
 polytene chromosomes, 821–822, 821f  
 repetitive DNA, 205t  
 RNA polymerase II, 450  
 sex determination, 493–494, 494f, 495f  
 TAFs (TBP-associated factors), 452  
 transposon occurrence and distribution, 394f  
*vnd* locus, 169f  
 whole-genome tiling arrays, 170, 170f  
*yellow (y)* locus, 759b–760b
- DSBs. *See Double-strand breaks*  
*Dscam* (Down syndrome cell-adhesion molecule) gene (*Drosophila*), 487–489, 487f  
 Ds (dissociator) element, 408b  
 D’Souza, Victoria, 112  
 DsRNA. *See Double-stranded RNA (dsRNA)*  
 Dyad axis, 224  
*dystrophin* gene (human), 468
- E**  
 E2F, 686  
 Early (E) complex, 474  
 Earth  
     seeding of life on, 606–607  
     when life arose on, 594–595, 596f  
 EcoRI, 149–150, 150f, 150t, 151f, 151t
- Ectodomain, 130b  
 Edges, in regulatory circuits, 776  
 Edman degradation, 177–178, 178f  
 EF-G, 542–544, 542f, 548, 548f  
 EF-Ts, 543, 544f  
 EF-Tu, 537–538, 538f, 539f, 543, 544, 544f, 546b  
 EGF (epidermal growth factor), 499, 499f  
 EIF1, 530, 531f, 535  
 EIF1A, 531, 531f, 535  
 EIF2, 531, 531f, 535, 556, 559, 560f  
 EIF3, 531, 531f  
 EIF4A, 531f, 532  
 EIF4B, 531f, 532  
 EIF4E, 532, 534b, 556  
 EIF4E-binding proteins (4E-BPs), 532, 556–557, 557f, 558f  
 EIF4G, 531f, 532, 532f, 533b, 534b, 556  
 EIF5, 531, 535  
 EIF5B, 535  
 80S initiation complex, 535  
 Electronegative atoms, 55  
 Electrophoresis  
     agarose gel, 148–149, 149f  
     SDS polyacrylamide for protein separation, 176, 176f  
 Electrophoretic mobility-shift assay (EMSA), 183–184, 183f  
 Electropositive atoms, 55  
 Electrostatic forces, 53  
 ELL family, 456  
 Elongation  
     transcription, 432, 433f, 434, 442–445, 444f, 455–457, 455f, 458f  
     translation, 535–544  
 Elongation factors, 455–456, 537.  
     *See also specific proteins*  
 Embryogenesis, 748b–749b  
     in *Arabidopsis*, 812  
     in *Drosophila*, 819  
     mouse, 826, 826f  
 Embryonic stem (ES) cells, 827–829  
 EMS (ethylmethane sulfate), 814, 821  
 EMSA (electrophoretic mobility-shift assay), 183–184, 183f  
 Endonuclease  
     HO, 369f, 370, 371, 372f  
     MutH, 316, 317f, 318–320, 319f, 320f  
     of poly-A retrotransposons, 396, 406  
     restriction, 149–151, 150f, 150t, 151f, 151t  
     RuvC, 361, 361f  
     Uvr(A)BC, 445  
 Endosperm, 812–813  
 End replication problem, 303–305, 304f  
 Energy  
     activation, 64–65, 64f, 65f  
     chemical-bond formation and, 53  
     first law of thermodynamics, 53  
     free, 54, 65, 66–69  
     hydrolysis, 66  
     second law of thermodynamics, 54  
     of weak bonds, 55  
*engrailed*, 822  
 Enhancers, 171–172, 173f, 390–391, 390f  
     description, 449, 658–659  
     eve, 758, 758f, 759f, 760–762, 761f, 762f  
     exonic splicing enhancers (ESEs), 482  
     of *H19 gene*, 692, 693f  
     identification of, 666b–667b  
     insulators, effect of, 672–673, 673f  
     location, 672  
     multiple, 754  
     *rhombo*, 750–751, 750f, 751f  
 Enhancesome, 677, 678f, 679f  
 ENS (ethylnitrosourea), 814  
 Entropy, 54, 66, 68  
 ENU (ethylnitrosourea), 185  
 Environment, plant response to, 815  
 Enzymes  
     allosteric regulation, 142f, 143  
     function of, 65  
     ligand binding, 142–143  
     lowering of activation energy by, 65, 65f  
     processivity of, 265–267, 267f  
     ribozymes, 114, 116–118, 116f, 117f  
     structure, 141–142  
     substrate binding, 62  
 Ephrussin, Boris, 16  
 Epidermal growth factor (EGF), 499, 499f  
 Epigenetics  
     in *Arabidopsis*, 814–815  
     gene regulation, 694–697, 695f  
     in mice, 829, 830f  
 Epitopes, 175–176  
 Equilibrium concentration, 54  
 Equilibrium constant, 53, 54, 54t, 65, 65t, 68  
*Escherichia coli*. *See also specific genes/proteins*  
     araBAD operon, 634, 634f  
     chromosome makeup, 201t  
     circular chromosome, 92  
     combinatorial control, 627  
     CRISPR, 710, 711f  
     Dam methylase, 318–319, 319f  
     DnaA, 289, 293, 294b–295b, 296f  
     DNA compaction, 200  
     DNA polymerases, 277, 278t, 284, 285f, 286  
     DSB-repair, 349–355  
     factors that catalyze recombination steps, 351t  
     gal genes, 627  
     Gal repressor, 633  
     gene density, 203–204, 203t, 204f, 205t  
     genome size, 203t  
     initiation of DNA replication, 293–295, 296f  
     lac system, 620–627, 628b–629b

*Escherichia coli* (Continued)  
 mismatch repair system, 316–320, 317f–320f  
 noise in gene expression, 778, 779f  
 nucleotide excision repair in, 328–329  
 origin of replication, 204, 288, 289, 293, 294b–295b  
 primase, 271  
 RecA, 355–359, 355f–359f  
 RecBCD pathway, 349–355, 352f–354f  
 repetitive DNA, 205, 208  
 replication fork enzymes, 277t  
 replication regulation by DnaA-ATP levels and SeqA, 294b–295b  
 replisome, 287–288  
 ribosomal protein operons, 554f  
 σ factors, 434–438, 435f, 437f–440f, 442  
 6S RNA, 701  
 sliding DNA clamp, 285f  
 sRNAs, 701–702  
 superhelical density, 229  
 translesion synthesis, 334  
 transposon occurrence and distribution, 394f  
 tryptophan operon of, 707b–708b  
 ESE (exonic splicing enhancer), 482, 491  
 E site, 525–526, 525f–528f, 541, 542, 548  
 ESS (exonic splicing silencer), 491  
 EST (expressed sequence tag), 169  
 Ester bond, 515  
 Ethidium, 102–103, 102f, 148, 324, 324f  
 Ethylmethane sulfate (EMS), 814, 821  
 Ethylnitrosourea, 185, 814  
 Euchromatin, 229, 232, 687, 688  
 Eukaryotes  
   chromosome makeup, 201–202, 201t  
   chromosome structure, 208–209, 208f  
   DNA polymerases, 277–278, 278t, 286  
   factors that catalyze recombination steps, 351t  
   gene density, 205  
   genome size, 203, 203t  
   homologous recombination, 342, 351t, 362–369  
   initiation of replication, 297–302  
   messenger RNA, 511, 511f, 512–513  
   mismatch repair systems, 319–320  
   regulatory RNAs in, 711–730  
   RNA polymerases, 431, 431t, 432f  
   transcriptional regulation in, 657–698  
   transcription in, 448–463  
   translation initiation, 530–535, 530f–534f  
   translation regulation, 556–561  
 Eukaryotic ribosome, 521, 522f  
*eve (even-skipped)* gene, 758, 758f, 759f, 760–762, 761f, 762f, 822  
 Evo-devo, 763

Evolution  
 anomalous genes, 769–770  
 comparative genome analysis, 769–773  
 conserved genes, 769  
 evolvability of a regulatory circuit, 683b  
 of flight limbs, 767–769, 768f  
 homeotic genes and, 763  
 human origins, 772  
 of the λ switch, 645b–646b  
 RNA catalysis and, 479–480  
 RNA interference, 721  
 RNA splicing, 479  
 synteny and, 770–772  
 systems biology, 775–776  
 Excision repair systems  
   base excision repair, 325t, 326–328, 327f, 328f  
   description, 325  
   nucleotide excision repair, 325t, 328–329, 329f, 330b, 330f  
 Exon definition, 476, 482  
 Exonic splicing enhancer (ESE), 482, 491  
 Exonic splicing silencer (ESS), 491  
 Exons, 467–468, 468f  
   cassette, 485–486  
   identification of, 169  
 Exon shuffling, 497–500, 499f  
 Exon skipping, 484f, 485  
 Exonuclease  
   in DNA mismatch, 318–319  
   proofreading, 268–269, 269f, 315  
 Exonuclease I, 319  
 Exonuclease VII, 318  
 Exponential phase of growth, 803, 803f  
 Exposed amino acid side chains, 130  
 Expressed sequence tag (EST), 169  
 Expression platform, riboswitch, 703, 703f  
 Expression vectors, 155, 634  
 Extended exon, 485, 485f  
 Extracellular gradient, 737  
 Extracellular matrix, 737f  
 Extrinsic noise, 778–779, 778f

**F**

FACT (facilities chromatin transcription), 457, 458f  
 Factor-binding center, ribosome, 537, 538f, 546b  
 Factor for inversion stimulation (Fis), 390–391, 390f  
 Familial isolated growth hormone deficiency type II, 497b  
 Feed-forward loops, 784–786, 785f, 787f  
 Ferritin, translation of, 558, 559f  
 Fertilization, in *Arabidopsis*, 812–813, 812f  
 F-factor, 803–804, 804f  
 FGF (fibroblast growth factor), 769  
 Fibroblast growth factor (FGF), 769  
 Fingerprint, DNA, 160b

Fire, Andrew, 722b  
 Fisher, Ronald A., 15  
 Fis protein, 390–391, 390f  
 5-bromouracil, 323, 324f  
 5-fluorouracil (5-FU), 268b  
 5 methyl group, 107  
 5 methyl-uracil, 107  
 5' cap, 530–532  
 5' splice site, 469–470, 469f, 473, 474, 476, 486  
 Flagella, 389, 389f  
 Flagellin, 389–390  
 Flanking host DNA, 397, 398f, 401  
*fliAB*, 389, 390f  
 Flight limbs, evolution of, 767–769, 768f  
 Floorplate, 744–745  
 FLP recombinase, use in *Drosophila*, 823–824, 823f  
 Fluorescent chain-terminating nucleotides, 161, 163b  
 Fluorescent labeled DNA, 151, 170, 170f  
 fMET (*N*-formyl methionine), 529, 529f  
 fMet-tRNA, 579  
 fMet-tRNAifMet, 529, 530  
 FMRP (fragile X mental retardation protein), 727b  
 Fold (topology), 130b, 132b  
 Folding chaperones, 134  
 Footprinting, 184–185, 184f  
 Forensics, polymerase chain reaction (PCR) and, 160b  
 Forespore, 743, 743f  
 48S preinitiation complex, 532, 535, 536f  
 43S preinitiation complex, 531, 531f, 532f  
 Forward genetics, in *Arabidopsis*, 814  
 454 sequencing, 167–168, 167f  
 FOXP1, 495–496, 496f  
 Fragile X mental retardation protein (FMRP), 727b  
 Fragile X syndrome, 316b, 727b  
 Frameshift mutations, 583, 584, 584f  
 Franklin, Rosalind, 24, 88  
 Frasier syndrome, 497b  
 Freedom of rotation, 52, 52f  
 Free energy  
   in biomolecules, 66–67  
   biosynthetic pathways, 67–69, 68f  
   degradative pathways, 66  
   description, 54  
   in DNA synthesis, 260  
   enzymes and, 65  
   hydrolysis, 70  
   in nucleic acid synthesis, 72  
 Free radicals, 322  
*Fritillaria assyriaca*  
   gene density, 203t  
   genome size, 203t  
 FtsH, 648  
 FtsK, 392b  
*Fugu rubripes*  
   chromosome makeup, 201t  
   gene density, 203t, 205t

- genome size, 203t  
repetitive DNA, 205t
- Fusidic acid, 552b
- Fusion proteins, 154
- G**
- G1 phase, 212f, 217  
G2 phase, 212f, 217
- GAGA-binding factor, 669
- GAL1* gene, 660, 660f, 666, 666f, 668, 682f, 686, 686f
- Gal4 activator, 660, 661f, 664b, 668, 669, 671, 671f, 672, 686, 686f
- Gal11 protein, 664, 669
- Gal80, 686, 686f
- gal* genes, 627, 681, 682f
- Gal repressor, 633
- galR* gene, 627
- Gametes, in *Arabidopsis*, 812–813, 812f
- Gametophyte, 812, 812f
- γ radiation, DNA damage from, 322
- Gap genes, 749b, 754, 757b
- Gap phases, 212f, 217
- Gap repair, in cut-and-paste transposition, 398–399
- Garrod, Archibald E., 16, 21
- G:C content, 90, 92f
- Gcn2, 559, 560f
- Gcn4, 129f, 137, 558–559, 560f, 561, 664, 669
- GCR (global control region), 674
- GcrA, 787, 788f
- Gel electrophoresis
- agarose, 148–149, 148f, 149f
  - of DNA topoisomers, 102, 102f, 103
  - ethidium intercalation into DNA, effect of, 102–103
  - pulsed-field, 149, 149f
  - SDS, 176, 176f
- Gel-filtration chromatography, 174–175, 174f
- Gel matrix, 148
- Gel shift assay, 183–184, 183f
- Gene(s)
- defined, 7
  - early speculations concerning nature of, 15–16
  - linked, 9–11
  - representation by letters or symbols, 7
- Gene conversion
- description, 368–369
  - from DSB repair, 349, 373–374, 374f
  - mating-type switching, 370–371
  - during meiotic recombination, 373
  - noncrossover recombination, 368
- Gene density, 203t, 204–206, 204f, 205t
- Gene–enzyme relationship, 16
- Gene expression
- differential, 733, 738–745
  - noise in, 777–779, 778f
  - regulation of translation, 549–561
- “Gene-for-gene” response, 815
- Gene fragments, 206, 206f
- Gene fusion, transposon-generated, 805–806, 806f
- Gene order, three-factor crosses to assign, 12, 12f
- Generalized transduction, 804, 805f
- General transcription factors (GTFs), 448, 449
- Gene regulation
- in development and evolution, 733–773
  - regulatory circuits, 776–790
  - autoregulation, negative, 777, 777f
  - autoregulation, positive, 777f, 779–780, 780f, 782b
  - bistable switches, 780–784, 780f, 782b, 784f
  - feed-forward loops, 784–786, 785f, 787f
  - AND gate logic, 776, 776f
  - nodes and edges, 776, 776f
  - noise, 777–779, 778f
  - oscillating, 786–790, 788f
  - robustness, 779
  - stochasticity, 778, 779
  - synthetic circuits, 789–790
- Gene silencing, 659, 687–693. *See also Silencing*
- Gene size, 205
- Genetic code, 573–590
- cracking, 37–38, 577–582
  - degeneracy, 573–577
  - expanded, 589b–590b
  - of mammalian mitochondria, 587–588, 588t
  - order in makeup of, 575
  - point mutations, 582–583
  - rules governing, 582–583
  - suppressor mutations, 584–587
  - table of, 38t, 574t, 588t
  - universality, 587–590
  - validity, 586–587
  - wobble concept, 575–577, 575t, 576f
- Genetic competence, 155, 804
- Genetic engineering, application of site-specific recombination to, 386b
- Genetic map
- description, 11–13, 15f
  - in *Drosophila*, 820–822, 821f
  - homologous recombination and, 373, 373f
- Genetic variability, origin through mutations, 13, 15
- Genome
- Arabidopsis*, 813–814
  - bacteriophage, 798
  - Drosophila*, 821
  - duplication, 813
  - Escherichia coli*, 204
  - gene density, 203–204, 203t, 204f, 205t
  - intergenic sequences, 205–208, 205t
- mouse, 825–826
- Saccharomyces cerevisiae*, 810
- synthetic, 594, 595f
- transposable elements in, 393–394
- Genome editing, 172–173
- Genome maps, in *Drosophila*, 820–822, 821f
- Genome sequencing
- Haemophilus influenzae*, 162
  - shotgun sequencing, 162–167, 164f–166ff
- Genome size, 769
- complexity of organism and, 202–203, 203t
- Genome-wide repeats, 207
- Genomic library, 156
- Genomics, 168–173
- annotation, 169
  - genome editing, 172–173
  - overview of, 40–41
  - regulatory DNA sequences, 171–172, 172f, 173f
  - whole-genome tiling arrays, 169–171, 170f
- Genotype, defined, 7
- Geological timeline, 596f
- Germline stem cell (GSC), 755b–756b
- Germline transformation, 827
- GGQ motif, 547
- giant enhancer, 754, 756, 757f
- Gibbs, Josiah, 54
- GINS, 299, 300f
- Gli activator, 747, 757b
- glnA* gene, 618, 631, 632f, 659
- Glucose, structure of, 52f
- Glutaminyl aminoacyl-tRNA synthetases, 518f
- Glycine, 57f, 124
- Glycoprotein, 134
- Glycosidic bond, 78, 89, 89f
- Glycosylase, 326–328, 327f, 328f
- Glycosylation
- of amino acid side chain, 134
  - defined, 130b
- Glyoxal, 149
- Goff, Stephen, 112
- Gradient thresholds, 757b
- GreB, 456, 457f
- Green fluorescent protein (GFP), 115, 115f
- Gre factors, 445, 456
- Griffith, Frederick, 22
- Groucho, 744, 745f
- Group I introns, 477, 477t, 478–480, 478f, 479b
- Group II introns, 477, 477t, 478f, 479, 480, 480f
- Group activation, 70
- Group transfer reactions, 69–74, 71f
- Growth curve
- bacterial, 803, 803f
  - single-step, 800–801, 800f
- GSC (germline stem cell), 755b–756b

- GTFs (general transcription factors), 448, 449
- GTPase  
EF-Tu, 537–538, 538f  
IF2, 529  
Ran, 505
- GTP-binding proteins, 546b
- GTP exchange factor, 543
- GTP, use in translation, 543–544, 546b, 556
- Guanine  
5' cap, 512  
alkylation, 322, 322f  
base pairing, 81–82, 81f  
binding to cytosine, 24  
Chargaff's rules, 26  
deamination of, 321, 322f  
depurination of, 320f, 321  
oxidation of, 322, 322f, 327f, 328  
structure, 25f, 80–81, 80f, 81f
- Guanylyltransferase, 459f
- G:U base pair, 109, 110f
- GUG, as start codon, 528
- Guide RNAs, 501–503, 502f, 713, 718
- Gyandromorphs, 822–823, 823f
- Gyrase, 229
- H**
- H19* gene, 692, 693f, 696b
- Hadean eon, 595, 596f
- Haeckel, Ernst, 6
- Haemophilus influenzae*, genome sequencing, 162
- Hairless, 744, 745f
- Hairpins  
DNA, 400–401, 400f, 409, 419, 419f  
RNA, 108, 109f  
in transcription terminators, 447, 447f
- Haldane, John Burden Sanderson, 15, 16
- Half-pint protein, 492
- Halteres, 767–768, 768F
- Hammarsten, Ola, 24
- Hammerhead ribozyme, 116–117, 117f
- Haplod, 201
- Haplod set of chromosomes, 199
- HATs (histone acetyltransferases), 248–249, 248t, 667, 668f
- HAT* transposon family, 401, 408b
- Hayflick, Leonard, 307b
- Hayflick limit, 307b
- HcPro, 724
- Heat, free energy and, 54, 66, 69
- Heat shock  $\sigma$  factor,  $\sigma^{32}$ , 630
- Helicase. *See also* DNA helicase  
eIF4A, 532, 535  
loading proteins, 293, 295, 296f, 298–300, 299f, 300f  
Mcm2-7, 298, 299f, 300f
- Helix-loop-helix proteins, 663, 663f
- Helix-turn-helix motif, 137, 437, 624–626, 625f, 661, 662f
- Hemimethylation, 318–319
- Hemoglobin  
cooperative binding and, 642b  
sickle-cell, 31
- Her1 protein, 788–789, 789f
- Her7 protein, 788–789, 789f
- Heredity  
chromosomal theory of, 8  
Mendelian laws, 6–8, 7f, 9f
- Hermaphrodites, 812, 817, 817f
- Hermes* transposon, 400f, 401, 408b
- Herpes virus activator VP16, 661
- Hershey, Alfred D., 23, 797, 807
- Heterochromatin, 229, 232, 673, 687, 687f, 813
- Heterodimers, 662
- Heteroduplex, 344, 345f, 374
- Heterogeneous nuclear ribonucleoprotein (hnRNP) family, 492
- Heterozygous, defined, 7
- Hexokinase, 136, 136f
- hfl* (high frequency of lysogeny), 648, 802
- Hfq protein, 702
- Hfr (high-frequency recombinant), 803–804, 804f
- High-energy bonds  
in ATP, 66  
in biosynthetic reactions, 67–69  
classes of, 67t  
defined, 63, 66  
free energy and, 66–67  
group-transfer reactions, 69–74  
hydrolysis, 66  
in nucleic acid synthesis, 71–74, 73f  
symbol for, 66
- High-performance liquid chromatography (HPLC), 177–178
- High-throughput sequencing, 163b
- HindIII, 150, 150f
- Hinges, protein, 133
- Hin recombinase, 389–391, 390f
- Histone acetyltransferases (HATs), 248, 248t, 249, 667, 668f
- Histone code, 691b
- Histone deacetylase, 248, 248t, 681–682, 682f, 688
- Histone demethylases, 248, 248t, 249
- Histone-fold domain, 222, 222f, 234
- Histone methyltransferases, 248, 248t, 249, 689, 690
- Histones. *See also* Nucleosome  
chaperones, 253, 253t, 254f  
core, 221–222, 222f, 222t  
deacetylation, 688, 688f  
defined, 199  
DNA binding, 224, 225f, 226–229, 227f, 232, 232f, 233f, 234f, 241f, 235f  
DNA replication and, 249–253, 251f, 252f, 253t, 254f  
DNA wrapping and negative superhelicity, 228–229  
effect on transcription, 456–457, 458f  
gene silencing and, 720
- inheritance, 251, 252f, 253
- linker, 221–222, 222t
- methylation, 689–690, 691b, 696–697
- modifications, 224, 241–245, 242f–244f, 248–249, 248t, 720
- nucleosome assembly and, 222–224, 223f
- properties of, 222, 222t
- structure, 221–224, 222f
- tails, 224, 227–228, 228f, 234, 234f, 241–244, 242f, 243f, 244f
- transcriptional regulation and, 657
- variants and nucleosome function, 234–236, 236f
- HIV. *See* Human immunodeficiency virus
- hix* sites, 389–391, 390f
- HMGA1 enhancer, 677
- HMG proteins, 663
- hnRNP1, 492, 493, 493f
- Hoagland, Mahlon B., 34, 509
- HO endonuclease, 369f, 370, 371, 372f
- HO gene, 675–676, 677f, 739–740
- Holliday junction  
cleavage, 346, 347f, 350b, 361  
description, 344  
generation, 344, 345f  
recombination intermediate, 348f, 350b
- RecQ helicases and, 368b
- resolving, 344, 346, 347f, 350b, 361
- RuvAB complex recognition of, 360, 360f
- RuvC, 361, 361f
- in tyrosine recombinase site-specific recombination, 383, 385f, 392b
- Holoenzyme, 434, 435f
- Homeodomain proteins, 662, 662f
- Homeotic genes, 762–769, 762f–768f, 764b–765b
- Homologous chromosomes, crossing over between, 11, 11f
- Homologous domains (or proteins), 130b
- Homologous proteins, 132
- Homologous recombination, 341–375  
in bacteria, 342, 349–361, 351t  
chromosome segregation and, 362–363, 362f
- double-strand break (DSB) repair, 331, 342, 346, 348f, 349
- in eukaryotes, 342, 351t, 362–369
- gene knockout via, 827–829, 828f
- genetic consequences of, 371, 373–374, 373f, 374f
- key steps, 343–346  
branch migration, 344, 345f  
heteroduplex DNA, 344, 345f
- Holliday junction, 344, 345f, 346, 350b, 347f, 359–361, 360f, 361f
- resolution, 344, 346, 347f, 350b, 361

- strand invasion, 344, 345f  
 mating-type switching, 369–371, 369f, 372f  
 during meiosis, 362–369, 362f–366f  
 models for, 342–349  
   double-strand break-repair, 346, 348f, 349  
   Holliday, 344, 345f, 346, 347f  
 overview, 341  
 protein machines, 349–361  
   RecA, 355–359, 357f–360f  
   RecBCD, 349–355, 352f–354f  
 yeast, 810
- Homologs, 201, 217, 219, 363  
 Homology modeling, 130b, 136  
*Homo sapiens*. *See* Human  
 Homozygous, defined, 7  
 Horvitz, Robert, 818  
 HOTAIR, 728  
*Hoxb-2* gene, 827, 827f  
*HoxD* genes, 674  
*Hox* genes, 188–189, 764b–765b, 769  
 HP1 protein, 689, 690, 691b  
 HpaI, 150, 150f  
 HPLC (high-performance liquid chromatography), 177–178  
 Hrp36, 490, 492  
 HSF, 669  
*HSP70* gene (*Drosophila*), 669–670  
 hsp70 promoter, 823  
 hSPT5, 458  
 Human  
   alternative splicing, 483  
   β-interferon gene, 676–677, 678f  
   centromere size and composition, 211f  
   chromosome makeup, 201  
   disease, RNAi and, 727b  
   ferritin translation, control of, 558, 559f  
   gene density, 203t, 204f, 205–206, 205t  
   mediator, 454, 454f  
   origins, 772  
   Rad51, 360f  
   regulatory elements, 658f  
   repetitive DNA, 205t  
   replication fork enzymes, 277t  
   splicing defects, 497b–498b  
   telomere-binding proteins, 308, 308f  
   telomere t-loop, 309, 310f  
   transposon occurrence and distribution, 393–394, 394f
- Human genome  
   intergenic sequences, 206  
   organization and content of, 206f  
   repetitive DNA, 207–208  
   sequencing, 164–168  
   size, 203t  
   transposable elements in, 207–208
- Human immunodeficiency virus (HIV)  
   deaminases and, 503b  
   promoter, 670
- tat* pre-RNA, 492  
*hunchback* gene, 753–754, 754f  
 Hunchback protein, 754, 755f, 756, 757f, 758, 759f, 760–761  
 Huntington's disease, 316b  
 HU proteins, 391  
 Hurwitz, Jerard, 36  
 Huxley, Julian, 15  
 Huxley, Thomas H., 5  
 Hybrid dysgenesis, 824, 824f  
 Hybridization  
   colony, 156–157  
   defined, 89  
   DNA, 151–153, 152f  
   to identify a specific clone in a DNA library, 156–157  
   in microarray analysis, 153, 153f  
   northern blot, 152–153  
   reannealing and, 91f  
   Southern blot, 152–153, 152f  
 Hybridization probes, 151–153, 152f  
 Hydrogen-bond acceptor, 85–86  
 Hydrogen-bond donor, 85–86  
 Hydrogen bonds  
   base pairing and, 81f, 82–83, 83f  
   bond lengths, 58t  
   complementary molecular structures, 58–59  
   description, 57–58  
   directional properties, 58, 59f  
   in DNA, 24  
   examples in biological molecules, 58f, 58t  
   ionic bonds as, 58  
   in nucleosome, 227  
   in protein secondary structure, 126–127, 127f, 128f  
   in water, 125, 126f  
   between water molecules, 59, 59f  
   in water soluble organic molecules, 60
- Hydrolases, 69  
 Hydrolysis  
   description, 66  
   DNA damage from, 320–322, 320f  
   free energy of, 70  
   group-transfer reactions, 69  
   GTP, 538, 539f, 547–548  
   hydrolases, 69  
   nucleic acids, 71  
   of peptide bonds, 68  
   of pyrophosphate, 259f, 260
- Hydrolytic editing, by RNA polymerase, 444–445
- Hydrophilic  
   amino acid side chain, 125–126  
   molecules, 60
- Hydrophobic  
   amino acid side chain, 125–126  
   bonds, 60–62, 61f  
   molecules, 60
- Hyperchromicity, 90  
 Hypoxanthine, 321, 514  
 Hysteresis, 782b
- I**
- ICR (imprinting control region), 692, 693f, 696b  
 IF1, 529–530, 530f  
 IF2, 529–530, 530f  
 IF3, 529–530, 530f  
*Igf2* gene, 692, 693f, 696b, 829, 830f  
 IHF (integration host factor), 388–389, 388f, 414, 631–632, 672  
 Imaginal disks, 820, 820f  
 Imino acid, 124  
 Immobilized metal affinity chromatography (IMAC), 182  
 Immune system, 416  
 Immunoaffinity chromatography, 175–176  
 Immunoblotting, 176–177, 177f  
 Immunoglobulin G (IgG), protein domains of, 133b  
 Immunoprecipitation, 176  
 Immunoprecipitation, chromatin. *See* Chromatin immunoprecipitation (ChIP)  
 Imprinting, 692, 693f, 814, 829, 830f  
 Imprinting control region (ICR), 692, 693f, 696b  
 Inborn errors of metabolism, 16  
 Inchworming model, 441, 441f  
 Incorporation assay, to measure DNA synthesis, 261b–262b  
 Independent assortment, principle of, 8, 9f  
 Independent segregation, principle of, 6–7, 7f  
 Induced pluripotent stem (iPS) cells, 666b, 733–735, 734b  
 Inducer, 143  
 Induction, of lytic phage pathway, 800  
 Ingram, Vernon M., 31  
 Initial transcribing complex, 434  
 Initiation  
   transcription, 432, 433f, 434, 440–442, 441f, 449–454, 450f  
   translation, 528–535  
 Initiation factors  
   eukaryotic, 530–535, 531f, 532f, 536f  
   prokaryotic, 529–530, 530f  
 Initiator (Inr), 448–449, 448f  
 Initiator protein, 288–289, 288f, 293  
 Initiator tRNA, 528–529, 530, 532  
 Inner cell mass, 733–734, 734b, 826, 826f  
 Insects, loss of abdominal limbs in, 766–767, 767f  
 Insertion  
   frameshift mutation from, 583  
   by site-specific recombination, 379, 379f  
 Insertional mutagenesis  
   in *Arabidopsis*, 814  
   transposon-generated, 805–806, 806f  
 Insertion sequence (IS), 410  
 Insulators, 449, 659, 672–673, 673f, 692

- Insulin-like growth factor 2 (*Igf2*) gene, 692, 693f, 696b, 829, 830f
- Integrase  
catalytic domain, 404–405, 405f  
cDNA recognition by, 403  
description, 395  
retroviral, 403, 405, 405f  
structure, 404–405, 405f
- Integration  
of bacteriophage λ, 378, 379f, 386–389, 388f  
retrotransposons and retroviruses, 403, 404f  
Ty elements, 414, 415f
- Integration host factor (IHF), 388–389, 388f, 414, 631–632, 672
- Interactome, 182, 183f
- Intercalating agents, 323–324, 324f
- Interference, 13
- Intergenic sequences, 205–208
- Intergenic suppression, 584–585
- Internal loops, RNA, 108, 109f
- Internal ribosome entry sites (IRESs), 533b–534b
- Interphase, 213, 213f, 216f
- Interwound writhe, 94, 230b
- int* gene, 651, 652f
- Intragenic suppression, 584, 584f
- Intrinsic noise, 778, 778f
- Intrinsic terminators, 446–447
- Intronic splicing enhancer (ISE), 491
- Intronic splicing silencer (ISS), 491
- Introns. *See also* RNA splicing  
AT-AC, 483  
description, 467–468, 468f  
genome size, contribution to, 205, 205t  
group I, 477, 477t, 478–480, 479bf  
group II, 477, 477t, 478f, 479, 480, 480f  
identification of, 169  
number per gene, 467, 468f  
removal by ribozymes, 116  
self-splicing, 477–480, 478t, 479b, 480f  
size, 467
- Introns early model, 497, 498
- Introns late model, 498
- Inversions  
by site-specific recombination, 379, 379f, 389–391, 390f  
from transposition, 403
- Inverted repeats  
site-specific recombination and, 379  
of transposons, 395, 395f, 396
- In vitro selection, 189, 189f
- Ion-exchange chromatography, 174, 174f
- Ionic bonds, 58
- Ionizing radiation, DNA damage  
from, 323
- iPS (induced pluripotent stem) cells, 495, 666b, 733–735, 734b
- IPTG, 626f
- IRESs (internal ribosome entry sites), 533b–534b
- Iron regulatory element (IRE), 558, 559f
- Iron regulatory proteins (IRPs), 558, 559f
- IS (insertion sequence), 410
- IS3 family transposons, 401
- IS4 family transposons, 409–410
- ISE (intronic splicing enhancer), 491
- Isoaccepting tRNAs, 517
- Isoforms, 469, 487–488
- Isoleucine, 518, 518f
- Isoleucyl-tRNA synthetase, 519
- Isomerization, of RNA polymerase, 438–440
- Isopods, 766
- ISS (intronic splicing silencer), 491
- J**
- Jacob, François, 288, 628b–629b, 804, 807
- Janssens, F. A., 9, 11
- JNK1 gene (human), 486
- Joint molecule, 358
- Jorgensen, Richard, 722b
- Jun, 676, 678f, 684
- Junctions, RNA, 108, 109f
- Junk DNA, 208
- K**
- Kendrew, John, 24
- Khorana, Har Gobind, 38
- Kinetochore, 208f, 209, 211, 213f, 214, 219
- Kirromycin, 552b
- Kluyveromyces lactis*, 683b
- Knirps* gene (*Drosophila*), 749b, 754, 756, 757f
- Knirps protein, 760–761, 761f
- Knotted DNA, 98–99
- Kornberg, Arthur, 26
- Kozak, Marilyn, 512
- Kozak sequence, 512
- Krüppel* gene, 754, 756, 757f
- Krüppel protein, 758, 760–761, 762, 762f
- Ku70, 332, 333f
- Ku80, 332, 333f
- Kuhn, A., 16
- L**
- Labeling DNA, 152
- labial* (*lab*) gene (*Drosophila*), 764b
- lacA* gene, 620, 620f
- lac* genes  
description, 620–621, 620f  
expression, 621, 621f, 626–627
- lacI* gene, 621, 628b
- lac* operator  
description, 620f–622f, 622, 625  
mutation in, 628b–629b
- lac* operon  
allosteric control, 626–627, 627f  
control region, 622f  
description, 620–621, 620f  
expression of *lac* genes, 621, 621f, 626–627
- AND gate logic, 776, 776f
- Jacob and Monod experiments, 628b–629b
- lac* promoter, 620–623, 620f–623f
- lac* repressor, 142f, 143  
allostery and, 626–627  
effect on RNA polymerase binding, 621, 621f, 633
- gene location, 620
- helix-turn-helix motif, 624–626, 625f
- response to lactose, 621, 626–627  
as tetramer, 625, 625f
- Lactase (*LCT*) gene (human), 760b
- Lactose permease, 621, 626f
- lacY* gene, 620, 620f, 621
- lacZ* gene, 620, 620f, 621, 661f, 805, 806f, 827, 827f
- Lagging strand, 270, 270f, 284, 285f, 286
- λ integrase protein (λInt), 387–389, 388f
- Lariat structure, 470, 470f, 477, 478, 478f
- Last Universal Common Ancestor, 595
- Latent period, 801
- LCR (locus control region), 671–672, 671f, 673–674, 674f
- Leading strand
- Lepidopterans, wings of, 768, 768f
- Lesions, DNA, 314
- let-7* miRNA, 714
- Leucine  
codons, 573  
tRNA, 574f
- Leucine zipper, 662, 663f
- Leukemia, 670b
- LexA, 335, 643, 660, 661f
- L-forms, 594
- Library  
cDNA, 156, 157f  
construction, 154, 156–157, 156f, 157f  
genomic, 156  
screening for specific clone, 156–157  
whole genome, 164, 164f
- Life  
origin and early evolution of, 593–607  
seeding Earth with, 606–607
- Life Sciences sequencing machine (454), 167–168, 167f
- Ligand, 142–143
- Ligase IV, 332, 333f
- Ligase ribozyme, 601, 601f, 602–603, 603f
- lin-4* gene, 722b–723b
- lin-14* gene, 722b–723b
- LINE (long interspersed nuclear element), 414–416, 415f
- Linked genes, 9–11
- Linker DNA, 220–221, 221f, 221t, 232–233, 233f
- Linker histone, 221–222, 222t
- Linkers, protein, 133
- Linking difference, 95–96, 103
- Linking number (LK), 93–96, 94f, 97, 98f, 101–102, 229, 230b–231b

- Lipid vesicles, 605, 605f, 606f  
 Liquid chromatography with mass spectrometry (LC-MS), 179, 180f, 181, 181f, 182  
*Listeria monocytogenes*, 110, 111f  
 $Lk^O$ , 94–96  
 Lock-and-key relationship, 58  
 Locus control region (LCR), 673–674, 674f  
*Locusta migratoria*  
 gene density, 203t  
 genome size, 203t  
 Long-intervening non-coding RNAs (lincRNAs), 207  
 Long non-coding RNAs (lncRNAs), 728  
 Long-terminal repeats (LTRs), 395, 396, 403, 404f  
 Loop, RNA, 108–109, 109f  
 Low-density lipoprotein (LDL) receptor gene, 499, 499f  
*lox* sites, 385  
 Luciferase, 635b  
 Luria, Salvador, 797, 803  
 LuxR, 635b–636b  
 Lymphocyte enhancer factor-1 (LEF-1), 140, 140f  
 Lysogenic induction, 636, 642–643  
 Lysogeny  
 bacteriophage, 798, 799f, 800, 802, 804  
 bacteriophage  $\lambda$ , 386–387, 636–649, 637f, 694, 695f  
 phage Mu, 411  
 Lysozyme, 586  
 Lytic growth  
 bacteriophage  $\lambda$ , 386, 636–649, 637f  
 phage Mu, 411  
 Lytic phage, 798, 799f
- M**
- Macho-1, 740–741, 742f  
 MacLeod, Colin M., 23  
 Macromutations, 15  
 Mad, 756b  
 MADS (MCM1, agamous, deficiens, and serum response)  
 transcription factors, 815  
 MAGE (multiplex automated genome engineering), 589b  
 Maintenance methylases, 694, 695f  
 Maize. *See Zea mays*  
 Major groove, DNA, 84–86, 85f, 138f  
 $malT$  gene, 618  
 Map units, 13  
 Maskin protein, 557  
 Mass action, 53, 65, 74  
 Mass spectrometry  
 LC-MS analysis, 179, 180f, 181–182, 181f  
 tandem (MS/MS), 178–179, 180f  
 Mating-type locus (*MAT* locus), 369–371, 369f, 372f, 680–681, 680f  
 Mating-type switching, 369–371, 372f, 738–740, 739f, 740f  
 Matthaei, Heinrich, 38  
 Matzke, Marjori, 722b  
 Maxillipedes, 766, 766f  
 Mayr, Ernst, 15  
 McCarty, Maclyn, 23  
 McClintock, Barbara, 11, 11f, 408b  
 Mcm1, 680–681, 680f  
 MeCP2, 692, 696, 696b  
 Medea, 756b  
 Mediator Complex, 453–454, 453f, 454f, 665, 665f, 666f, 669  
 Megakaryocytes, 202  
 Meganucleases, 173  
 Meiosis  
 chromosome number reduction, 217–219  
 chromosome segregation, 362–363, 362f  
 phase, 218f, 219  
 Meiosis I, 218f, 219  
 Meiosis II, 218f, 219  
 Meiotic cell cycle, 217–219, 218f  
 Meiotic recombination, 363–369, 363f–366f, 373  
 Mello, Craig, 722b  
 Melting point, of DNA, 90, 92f  
 Mendel, Gregor, 6–8, 803  
 Mendelian laws  
 history, 6  
 independent assortment, 8, 9f  
 independent segregation, 6–7, 7f  
 Mercaptoethanol, 176  
 MerR, 618, 630–631, 632–633, 632f  
 $merT$  gene, 632  
 $merT$  promoter, 618, 632, 632f, 633f  
 Meselson, Matthew, 27–30, 342, 807  
 MesP, 752b  
 Messenger RNA (mRNA)  
 5' cap, 512  
 amino acid incorporation into synthetic, 578  
 broken, 563, 564f  
 circularization of eukaryotic, 532, 532f, 535  
 description, 510  
 discovery of, 35  
 eukaryotic, 511, 512–513  
 eukaryotic ribosome recruitment to, 530–532, 531f  
 function of, 108  
 localization, 735, 740–741  
 monocistronic, 511, 511f  
 nonsense-mediated decay, 565–567, 566f  
 nonstop-mediated decay, 567, 568f  
 northern blot hybridization, 152–153  
 open reading frame (ORF), 510–512  
 poly-A tail, 513  
 polycistronic, 511, 511f  
 prokaryotic, 511f, 512  
 recruitment to ribosome in prokaryotes, 528, 528f  
 ribosome-binding site (RBS), 511f, 512, 528, 528f, 549, 551, 553, 554  
 ribosome entry/exit, 527  
 structure, 511f  
 translation and, 510–513, 511f  
 translation-dependent regulation of stability, 563–567  
 transport, 503–505, 504f  
 Metabolite sensors, 115, 115f  
 Metal ions, in DNA polymerase, 263, 264, 265f  
 Metaphase, 216f, 217  
 Metaphase II, 218f, 219  
 Methionine, 528–529, 529f  
 Methylation  
 DNA, 814  
 effect on restriction enzymes, 150  
 epigenetic regulation and, 694, 695f, 696–697  
 in mismatch repair process, 318–319, 319f  
 silencing by, 692, 693f  
 histone, 242, 242f–244f, 244–245, 248–249, 248t, 689–690, 691b, 696–697  
 RNA-dependent DNA, 814  
 Methylcytosine, 321  
 Methylguanine, 514  
 Methyltransferase, in DNA repair, 325–326, 326f  
 Mica experiment, 84  
 Microarray analysis, 153, 153f  
 whole-genome tiling microarray, 169–171, 170f  
 Micrococcal nuclease (MNase), 226b, 245b–247b  
 Microinjection, in mice, 827, 827f  
 Microprocessor complex, 716, 717f  
 MicroRNA (miRNA), 712–718, 712f, 715f–717f, 717b, 722b–723b  
 in *Arabidopsis*, 814  
 in *C. elegans*, 819  
 genes, 170–171  
*Hox* activity modulation, 765b  
 human disease and, 727b  
 as intergenic sequences, 206–207  
 production of, 716–718  
 structure, 714–716, 715f  
 Microsatellite DNA, 207  
 Microtubule-organizing centers, 211, 213, 214, 216f, 217  
 Microtubules, 209, 211, 214, 216f, 735, 741b–742b  
 Mig1, 681–682, 682f  
 Miller, Stanley, 596–597  
 Miller–Urey experiment, 596–597, 596f  
 Minor groove, DNA, 84–86, 85f, 138f, 227, 227f, 451  
 Minor spliceosome, 483, 483f, 486, 487f

- miRNA. *See* microRNA
- Mismatch repair system
- description, 316–321, 317f–320f
  - gene conversion, 374, 374f
- Missense mutations, 582, 584
- Mitochondria, genetic code of, 587–588, 588f
- Mitogen-activated protein kinase (MAPK) pathway, 684, 685f
- Mitosis (M phase), 211, 212f, 213f
- Mitotic cell divisions, 211–217, 212f, 213f, 215f, 216f, 811f
- Mitotic spindle, 208f, 211
- MNase (micrococcal nuclease), 226b, 245b–247b
- Model organisms, 797–830
- Arabidopsis*, 811–816
  - bacteria, 802–808
  - bacteriophage, 798–802
  - baker's yeast (*Saccharomyces cerevisiae*), 808–811
  - Caenorhabditis elegans*, 816–819
  - choice of, 797–798
  - Drosophila melanogaster*, 819–825
  - mouse (*Mus musculus*), 825–830
- Moi (multiplicity of infection), 647–648
- Molecular biology techniques, 147–189
- nucleic acid, 148–168
    - DNA cloning, 154–157, 155f
    - DNA hybridization, 151–153, 152f, 156–157
    - DNA library, 156–157, 156f, 157f
    - DNA segment isolation, 153–154
    - DNA sequencing, 158–168
    - electrophoresis, 148–149, 148f, 149f
    - oligonucleotide synthesis, 157–158
    - plasmid vectors, 154–155, 155f
    - polymerase chain reaction (PCR), 158, 159f, 160b
    - restriction endonucleases, 149–151, 150f, 150t, 151f
    - transformation, 155, 155f
  - nucleic acid–protein interactions, 182–189
  - chromatin immunoprecipitation, 185–187, 186f
  - chromosome conformation capture assay, 187–189, 188f
  - electrophoretic mobility shift assay, 183–184, 183f
  - footprinting, 184–185, 184f
  - in vitro selection (SELEX), 189, 189f
  - proteins, 173–179
    - immunoblotting, 176–177, 177f
    - purification from cell extracts, 173–174
    - separation
      - affinity chromatography, 175–176
- gel-filtration chromatography, 174–175, 174f
- ion-exchange chromatography, 174, 174f
- polyacrylamide gels, 176, 176f
- sequencing, 177–179, 178f
- Edman degradation, 177–178, 178f
  - tandem mass spectrometry (MS/MS), 178–179, 180f
  - proteomics, 179–182, 180f, 181f, 183f
- Molecular mimicry, 440, 543
- Molecules, described, 51
- Monocistronic mRNA, 511, 511f
- Monod, Jacques, 628b–629b, 804, 807
- Monosiga*, 769
- Monovalent attachment, 214, 216f, 218, 219
- Morgan, Thomas Hunt, 10, 11–13, 819, 820–821
- Morphogen, 737f, 738, 744
- Morula, 826, 826f
- Mosaics, 728–729, 822–823, 823f
- Motif (sequence), 130b
- Motif (structural), 130b
- Mouse (*Mus musculus*)
  - chromosome makeup, 201t
  - gene density, 203t
  - genome, 825–826
  - genome size, 203t
  - Hox* genes, 764b–765b
  - life cycle, 825
  - as model organism, 825–830
    - embryonic development, 826, 826f
    - epigenetics, 829, 830f
    - knockout, 827–829, 828f
    - microinjection, 827, 827f
    - synteny with humans, 826
    - transgenic mice, creation of, 827, 827f
  - patched* mutants, 825

M phase. *See* Mitosis

mRNA. *See* Messenger RNA (mRNA)

MRX, 364–365, 365f, 370

MS/MS, 178–179, 180f

MTE (motif ten element), 448

mTor, 556

MuA protein, 411, 413b

MuB protein, 411, 413b

Muller, Hermann J., 10, 13, 16, 821

Multimeric circular DNA, 391, 391f

Multiplicity of infection (moi), 647–648

Murchison meteorite, 598, 598f

Murine leukemia virus (MLV), 112

Muscle differentiation in seq squirt embryo, 740–741, 742f

Mutagenesis, 323, 335

  - EMS, 814, 821
  - insertional
    - in *Arabidopsis*, 814
    - transposon-generated, 414, 805–806, 806f
  - site-directed, 157

*Sleeping Beauty*, 414

Mutagens, 320, 321b

Mutant genes, 10

"Mutate-and-map" strategy, for RNA structure prediction, 113

Mutation

  - Ames test, 321b
  - from base analogs and intercalating agents, 323–324, 324f
  - complementation, 801, 809
  - consequences of, 314
  - described, 13, 314
  - from DNA damage, 320–324
  - epigenetic variation, 815
  - frameshift, 583, 584, 584f
  - hot spots, chromosome, 315
  - macromutations, 15
  - missense, 582, 584
  - nonsense, 582–583
  - null, 805
  - origin of genetic variability, 13, 15
  - point, 582–583
  - recombinational transformation, 810
  - from replication errors, 314–320, 315f, 317f
  - reverse (back), 584
  - sources of, 313
  - splicing defects, 497b–498b
  - suppressor, 584–587, 584f, 585f
  - transition, 575
  - from transposition, 393
  - transversion, 575
  - types
    - point mutations, 314
    - transitions, 314, 314f
    - transversions, 314, 314f
    - triplet repeat expansion, 316b
  - X-ray induced, 16
  - yeast, 810

MutH, 316, 317f, 318–320, 319f, 320f

MutL, 316, 317f, 318–319, 320f

Mu transposase, 405f

MutS, 316, 317f, 318, 318f, 320f

*Mycoplasma*, 594

  - M. capricolum*, 588
  - M. genitalium*
    - chromosome makeup, 201t
    - gene density, 203t
    - genome size, 203t

Myosin, 735f

**N**

NANOG protein, 495, 496

Nanos protein, 557, 754, 754f

Neanderthal DNA, 772

Nearest neighbors, 59

Negative autoregulation, 643–644, 777, 777f

Negative superhelicity, 228–229, 230b, 231b

*Nematostella*, 771–772

Neural tube development, 744–745, 745f

Neurogenic ectoderm, 744, 744f

- Neurospora crassa*, riboswitches in, 705  
 NF-κB, 676, 678f, 686  
*N*-formyl methionine (fMET), 529, 529f  
 NHEJ (nonhomologous end joining), 331–333, 332b, 333f  
 Nirenberg, Marshall, 38  
 Nitrosamines, 322  
 NMD (nonsense-mediated decay), 486–487, 487f, 565–567, 566f  
 NMR (nuclear magnetic resonance), for RNA structure prediction, 111, 112  
 Node, of primitive streak, 826  
 Nodes, in regulatory circuits, 776  
 No-go decay, 567, 568f  
 Noise  
   defined, 778  
   extrinsic, 778–779, 778f  
   intrinsic, 778, 778f  
 Nonautonomous transposons, 396  
 Non-crossover products, 346, 347f  
 Nondisjunction, 363  
 Nonhistone proteins, 199  
 Nonhomologous end joining (NHEJ), 331–333, 332b, 333f  
 Nonpolar molecules, 55–56  
 Nonsense codons, 40  
 Nonsense-mediated decay (NMD), 486–487, 487f, 565–567, 566f  
 Nonsense mutations, 582  
 Nonsense suppressors, 585–586, 585f  
 Nonstop-mediated decay, 567, 568f  
 Nontransferred strands, in DNA transposition, 399–401, 400f  
 Nonviral retrotransposons. *See* Poly-A retrotransposons  
 Northern blot hybridization, 152–153  
 Notch, 744, 745f, 789  
 Notch-Su(H) regulatory switch, 744, 745f  
 NotI, 150, 150t  
 N protein, λ, 649–650  
 NREs (Nanos response elements), 754  
 NtrC, 618, 630–632, 632f, 684  
 Nuclear magnetic resonance (NMR), for RNA structure prediction, 111, 112  
 Nuclear pore complex, 504  
 Nuclear scaffold, 231b, 234, 235f  
 Nuclease. *See also* Endonuclease; Exonuclease  
   RecBCD helicase/nuclease, 351–355, 352f–354f  
 Nuclease protection footprinting, 184, 184f  
 Nucleic acids  
   hydrolysis, 71  
   synthesis, 71–72, 73f  
 Nucleic acid techniques, 148–168  
   DNA cloning, 154–157, 155f  
   DNA hybridization, 151–153, 152f, 156–157  
   DNA library, 156–157, 156f, 157f  
   DNA segment isolation, 153–154  
   DNA sequencing, 158–168  
   electrophoresis, 148–149, 148f, 149f  
   oligonucleotide synthesis, 157–158  
   plasmid vectors, 154–155, 155f  
   polymerase chain reaction (PCR), 158, 159f, 160b  
   restriction endonucleases, 149–151, 150t, 151f  
   transformation, 155, 155f  
 Nucleoid, 201, 202f  
 Nucleoside, 78  
 Nucleoside phosphates, 71  
 Nucleoside triphosphates, 72, 73f, 258–260, 258f  
 Nucleosome core particle, 224, 226b  
 “Nucleosome linking number paradox,” 231b  
 Nucleosome modifiers, 453, 453f, 454, 657–658, 667–669, 668f  
 Nucleosome-remodeling complexes, 237–238, 237f 242, 238f, 240t  
 Nucleosomes  
   acetylated, 668  
   arrays, 232–233, 237  
   assembly, 222–224, 223f, 230b–231b, 249–253, 254f  
   atomic structure, 224  
   axis of symmetry, 224, 225f  
   core DNA, 220, 221f  
   description, 200  
   DNA access, 236–237, 237f, 249  
   epigenetic inheritance and, 696–697  
   histones  
    core, 221–222, 222f, 222t  
    DNA binding/interaction, 224, 225f, 226–229, 227f, 232, 233f, 234f, 236–237, 241f  
    linker, 221–222, 222t  
    structure, 221–224, 222f  
    tails, 224, 227–228, 228f, 234, 234f, 241–244, 242f, 243f, 244f  
    variants, 234–236, 236f  
   lacking H2A and H2B, 227f  
   linker DNA, 220, 221, 221f, 221t, 232–233, 233f  
   micrococcal nuclease (MNase)  
    treatment, 222b  
   modification, 688f, 696–697  
   movement, 237–238, 238f, 239f, 240t  
   negative supercoiling, 96–97  
   negative superhelicity, 230b  
   positioning, 240–241, 240f, 241f, 245b–247b  
   regulatory role of, 200  
   superhelical density, 230b–231b  
   30-nm fiber, 232–234, 233f, 234f, 244  
   transcriptional regulation and, 657  
   transcription and, 457, 458f  
 Nucleotide analogs, 268b  
 Nucleotide excision repair, 325t, 328–329, 329f, 330b, 330f  
 Nucleotides  
   description DNA, 27f  
   DNA and RNA compared, 32–33, 33f  
   formation of, 78–79, 79f  
   generation from simple organic molecules, 597–598, 597f  
   genetic code, 37–38, 38t  
   phosphodiester linkage, 79, 80f  
   sequence as source of genetic information, 30–31  
 Nucleus, chromosomes within, 202, 202f  
 Null mutation, 805  
 Nurse cells, 752  
 Nus proteins, 445, 456  
 Nüsslein-Volhard, Christiane, 749b  
 nut (N utilization), 650
- O**
- OCT4 protein, 495, 496  
 Okazaki fragments, 270, 270f, 271, 285f, 286, 287, 303, 319  
 Oligomer, protein, 129  
 Oligonucleotides  
   chemically synthesized, 157–158  
   defined, 157  
   primers in PCR procedure, 158, 159f  
 Oligoribonucleotides, preparation of, 581f  
 Oncogenes, miRNAs as tumor suppressors of, 727b  
 One gene–one enzyme hypothesis, 21  
 Oocyte, 752  
 Open complex, 433f, 434, 438–440, 440f  
 Open reading frame (ORF), 169, 510–512, 533b, 559, 560f, 561  
 Operator, 143  
   bacteriophage λ, 639–640, 639f, 640f, 642–644, 647, 647f  
   cI gene, 777, 779  
   lac, 620f–622f, 622, 625  
   repressor binding, 617  
 Operon  
   *araBAD*, 634, 634f  
   *lac*  
    allosteric control, 626–627, 627f  
    control region, 622f  
    description, 620–621, 620f  
    expression of *lac* genes, 621, 621f, 626–627  
    AND gate logic, 776, 776f  
    Jacob and Monod experiments, 628b–629b  
    ribosomal proteins, 553–554, 554f  
    *trp*, 707b–708b  
 Optical density of DNA, 90  
 Organic molecules  
   in Murchison meteorite, 598  
   pre-biotic, 595–598  
*oriC*, 289, 289f, 293, 294b–295b

- Origin of replication  
defined, 288  
*Escherichia coli*, 204, 288, 289, 294b–295b  
of eukaryotes, 297, 298f  
function of, 209  
identification of, 290b–292b  
location, 208f, 209
- Origin recognition complex (ORC), 293, 298, 301
- Oryza sativa*  
gene density, 203t, 205t  
genome size, 203t  
repetitive, 205t
- Oscillating regulatory circuit, 786–790, 788f
- Oskar, 556–557, 558f  
*oskar* mRNA, 751–753, 753f
- Oxidation of DNA, 322
- OxoG, 322, 326, 327f, 328
- P**
- p53* gene, 827  
p73, 497b
- Paige, Jeremy, 115
- Paired-end sequencing, 165–167, 166f
- Pair-rule gene, 749b, 758
- Palindromic sequence, 137
- Paper chromatography, 26
- par-1* gene, 722b
- Parental imprinting in mice, 829, 830f
- Partially diploid cells, 628b, 629f
- Pasha, 716
- Passenger RNA, 713
- Pasteur, Louis, 808–809
- patched* gene, 825
- Patch products, 346, 347f
- Pattern formation, in *Arabidopsis*, 815
- Pauling, Linus, 24
- PAZ domain, 717, 718f
- PCNA, 253, 254f, 280, 281f, 320, 336
- PCR. See Polymerase chain reaction
- PCR2, 728
- P-elements, 824, 824f, 825f
- Peptide bond  
defined, 122  
formation of, 122, 123f, 599, 600f  
formation within ribosome, 524, 524f, 538, 540–541, 541f  
hydrolysis, 68  
planar shape of, 52f  
structure, 122–123, 123f
- Peptide recognition, 140, 140f
- Peptidyl transferase, 600f
- Peptidyl transferase center, ribosome, 118, 521, 525–527, 538, 540f, 547
- Peptidyl transferase reaction, 524, 524f, 537, 540
- Peptidyl-tRNA, 524, 524f, 537
- per gene, 787–788
- Permissive cells, 503b
- Perutz, Max, 24
- Petunia, 722b
- Phage. *See* Bacteriophage
- Phage Group, 797
- PHD (plant homeodomain) fingers, 244, 245
- Phenotype, defined, 7
- Phenylalanine, 16, 518, 518f, 579
- Phenylisothiocyanate (PITC), 177, 178f
- Phosphamidines, 157, 158f
- Phosphodiester bond  
cleavage and formation in RNA splicing, 470  
in DNA, 24, 25f, 79, 80f  
energy in, 71–72, 74–75  
formation of, 259  
recombinase cleavage, 380–381
- Phosphoproteome, 182
- Phosphorylation  
of amino acid side chain, 134  
histone, 242, 242f, 243f  
of RNA polymerase II, 450–451, 450f
- Phosphotyrosine linkage, 99, 100f
- Photoreactivation, 325, 325t, 326f
- Photosynthesis, 66
- PITC (phenylisothiocyanate), 177, 178f
- Pitx1 gene, 759b–760b
- Piwi-interacting RNAs (piRNAs), 713, 718, 724
- Plants. *See also* *Arabidopsis*  
development and pattern formation, 815–816  
environmental response, 815  
RNAi in, 722b, 724  
RNA polymerases, 431
- Plaque, 649b, 800, 800f
- Plaque assay, 800, 800f
- Plasmids  
chromosomes compared, 201  
as circular genetic elements, 92  
as cloning vectors, 805  
conjugation, 803–804, 804f  
description, 154
- Plasmid vectors, 154–155, 155f, 805
- Plectonemic writhe, 94
- Pluripotency, 495–496, 496f
- Pluripotent, 734
- Point mutations, 314, 582–583
- Polar granules, 751, 753
- Polarity, of DNA helicase, 272f, 273
- Polar molecules, 55–56, 60
- Pol I, II, or III. *See* DNA polymerase
- Poly-A-binding protein, 532, 532f
- Polyacrylamide, 148
- Polyacrylamide gels, separation of proteins on, 176, 176f
- Polyadenylation, 458–460, 460f
- Polyadenylic acid (poly-A), 578
- Poly-A polymerase, 459, 460f, 513
- Poly-A retrotransposons, 395f, 396, 405–406, 407f
- Poly-A sequence, 396, 415
- Poly-A signals, 459
- Poly-A tail, translation efficiency and, 513, 535
- Polycistronic mRNA, 511, 511f
- Polycomb repression complex (PRC), 690, 690f, 728, 765b, 815
- Polycomb Response Element (PRE), 690, 690f
- Polymerase. *See also* DNA polymerase; RNA polymerase  
switching, 278, 279f  
translesion, 325
- Polymerase chain reaction (PCR)  
in ChIP assay, 186, 186f  
forensic use, 160b  
procedure, 158, 159f
- Polymorphism, CA repeats, 315
- Polynucleotide chains, 78–79, 78f, 80f, 81
- Polynucleotide phosphorylase, 578, 578f
- Polypeptide chain, 122, 123–124, 510
- Polyphenylalanine, 579
- Polyplloid, 201–202
- Polypyrimidine (Py) tract, 469, 474
- Polypyrimidine (Py) tract-binding protein, 493, 493f
- Polyribonucleotide chain, 32f
- Polyribosomes (polysomes), 35, 36f, 522–523, 523f
- Polysome profiling, 561b–562b
- Polytene chromosomes, 821–822, 821f
- Positional information, 737
- Position-weighted matrix (PWM), 172
- Positive autoregulation, 643, 777f, 779–780, 780f, 782b, 783f
- Positive control (*pc*) mutants, 622
- Post-translational modifications, 133–134
- POT1, 308, 308f
- Potato virus Y, 724
- PRC (polycomb repression complex), 690, 690f, 765b, 815
- Prebiotic chemistry, 597
- Precursor RNAs (pri-RNAs), 170–171
- Preinitiation complex, 449, 450f, 453f
- pre-mRNA  
description, 468  
splicing of, 468–469, 468f, 480f, 484f
- Prey, 664b
- PrfA transcription factor, 110, 111f
- Primary structure, protein, defined, 126, 130b
- Primase, 271  
DNA helicase interaction, 271, 284, 285f, 286, 287
- Primer  
in DNA replication  
extension of, 259  
removal, 271–272, 271f  
structure of, 258–259  
synthesis, 270–271
- in PCR procedure, 158, 159f
- priming protein, 303
- Primer:template junction, 258–259, 258f, 263–265, 264f, 267, 285f, 286

- pri-miRNA, 715–717, 717f  
 Primitive streak, 826  
 Principle of Independent Assortment, 8, 9f  
 Principle of Independent Segregation, 6–7, 7f  
*proα2* collagen gene (chicken), 467  
 Probe, hybridization, 151–153, 152f  
*proboscipedia (pb)* gene (*Drosophila*), 764b  
 Processed pseudogenes, 416  
 Processivity  
     degree of, 265  
     DNA helicase, 272  
     DNA polymerase, 265–267, 267f, 277–281, 281f  
 Proflavin, 324, 324f  
 Programmed rearrangements, 389  
 Prokaryotes. *See also* Bacteria; specific species  
     chromosome diversity, 200, 201t  
     coupling of transcription and translation, 520–521, 520f  
 DNA compaction, 200  
 DNA gyrase of, 97  
 DNA polymerases, 278t  
 factors that catalyze recombination steps, 351t  
 gene density, 204  
 genome size, 203, 203t  
 messenger RNA, 511f, 512  
 ribosome, 521–522, 522f  
 RNA polymerases, 431, 431t, 432f  
 supercoiling of DNA, 230b–231b  
 transcriptional regulation in, 615–653  
 translation initiation, 528–530, 530f  
 translation regulation, 549–555, 551f, 554f, 555f  
 Proline, 124  
 Promoter  
     –10 region, 435–438, 435f, 436b, 438f–441f  
     –35 region, 435–438, 435f, 436b, 438f–441f  
     activator regulation of, 616–618, 617f, 618f  
     alternatives, 485, 630, 631f  
     bacterial, 434–442, 435f–440f  
     bacteriophage λ, 638, 638f, 639f, 640, 642–643, 645, 645f, 647, 647f, 650–651  
     consensus sequence, 435–436, 436b  
     constitutive expression, 617  
     description, 432, 658  
     eukaryotic core, 448, 448f  
     flipping by site-specific recombination, 389  
     hsp70, 823  
     lac, 620–623, 620f–623f  
     LTR, 403  
     MerR twisting of DNA, 632–633, 632f  
     merT promoter, 618  
     phage, 630, 631f, 633  
     Pol I, 462, 462f  
     Pol III, 463, 463f  
     RNA polymerase binding, 432, 437–438  
     Tn10, 410, 410f  
 Promoter escape, 434, 442, 449–451, 633  
 Promoter proximal elements, 449  
 Proofreading, 315  
     by aminoacyl-tRNA synthetases, 518–519  
     by DNA polymerase, 260, 267–269, 269f  
     by ribosome, 537–538  
     by RNA polymerase, 444–445, 456  
 Proofreading exonuclease, 268–269, 269f, 315  
 Propeller twist, in DNA structure, 86, 87f  
 Prophage, 636, 798, 799f, 800  
 Prophase, 214, 216f  
 Protein(s)  
     fusion, 154  
     gene control over amino acid sequence in, 31  
     modification states, 181–182  
     peptide bonds hydrolysis, 68  
     structure (*see* Protein structure)  
     synthesis  
         adaptor hypothesis of Crick, 34  
         direction of, 38–39  
         ribosomes as location for, 34, 34f  
         riboswitch control of, 112  
     techniques, 173–179  
         immunoblotting, 176–177, 177f  
         purification from cell extracts, 173–174  
         separation  
             affinity chromatography, 175–176  
             gel-filtration chromatography, 174–175, 174f  
             ion-exchange chromatography, 174, 174f  
             polyacrylamide gels, 176, 176f  
             sequencing, 177–179, 178f  
             Edman degradation, 177–178, 178f  
             tandem mass spectrometry (MS/MS), 178–179, 180f  
 Protein–DNA interactions  
     cooperative binding in, 617, 619f  
     in replication initiation, 293, 295  
     weak bonds in, 62–63  
 Protein folding  
     Anfinsen Experiment, 134, 135b  
     domains, 130–134, 131f  
 Protein–protein interactions  
     interactomes, 182, 183f  
     at replication fork, 286–288  
     in replication initiation, 293, 295  
     weak bonds in, 62–63  
     yeast two-hybrid assay for determining, 182  
 Protein–protein interfaces, 140–141, 140f  
 Protein structure, 121–144  
     activity regulation and, 142–143, 142f  
     amino acids  
         hydrophilic and hydrophobic side chains, 125  
         with special conformational properties, 124–125, 125f  
     structure of, 121–122, 122f  
     coiled-coil, 129, 129f  
     conformational changes, 136–137  
     disulfide bonds, 124–125, 124b  
     domains, 130–134, 131f  
         classes of, 132, 132f  
         closure, 136f  
         immunoglobulin G example, 133–134  
         linkers and hinges, 133  
         post-translational modifications, 133–134  
     enzyme, 141–142  
     folding, 134, 135f  
     levels of, 126–129, 127f  
         primary structure, 126, 127f  
         quaternary structure, 127, 127f, 129, 129f  
         secondary structure, 126–127, 127f, 128f  
         tertiary structure, 127, 127f  
     molecular recognition and, 137–141  
         DNA recognition, 137, 138f, 139–140, 139f, 140f  
         protein–protein interfaces, 140–141, 140f  
         RNA recognition, 141, 141f  
     peptide bond, 122–123, 123f  
     polypeptide chains, 123–124, 124b  
     predicting from amino acid sequence, 135–136  
     Ramachandram plot, 124b  
     random coils, 126  
     water effect on, 125–126  
 Protocells, 593  
     self-replicating, 603–606  
 Protonated phosphamidines, 158f  
 Proton sensor, 112  
 Proton shuttle, 541, 541f  
 Prp22, 477  
 Pseudogenes, 206, 206f, 207f, 416  
 Pseudoknots, 109, 110f, 112  
 Pseudouridine ( $\psi$ U), 514, 514f  
 Pseudouridine loop ( $\psi$ U loop), of tRNA, 514–515, 514f, 515f  
 P site, 525–527, 525f–528f, 528, 529, 531, 531f, 535, 537, 537f–540f, 540–543  
 PstI, 150, 150f  
 P-TEF (positive transcription elongation factor), 669–670, 671f  
 PTEN, 727b  
 Pulse chase experiment, 37

Pulsed-field gel electrophoresis, 149, 149f  
 Pulse-labeling, 39  
 Purification, protein, 173–174  
 Purines, 80, 80f  
 Puromycin, 552b–553b  
 Pyrazinamide, 565b  
 Pyrimidine dimers, 325  
 Pyrimidines, 80, 80f, 597f  
 Pyrophosphatase, 259f, 260  
 Pyrophosphate bonds, 66, 67t, 73–74, 259f, 260  
 Pyrophospholytic editing, 444  
 Pyruvate kinase, 131f

**Q**

QBE sequence, 650, 650f  
 Q proteins, 649–651  
 Quantum mechanics, 52–53  
 Quaternary structure, protein, 127, 127f, 129, 129f, 130b  
 Quorum sensing, 635, 635b

**R**

Rad51, 359, 360f, 365f, 366–367, 366f, 367b, 370  
 Rad52, 366–367, 370  
 RadA, 359, 359f  
 Radiation, DNA damage from, 322–323  
 Radioactively labeled DNA, 152  
 RAG1, 332b, 418, 419, 420, 770  
 RAG2, 332b, 418, 770  
 Ramachandran plot, 124b  
 Ran, 505  
 Random coils, 126  
 Rap1, 688, 688f  
 Rat1, 461  
 Rb (retinoblastoma protein), 686  
 RB2, 727b  
 RBS. *See* Ribosome binding site  
 RdRP (RNA-dependent RNA polymerase), 713–714  
 Reading frames, 510–511, 511f  
 RecA, 355–359  
     assembly, 355–356, 357f  
     base-paired partners within, 356–359  
     homologs, 359, 360f  
     joint molecule, 358  
     mechanism of action, 356–359, 357f  
     in SOS response, 335  
     ssDNA-coating, 354  
     stimulation of proteolytic autocleavage by, 642–643  
     structure, 355–356, 356f, 358f, 359f  
     substrates for, 355f  
 RecBCD helicase/nuclease, 349, 351–355, 352f–354f  
 Recessive, 6–7  
 Recognition helix, 137, 624–625, 625f  
 Recombinant DNA, bacteriophage, 801–802  
 Recombinants

chromosome mapping, 11–13, 12f  
 defined, 8  
 Recombinase recognition sequences, 379, 380f  
 Recombinases  
     covalent-intermediate formation, 380–381, 381f  
     DDE-motif transposase/integrase proteins, 404  
     description, 377  
     Hin, 389–391, 390f  
     λ integrase, 387–389, 388f  
     mechanism of action, 380–381, 381f  
     serine, 380, 381f, 381t, 382–383, 382f, 384f  
     tyrosine, 380, 381f, 381t, 383–385, 385f  
     XerCD, 391f, 392b  
 Recombination  
     bacteriophage, 802  
     classes of, 378f  
     FLP-FRT system, 823–824, 823f  
     hot and cold spots, 373  
     mating-type switching, 369–371, 369f, 372f  
     meiotic, 363–369, 363f, 366f, 373  
     site-specific, 377–393  
     transposition, 377, 393–416  
     V(D)J, 416–420  
 Recombinational repair, 325, 330–331  
 Recombinational transformation in yeast, 810, 810f  
 Recombination factories, 366, 366f  
 Recombination signal sequences, 418–420, 418f  
 Recombination sites, 378–379, 378f, 379f, 380f  
 RecQ helicases, 367, 368f  
 Recruitment, of RNA polymerase, 617, 617f  
 Refolding, protein, 134, 135b  
 Regulation by RNA, 701–731  
     in bacteria, 701–711  
         attenuation, 702, 704–705, 707b–708b  
         CRISPRs, 706, 709–710  
         riboswitches, 703–705, 703f–706f  
         small RNAs (sRNAs), 701–702, 703f  
     in eukaryotes, 711–730  
         chromatin modification, 719–720, 721f  
         long non-coding RNAs (lncRNAs), 728  
         miRNAs, 712–718, 712f, 715f–717f, 722b–723b, 727b  
         RNA interference (RNAi)  
             efficiency of, 713  
             evolution of, 721  
             history of, 722b–723b  
             human disease and, 727b  
             overview, 712–714, 712f

as tool for manipulating gene expression, 725–726, 726f  
 siRNA, 712–714, 712f, 714f, 718, 725  
 X-inactivation, 728–730, 729f, 730f  
 Regulator binding sites, 658  
 Regulatory circuits, 776–790  
     autoregulation, negative, 777, 777f  
     autoregulation, positive, 777f, 779–780, 780f, 782b  
     bistable switches, 780–784, 780f, 782b, 784f  
     evolution of, 683b  
     feed-forward loops, 784–786, 785f, 787f  
     AND gate logic, 776, 776f  
     nodes and edges, 776, 776f  
     noise, 777–779, 778f  
     oscillating, 786–790, 788f  
     robustness, 779  
     stochasticity, 778, 779  
     synthetic circuits, 789–790  
 Regulatory DNA sequences  
     description, 206, 658  
     identification of, 171–172  
     transcription, 449  
 Relaxed DNA, 95–96, 95f  
 Release factors (RFs), 544, 545f, 547–549, 547f, 577  
 Renaturation, of DNA, 90  
 Repeated sequences, 709  
 Repetitive DNA, 207–208  
     in telomeres, 209, 211f  
 Replicase ribozyme, 602, 603–606, 604f  
 Replication bubble, 290b–292b  
 Replication fork, 269–277  
     Dam methylation at, 319f  
     description, 269–270, 270f  
     DNA synthesis at, 283–288  
     enzymes active at, 275–277, 277t  
     initiation of new strand, 270–271  
     lagging strand, 270, 270f, 284, 285f, 286  
     leading strand, 270, 270f, 284, 285f, 286  
     primer removal, 271–272, 271f  
     primer synthesis, 270–271  
     ssDNA stabilization, 273–274, 275f  
     topoisomerase action at, 275, 276f  
     unwinding of DNA, 272–273, 272f, 275, 276f, 287f  
 Replicative transposition, 401–403, 402f  
 Replicator, 288, 289, 289f, 290b, 292, 297, 297f  
 Replicon model, 288, 288f  
 Replisome, 287–288  
 Reporter gene, 659, 805–806  
 Reporter plasmid, 660, 661f  
 Repressilator, 789–790  
 Repressor(s). *See also specific repressors*  
     allostery and, 618  
     of bacteriophage λ, 137, 138f, 139

- binding to operator, 617  
 description, 616  
 effect on RNA polymerase, 633  
 eukaryotic, 681–682, 682f  
*Hunchback*, 754, 755f, 756, 757f, 758, 759f, 760–761  
*Krüppel*, 762, 762f  
 negative autoregulation, 643–644  
 positive autoregulation, 643  
*Runt*, 757b  
 short-range transcriptional, 761–762  
*Snail*, 750–751  
 of transcription, 616  
**Resolution**, in homologous recombination, 344, 346, 347f, 350b, 361  
**Resolvases**, 391  
**Restriction endonucleases**, 149–151, 150–152, 150f, 150t, 151f  
**Retinitis pigmentosa**, 497b  
**Retroregulation**, 651, 652f  
**Retrotransposons**  
 poly A  
   genetic organization, 395f, 396–397  
   reverse splicing mechanism, 405–406, 407f  
 virus-like  
   genetic organization, 395f, 396  
   mechanism of transposition, 403, 404f  
**Retroviruses**  
 cDNA formation, 403, 404f  
 integration of, 403, 404f  
 murine leukemia virus (MLV), 112  
 virus-like retrotransposons compared, 396  
**Rett syndrome**, 696b  
**Reverse genetics**  
 in *Arabidopsis*, 813  
 tilling strategies, 814  
**Reverse gyrase**, 229  
**Reverse (back) mutations**, 584  
**Reverse transcriptase**, 396  
 in cDNA library construction, 156, 157f  
 pseudogenes and, 206, 207f  
 of retrotransposons and retroviruses, 396  
 TERT (telomerase reverse transcriptase), 305  
**Reverse transcription**  
 of retrotransposons and retroviruses, 403, 406, 407f  
 target-site-primed, 405–406, 407f  
**RFs (release factors)**, 544, 545f, 547–549, 547f  
**R group**, amino acid, 121, 122f  
**Rho-dependent terminators**, 445–446, 446f, 447f  
**Rho factor**, 445–446, 446f  
**Rho-independent terminators**, 445–447, 446f, 447f  
*rhomboid* gene, 750–751, 750f, 751f  
 “Ribbon diagrams,” 130b, 131f  
**Ribonuclear protein (RNP) motif**, 141  
**Ribonuclease**, 23  
**Ribonuclease A**, 134, 135b  
**Ribose**, 33, 33f, 107  
**Ribosomal proteins**, regulation of translation, 551, 553–555, 554f, 555f, 556f  
**Ribosomal RNA (rRNA)**, 35, 525, 528  
**Ribosome binding site (RBS)**, 511f, 512, 528, 528f, 549, 551, 551f, 553, 554, 704  
**RNA secondary structure** and, 110, 111f  
**Ribosome cycle**, 522–523, 523f  
**Ribosome profiling**, 561b–562b  
**Ribosome recycling factor (RRF)**, 548, 548f  
**Ribosomes**, 519–528  
 aminoacyl-tRNA delivery and binding, 537, 538f  
 association and dissociation cycle, 522–523, 523f  
 channels, 527, 527f  
 cycle, 522–523, 523f  
 decoding center, 521, 525, 527  
 discrimination of aminoacyl-tRNAs, 519, 537–538, 539f  
 description, 510, 519–521  
 factor-binding center, 537, 538f  
 peptide bond formation, 524, 524f, 538, 540–541, 541f  
 peptidyl transferase center, 521, 525–527, 547  
 peptidyl transferase ribozyme, 118  
 polyribosomes, 35, 36f, 522–523, 523f  
 prokaryotic mRNA recruitment to, 528, 528f  
 proofreading, 537–538  
 recruitment to mRNA by 5' cap, 530–532  
 recycling, 548–549, 550f  
 as ribozyme, 538, 540–541  
 RNA in, 108  
 rRNA functions, 525  
 scanning by, 512, 530, 535  
 sedimentation velocity, 521, 521f  
 as site of protein synthesis, 34, 34f  
 stalled, 563, 564f  
 structure, 35, 521–522, 522f, 524–527, 525f–528f  
 translocation, 522, 537, 541–543, 542f  
 tRNA-binding sites, 525–527, 525f–528f  
 tRNA discrimination, 519  
**Riboswitch**, 702, 703–705, 703f–706f  
 control of protein synthesis by murine leukemia virus, 112  
 described, 114  
 structure prediction, 113–114, 113f  
**Ribozyme**, 114, 116–118, 116f, 117f, 479b, 538, 540–541  
 discovery of, 599  
 replicase, 602, 603–606, 604f  
 self-replicating, 599–603  
**Richardson**, Jane, 130b, 131f  
**Ricin**, 110  
**rII locus**, of T4, 808  
**RISC**. *See* RNA-induced silencing complex (RISC)  
**RITS** (RNA-induced transcriptional silencing) complex, 720, 721f  
**R-loop mapping**, 471b–472b  
**RNA**  
 directed evolution, 114, 114f, 115  
 DNA compared, 32–33, 33f  
 functions of, 108, 115  
 precursors of, 71–72  
 protein recognition of, 141, 141f  
 regulatory functions (*see* Regulation by RNA)  
 separation by electrophoresis, 149  
 structure (*see* RNA structure)  
 synthesis, location of, 35–37, 38f  
 transcription (*see* Transcription)  
**RNA-dependent RNA polymerase** (RdRP), 713–714  
**RNA editing**  
 deamination, 500–501, 503b  
 guide RNAs, 501–503, 502f  
**RNA-induced silencing complex (RISC)**, 713–714, 714f, 718  
**RNA-induced transcriptional silencing (RITS) complex**, 720, 721f  
**RNA interference (RNAi)**  
 in *Arabidopsis*, 814  
 in *C. elegans*, 818–819  
 discovery of, 722b–723b  
 efficiency of, 713  
 evolution of, 721, 724  
 human disease and, 727b  
 overview, 712–714, 712f  
 in *Schizosaccharomyces pombe*, 713, 719–720, 721f  
 as tool for manipulating gene expression, 725–726, 726f  
**RNA ligase ribozyme**, 601, 601f  
**RNA polymerase**  
 abortive synthesis, 442  
 action of, 36–37, 37f  
 allostery and, 618, 618f, 631  
 bacteriophage T7, 443b  
 carboxy terminal domain, 623, 623f, 624b  
 discovery of, 36  
 eukaryotic, 431, 431t, 432f  
 holoenzyme, 434, 435f  
 initiation of RNA chain, 440–441, 441f  
 initiation of transcription, 429

- RNA polymerase (*Continued*)  
 isomerization, 438–440  
*lac* operon and, 622–623, 623f  
 phosphorylation, 669  
 primase, 271  
 processivity, 442  
 prokaryotic, 431, 431t, 432f, 435f  
 promoter binding, 432, 437–438  
 proofreading, 444–445, 456  
 recruitment, 617, 617f  
 RNA polymerase I (Pol I), 431, 431t, 462–463, 462f  
 RNA polymerase II (Pol II), 431, 431t, 448–462  
 RNA polymerase III (Pol III), 431, 431t  
 RNA polymerase III (Pol III) enzyme, 414, 431, 431t, 462, 463, 463f  
 RNA polymerase IV (Pol IV), 431  
 RNA polymerase V (Pol V), 431  
 $\sigma$  factors, 434–438, 435f, 437f–440f, 442  
 single-subunit, 443b  
 stalled, 669  
 structure, 430–432, 431t, 432f  
 transcription-coupled repair, 329, 330f  
 in transcription process, 432, 433f, 434  
 translocation models, 441–442, 441f  
 RNA polymerase ribozyme, 599–603  
 RNA primer  
   removal by RNase H, 271–272, 271f  
   synthesis by primase, 271  
 RNA processing, 457–460  
   capping, 458, 459f  
   polyadenylation, 458–460, 460f  
   splicing, 457  
 RNA-recognition motif (RRM), 141, 492  
 RNA:RNA hybrids, 474f  
 RNase domain  
   of Argonaute, 718, 719f  
   of Dicer, 717, 718f  
 RNase H, 271, 271f  
 RNase III family, 717  
 RNase P, 114, 116, 116f  
 RNA splicing, 467–500  
   of adenovirus, 471b–472b  
   alternative splicing, 469, 483–496  
   chemistry of, 469–470, 473  
     splice sites, 469–470, 469f  
     splicing reaction, 470, 470f, 471f, 473, 474f  
     transesterification, 470, 480  
   classes of, 477t  
   defects, 497b–498b  
   description, 205, 457, 468, 468f  
   discovery of, 471b–472b  
   errors in, 481, 481f  
   exon shuffling, 497–500  
   pathways, 474–482  
   of pre-mRNAs, 468–469, 468f, 480f, 484f  
     ribozyme role in, 116  
     schematic of, 205f (*See also* Introns)  
     by self-splicing introns, 477–480, 478f, 479b, 480f  
   spliceosome  
     alternative (minor), 483, 483f, 486, 487f  
     splice site selection, 480–482, 481f, 482f  
     splicing reaction and, 474–476, 475f  
     structure, 473–474  
     variants, 482  
 RNA structure, 107–118, 108f  
   description, 32–33, 33f  
   DNA compared, 107–108  
   double-helical characteristics, 108–110  
     G:U base pairing, 109, 110f  
     pseudoknot, 109, 110f, 112  
     stem-loop structures, 108–109, 109f, 110f  
     tetra loop, 109, 109f  
   features of, 107–108, 108f  
   prediction of, 111–114, 113f  
   of ribozymes, 114, 116–118, 116f, 117f  
     tertiary structures, 110–111, 111f  
 RNA World Hypothesis, 118, 599  
 Robustness, 779  
 RPA, 366  
 Rpd3, 681  
*rpoS* gene (*Escherichia coli*), 702  
 RpsA, 565b  
 RRF (ribosome recycling factor), 548, 548f  
 rRNA (ribosomal RNA), 35, 525, 528  
 RSC, 671, 671f  
 RS domain, 492  
 Rtt103, 461  
 Runt repressor, 757b  
 Rut sites, 445  
 RuvAB complex, 359–360, 360f  
 RuvC, 361, 361f  
 Ruvkun, Gary, 722b  
 RyB RNA, 702
- S**
- Saccharomyces cerevisiae*  
   centromere size and  
     composition, 211f  
   chromosome makeup, 201t  
   combinatorial control of the  
     mating-type genes, 680–681, 680f  
   Gal4, 660, 660f, 661f  
   GAL genes, 681, 682f  
   gene density, 203t, 204, 204f, 205t  
   genome size, 203t, 810  
   HO gene, 675–676, 677f  
   interactome, 182, 183f  
   introns, 205  
 life cycle, 808, 809f  
 mating type, 738–740, 739f, 740f  
 mating-type switching, 369–371, 369f, 372f  
 as model organism, 808–811  
   cell shape changes, 810–811, 811f  
   diploid state, 809, 809f  
   genetic analysis, 809  
   genome size and characterization, 810  
   haploid state, 809, 809f  
   mutation generation, 810  
 nucleosome positioning, 241  
 repetitive DNA, 205t, 208  
 replication fork enzymes, 277t  
 replicators, 288, 288f, 292  
 RNA polymerase, 432f  
 silencing in, 688–689, 688f  
 telomere-binding proteins, 307–308, 308f  
 transposon occurrence and distribution, 394f  
   Ty elements, 414, 414f, 415f  
 Safe havens, for transposons, 409, 414  
 SAGA (Spt-Ada-Gcn5-acetyltransferase) complex, 452, 669  
*Salmonella*  
   in Ames test, 321b  
   flagella, 389, 389f  
   Hin recombinase, 389–391, 390f  
 SAM (*S*-adenosylmethionine)-sensing riboswitch, 703–704, 704f, 705f  
 SANT domain, 244  
 Sau3A1, 150, 150t  
 SBF, 675–676, 677f  
 SC35, 492, 493f  
 Scaffolds, 165–167  
 Scanning, 530, 535  
*Schizosaccharomyces pombe*  
   centromere size and composition, 211f  
   chromosome makeup, 201t  
   gene density, 203t  
   genome size, 203t  
   RNAi in, 713, 719–720, 721f  
 Schnurri, 756b  
 Scott-Moncrieff, Rose, 16  
 Scr gene, 764b  
 Scrunching model, 441–442, 441f  
 SDSA (synthesis-dependent strand annealing), 371, 372f  
 SDS gel electrophoresis, 176, 176f  
 Sea anemones, 771–772, 771f  
 Sea squirts, 740–741, 742f, 769–770, 770f  
 Sea urchin genome, 770  
 SEC (Super Elongation Complex), 669, 670b  
 Secondary structure, protein, 126–127, 127f  
 $\beta$ -sheet, 126–127, 128f

- defined, 130b  
 $\alpha$ -helix, 126, 128f  
 Selective reagents, 177  
 Selective stickiness, 61  
 Selector sequence, 489f, 490, 490b–491b  
 Selenocysteine, 520b  
 SELEX (systematic evolution of ligands by exponential enrichment), 114, 114f, 115, 189, 189f  
 Self-replicating protocells, 603–606  
 Self-replicating ribozymes, 599–603  
 Self-replication, 594  
 Self-splicing introns, 477–480, 478f, 479b, 480f  
 SeqA, 294b  
 Sequenators, 162, 163b, 168  
 Sequence coverage, 10x, 162  
 Sequencing  
   DNA  
 454 sequencing machine, 167–168, 167f  
 chain-termination method, 160–162, 161f, 163b  
 contigs, 165–167, 165f, 166ff  
 gel, 162f  
 high-throughput, 163b  
 of human genome, 164–168  
 paired-end strategy, 165–167, 166f  
 principle of nested sets of fragments, 159–162  
 readout of sequence, 163f  
 scaffolds, 165–167  
 Sequenators, 162, 163b, 168  
 sequence coverage (10 $\times$ ), 162  
 shotgun, 162–167, 164f–166f  
 proteins, 177–179, 178f  
   Edman degradation, 177–178, 178f  
   tandem mass spectrometry (MS/MS), 178–179, 180f  
 Sequencing machines, 162, 163b, 167–168, 167f  
 Serine recombinases, 380, 381f, 381t, 382–383, 382f, 384f  
   Hin recombinase, 389–391, 390f  
 Serine, tRNAs for, 518  
 Serine-tRNA synthetase, 520b  
 70S initiation complex, 530, 530f  
*sex combs reduced (Scr)* gene (*Drosophila*), 764b  
*sex-lethal (Sx1)* gene (*Drosophila*), 493–494, 494f, 495f  
 Sex-linked gene, 10  
 SF2/ASF splicing regulator, 485, 492  
 SHAPE (selective 2'-hydroxyl acylation analyzed by primer extension) procedure, 113  
*shavenbaby* gene, 754  
 Shh (Sonic hedgehog), 744–745, 745f  
 Shine-Dalgarno sequence, 512  
 Short hairpin RNA (shRNA), 726  
 Shotgun sequencing, 162–167, 164f–166f
- Sickle-cell anemia, 31  
 $\sigma$  factors, 434–438, 435f, 437f–440f, 442  
   alternative, 630, 631f, 701  
   autoregulation, 780, 787f  
   heat shock, 630  
 Signaling molecules  
   cell-to-cell contact and, 736, 736f  
   secreted, 735f, 737–738, 737f  
 Signal integration, 627, 675–676, 677f  
 Signal transduction pathway, 682–686, 685f, 736, 736f  
 Silencers, 449  
 Silencing  
   by DNA methylation, 692, 693f  
   heterochromatic, 673, 687  
   by histone modification, 681, 688–690, 688f, 690f, 691b  
   RNA interference, 712–714, 718–726  
   in yeast, 688–689, 688f  
 Silent cassettes, 370  
 Simpson, George Gaylord, 15  
 SINE (short interspersed nuclear element), 415, 415f  
 Single-step growth curve, 800–801, 800f  
 Single-stranded DNA (ssDNA)  
   in homologous recombination process, 346, 349, 351, 354  
   Rad51 and, 370  
   RecA, 354, 355–358, 358f, 359f, 367  
 Single-stranded DNA-binding proteins (SSBs), 273–274, 275f, 354  
*Sinorhizobium meliloti*  
   chromosome makeup, 2011  
   gene density, 203t  
   genome size, 203t  
 SinR, 783, 784f  
 SIR (silent information regulator), 688  
 Sir2, 688–689, 688f  
 siRNA, 431, 712–714, 712f, 714f, 725, 814  
 Sister-chromatid cohesion, 211, 213, 214, 217, 219  
 Sister chromatids, 211, 363, 363f  
 Sister-chromatid separation, 212, 213f  
 Sister chromosomes, 209  
 Site-directed mutagenesis, 157  
 Site-specific recombination. *See* Conservative site-specific recombination (CSSR)  
 6-mercaptopurine (6-MP), 268b  
 6S RNA, 701  
 Skin-nerve regulatory switch, 743–744, 744f  
 SL1, 462f, 463  
*Sleeping Beauty* element, 414  
 Slicer, 713, 722b  
 Sliding clamp loaders, 281, 282b–283b, 283, 295, 334f, 336, 337  
 Sliding DNA clamp, 278–283, 279f–281f, 295, 296f  
 Sliding of DNA, 237–238, 238f  
*Slo* gene (human), 484
- SlrR, 783–784, 784f  
 Small interfering RNAs (siRNAs), 431, 712–714, 712f, 714f, 725, 814  
 Small nuclear ribonuclear proteins (snRNPs), 473–474, 474f, 475f, 476  
 Small nuclear RNAs (snRNAs), 473–474, 476  
 Small RNAs (sRNAs), 701–702, 703f.  
   *See also specific types*  
 SMC proteins, 214, 215f, 234  
 SMN (survival motor neuron), 497b–498b  
*snail* gene (*Drosophila*), 786  
 Snail protein, 750–751  
 Snapdragon, inheritance of flower color in, 8, 8f  
 Sodium dodecyl sulfate (SDS), 176, 176f  
*sog* gene, 746, 750, 750f, 751f  
 Solenoid model, nucleosome, 232–233, 233f  
 Somite gene expression, 788–789, 789f  
 Sonic hedgehog (Shh), 744–745, 745f  
 SOS response, 335, 643  
 Southern blot hybridization, 152–153, 152f  
 Southern, Edward, 152  
 Sox2, 663  
*SoxB2* gene, 172f  
 SP1 activator, 316b, 664, 670  
 Spacer sequences, 709, 710, 710f  
 Spätzle, 747, 747f  
 Specialized transduction, 802, 804  
 S phase, 297  
 Spinach, 115  
 Spinal muscular atrophy, 497b–498b  
 Spindle pole bodies, 211  
 Spiral writhe, 94, 96  
 Spliceosomal protein–RNA complex, 141f  
 Spliceosome  
   alternative (minor), 483, 483f, 486, 487f  
   assembly/disassembly, 476–477  
   splice site selection, 480–482, 481f, 482f  
   splicing reaction and, 474–476, 475f  
   structure, 473–474  
 Splice recombination products, 346, 347f  
 Splice sites, 469–470, 469f, 474, 476, 486  
 Splice site selection, 480–482, 481f, 482f  
 Splicing. *See* RNA splicing  
*SPO11*, 363–365, 364f, 365f  
 Sporophyte, 812, 812f  
 Sporulation, 743, 743f  
 SPT5, 456, 458  
 Spt16, 457  
 SRNAs (small RNAs), 701–702, 703f  
 SR proteins, 482, 482f, 491–492, 504  
 SSBs. *See* Single-stranded DNA-binding proteins  
 ssDNA. *See* Single-stranded DNA

- SsrA RNA, 563, 564f, 565b  
 SSRP1, 457  
 Stadler, L.J., 16  
 Stahl, Franklin W., 27–30, 342, 807  
 Start codon, 510, 528, 530, 533b, 535  
 Stationary phase, of bacterial growth, 803, 803f  
 STAT (signal transducer and activator of transcription) pathway, 684, 685f  
 Steady state, 779  
 Stem cells  
   germline (GSC), 755b–756b  
   induced pluripotent stem (iPS) cells, 666b, 733–735, 734b  
   mouse embryogenesis and, 826  
 Stem-loop structures, 108–109, 109f, 110f  
 Stereoisomers, of amino acids, 61  
 Steric hindrance, in mutually exclusive splicing, 486, 486f  
 Stochasticity, 778, 779  
 Stop codons, 40, 510, 554, 583, 584–586, 585f, 587–588  
 Stop mutation, 582  
 Strand-exchange proteins, 344, 355, 359  
 Strand exchange, substrates for, 355f  
 Strand invasion, 344, 345f, 365f, 370, 372f  
*Streptococcus pneumoniae*  
   gene density, 203t  
   genome size, 203t  
 Stress, plant response to, 815  
 Structural maintenance of chromosome (SMC) proteins, 214, 215f, 234  
 Sturtevant, Alfred H., 10, 13, 821  
 Su(H), 744, 745f  
 Su(Var)3-9 gene/protein, 689–690  
 Sulston, John E., 816  
 Supercoiling, 94  
   DNA helicase production of, 275, 276f  
 DNA migration in gel electrophoresis, 102, 102f  
 linking difference, 95–96  
 negative, 94f, 96, 228–229, 230b  
 positive, 96, 229, 230b  
 reduction by topoisomerase at replication fork, 275, 276f  
 removal of, 95, 97  
 writhe cccDNA, 95f  
 Superhelical density, 96, 229, 230b–231b  
 Suppressor genes, 584  
 Suppressor mutations, 584–587, 584f, 585f  
   of frameshift mutation, 584f  
   intergenic suppression, 584–585  
   intragenic suppression, 584, 584f  
   nonsense, 584–586, 585f  
 Sutton, Walter S., 8  
 SV40 T-antigen, alternative splicing of, 485, 485f  
 Svedberg (*S*), 521  
 Svedberg, Theodor, 521  
 SWI5, 675–676, 677f  
 SWI/SNF, 669, 672, 676  
 Switch  
   bistable, 780–784, 780f, 784f  
   nodes and edges, 776, 776f  
   two-node, 776  
 Synapsis, 9, 11, 11f  
 Synaptic complex, 377, 397, 398f, 405  
 Syn conformation, glycosidic bond, 87, 89, 89f  
 Syncytium, 746  
 Synergy, of activators, 675  
 Synonyms, codons as, 573  
 Synteny, 770–772, 826  
 Synthesis (S phase), 211, 212f  
 Synthesis-dependent strand annealing (SDSA), 371, 372f  
 Synthetic biology, 775  
 Synthetic circuits, 789–790  
 Systematic Evolution of Ligands by Exponential Enrichment (SELEX), 600  
 Systems biology, 775–791  
   autoregulation  
    negative, 777, 777f  
    positive, 777f, 779–780, 780f, 782b  
   bistable switches, 780–784, 780f, 782b, 784f  
   feed-forward loops, 784–786, 785f, 787f  
   AND gate logic, 776, 776f  
   nodes and edges, 776, 776f  
   noise, 777–779, 778f  
   oscillating, 786–790, 788f  
   overview, 775–776  
   regulatory circuits, 776–790  
   robustness, 779  
   stochasticity, 778, 779  
   synthetic circuits, 789–790
- T**
- T7 DNA polymerase, 264f  
 TAFII30, 316b  
 TAFs (TBP-associated factors), 449, 452, 452t, 669  
*Tam* elements, 408f  
 Tandem mass spectrometry (MS/MS), 178–179, 180f  
 Target DNA, of transposons, 397  
 Target immunity, transposition, 409, 411, 413b  
 Target-site duplications, 395f, 396, 403  
 Target-site-primed reverse transcription, 405–406, 407f  
*tasi* (*trans*-acting, short interfering RNA), in *Arabidopsis*, 814  
 TATA element (or box), 448–449, 448f, 450f, 452, 452f, 462  
 TAT protein, 670  
 TAT-SF1, 458  
 Tatum, Edward, 21  
 Tautomeric states, of DNA bases, 80–81, 81f  
 Tautomers, 267, 267f  
 TBP (TATA-binding protein), 448f, 449, 451, 452f, 452t, 462–463  
*Tc1/mariner* elements, 411–414  
 T-cell receptors, 416  
 T-DNA (transfer DNA), 813  
 Telomerase, 209, 305–309, 306f, 307f  
 Telomerase reverse transcriptase (TERT), 305  
 Telomere-binding proteins, 307–309, 308f, 309f  
 Telomeres  
   cellular aging and, 307b  
   description of, 209, 211f, 304  
   length regulation, 307–308, 309f  
   location, 208f, 209  
   protection of, 308–309, 310f  
   repetitive DNA in, 209, 211f  
   replication, 304–306, 306f, 307f  
   silencing at yeast, 688, 688f  
 Telophase, 216f, 217  
 Temperate phage, 798  
 Template, for DNA synthesis, 258, 266f  
 TER, 305  
 Terminal uridylyl transferase (TUTase), 503  
 Termination  
   transcription, 432, 433f, 434, 445–446, 447f, 460–462, 461f  
   translation, 544–549  
 Terminators, 445–446, 446f, 447f  
 Ternary complex, 531, 559, 561  
 TERT (telomerase reverse transcriptase), 305  
 Tertiary structure, protein, 127, 127f, 130b  
*Tetrahymena*  
   chromosome makeup, 200, 201t  
   gene density, 203t  
   genome size, 203t  
 Tetraloop, 109, 109f  
 TFIIA, 449, 452t  
 TFIIB, 449, 452–453, 452f, 452t  
 TFIIB recognition element (BRE), 448, 448f  
 TFIID, 316b, 448f, 449, 453, 665f, 668–669  
 TFIIE, 449, 452t, 453  
 TFIIF, 449, 452–453, 452t  
 TFIIH, 329, 449, 451, 452t, 453, 454, 669  
 TFIILA, 463  
 TFIIB, 463, 463f  
 TFIIC, 463, 463f  
 TFIIS, 445, 455, 456, 457f  
 T:G mismatch, 328  
 Thermodynamics  
   first law of, 53

- second law of, 54
- Thermophiles**  
positive supercoiling, 229  
positive supercoiling of DNA, 96  
reverse gyrase, 229
- Thermus aquaticus*, RNA polymerase of, 432f, 435f
- Thiogalactoside transacetylase, 621
- 30-nm fiber, 232–234, 233f, 234f, 244
- Three-factor crosses to assign gene order, 12, 12f
- Three-node network, 784–786, 785f
- 3' splice site, 469–470, 469f, 474, 476, 486
- 3' untranslated trailer region (3' UTR), 735, 735f
- Thymidine kinase (TK) marker gene, 828
- Thymine**  
base analog, 323, 324f  
base pairing, 81–82, 81f  
binding to adenine, 24  
Chargaff's rules, 26  
from deamination of 5-methylcytosine, 320f, 322  
structure, 25f, 27f, 80, 80f  
in tRNA, 514
- Thymine dimers, 322, 322f, 323b, 326f
- Tiling array, whole-genome, 169–171, 170f, 170
- Tilling, 814
- TIM barrel**, 131f
- Ti** (tumor-inducing) plasmid, 813
- Titin* gene (human), 467
- TK (thymidine kinase) marker gene, 828
- t-loop, 309, 310f
- tmRNA, 563, 564f
- Tn5, 400f, 405, 405f, 406f, 805
- Tn7, 399–400, 400f
- Tn10, 400, 400f, 409–410, 409f, 410f
- TnsA, 400
- Todd, Alexander, 24
- Toll, 747, 747f, 750
- Topoisomerases**, 97–101, 97f–101f  
catenation/decatenation, 98–99, 99f  
decatenation, 303, 303f  
DNA cleavage by, 99–101, 100f  
DNA gyrase, 97  
DNA rejoicing by, 99, 101  
DNA relaxation, 97, 100–101  
gyrase, 229  
knotted DNA, action on, 98–99  
linking number change by, 97, 98f, 101  
nuclear scaffold and, 231b, 234  
nucleosome assembly, 228–229, 230b–231b  
at replication fork, 275, 276f  
strand passage, 100–101  
type I, 97–101, 98f, 99f, 100f, 101f  
type II, 97–101, 97f, 99f, 234, 276f, 303, 303f
- Topology (fold), 130b, 132
- Toroid writhe, 94, 96, 230b
- Torpedo model of termination, 461, 461f
- Torsion angles, polypeptide chain, 123f, 124, 124f
- Totipotent, 746
- tra* gene (*Drosophila*), 494, 495f
- Trans-acting RNAs, 702, 713
- trans-acting, short interfering RNA (tasiRNA), in *Arabidopsis*, 814
- Transcription, 429–464, 430f  
accuracy, 429–430  
in bacteria, 434–447  
arrest of transcription, 445  
elongation, 442–445  
initiation, 440–442, 441f  
isomerization, 438–440  
polymerase translocation models, 441–442, 441f  
promoter escape, 442  
promoter features, 434–437, 435f  
 $\sigma$  factor-mediated binding, 437–438, 437f, 438f  
termination, 445–446, 446f, 447f  
in central dogma, 34  
coupling of transcription and translation, 520–521, 520f  
description, 36f  
DNA replication compared, 429–430  
in eukaryotes, 448–463  
elongation, 455–457, 455f, 458f  
histone removal and replacement, 456–457, 458f  
initiation, 449–454, 450f  
Mediator Complex, 453–454, 453f, 454f  
by Pol I, 462–463, 462f  
by Pol II, 448–462  
by Pol III, 462, 463, 463f  
preinitiation complex, 449, 450f, 453f  
processing of RNA, 457–460, 459f, 460f  
promoter escape, 449–451  
termination, 460–462, 461f  
initiation  
activator and repressor activity, 616  
in bacteria, 440–442, 441f  
in eukaryotes, 449–454, 450f  
regulation of in prokaryotes, 620–636  
mechanism, 429–464  
RNA polymerase structure, 430–432, 431t, 432f  
speed of, 521  
steps  
elongation, 432, 433f, 434, 442–445, 444f, 455–457, 455f, 458f  
initiation, 432, 433f, 434, 440–442, 441f, 449–454, 450f
- termination, 432, 433f, 434, 445–446, 446f, 447f, 460–462, 461f  
termination, riboswitches and, 703–704, 704f
- Transcriptional activator-like effector nucleases (TALENs), 173
- Transcriptional regulation  
bacteriophage  $\lambda$ , 636–652  
circuits, 776, 776f  
in eukaryotes, 657–698  
combinatorial control, 678–681, 679f, 680f  
conserved mechanisms, 659–665  
activating regions, 660–661, 661f, 663–665  
DNA-binding domains, 660–663, 661f–663f  
DNA looping, 672  
epigenetic regulation, 694–697, 695f  
gene silencing, 687–693, 688f–690f, 693f  
human disease and, 696b  
insulators, 672–673, 673f  
locus control region (LCR), 673–674, 674f  
overview, 657–659  
prokaryotes compared to, 657–659, 658f  
recruitment of protein complexes  
by activators, 665–672  
repressors, 681–682, 682f  
signal integration, 675–677  
signal transduction, 682–686, 685f  
in prokaryotes, 615–653  
general principles, 615–620  
allostery, 618, 618f, 619–620  
cooperative binding, 617, 619–620, 619f  
DNA looping, 618–619, 619f  
promoter activation, 616–618, 617f, 618f  
recruitment of RNA polymerase, 617, 617f  
transcription initiation, regulation of, 620–636  
alternative  $\sigma$  factors, 630, 631f  
*araBAD* operon of *E. coli*, 634, 634f  
bacteriophage  $\phi$ 29 P<sub>4</sub> protein, 633  
Gal repressor  
*lac* system of *E. coli*, 620–627, 628b–629b  
MerR, 630–631, 632–633, 632f  
NtrC, 630–632, 632f
- Transcriptional silencing, 673, 687
- Transcription bubble, 432
- Transcription-coupled repair, 329, 330f, 445

- Transcription factors  
in *Arabidopsis*, 815  
MADS, 815
- Transcriptome, 170
- Transduction, 802  
generalized, 802, 804, 805f  
specialized, 804
- Transesterification, 400f, 401, 470, 479b, 480
- Transferases, 69
- Transfer DNA (T-DNA), 813
- Transfer RNA (tRNA), 513–519  
as adaptors between codons and amino acids, 513–514  
amino acid attachment, 515–517, 516f  
binding to defined trinucleotide codons, 579–580, 580t  
CCA-adding enzymes, 513b  
charged, 515–519, 516f, 524, 528  
description, 510  
discovery, 34, 509  
discriminator base, 518  
frameshifting, 542  
function of, 108  
initiator, 528–529, 530, 538  
intergenic suppression by mutant tRNA, 584–585  
isoaccepting, 517  
mitochondrial, 587–588  
modified bases in, 513–514, 514f  
percentage of total cellular RNA, 34  
ribosomal discrimination, 519, 537–538, 539f  
ribosome binding sites for, 525–527, 525f–528f  
ribosome recycling factor as mimic of, 548  
structure, 35f, 513–515, 514f, 515f, 517f  
uncharged, 515, 705
- Transformation  
*Arabidopsis*, 813  
bacteria, 804–805  
DNA-mediated, 804–805  
Frederick Griffith's experiments, 22–23, 22f  
germline, 827  
homeotic, 765b  
in mice, 827  
P-element, 824, 824f, 825f  
recombinational in yeast, 810, 810f  
of vector DNA into host organism, 155, 155f  
*transib*, 420
- Transient excursion model, 441, 441f
- Transitions, 314, 314f, 575
- Transition state, 141–142
- Translation, 509–569  
antibiotics targeting, 552b–553b  
in central dogma, 34  
coupling of transcription and translation, 520–521, 520f  
description, 36f, 509–510  
directionality, 524  
elongation, 535–544  
energetic cost of, 509  
energy use, 543–544  
error rate, 537  
initiation, 528–535, 528f  
in eukaryotes, 530–535, 530f–534f  
at internal ribosome entry sites (IRESs), 533b–534b  
in prokaryotes, 528–530, 530f  
riboswitch control, 703, 704, 704f
- mRNA, 510–513  
eukaryotic, 511, 511f, 512–513  
open reading frames, 510–512  
prokaryotic, 511, 511f  
reading frame, 510–512, 511f  
regulation, 549–561  
of bacterial initiation by inhibiting 30S subunit binding, 549, 551, 551f  
in eukaryotes, 556–561  
of ferritin, 558, 559f  
global regulators, 556, 557f  
by masking ribosome-binding site (RBS), 549, 551  
in prokaryotes, 549–555, 551f, 554f, 555f  
of ribosomal proteins in bacteria, 551, 553–555, 555f  
spatial control by mRNA-specific 4E-BPs, 556–557, 558f  
of yeast Gcn4, 558–559, 560f, 561
- ribosome, 519–528  
speed of, 520  
start codon, 510  
stop codon, 510  
termination, 544–549
- tRNAs, 513–519  
amino acid attachment, 515–517, 516f  
structure, 513–515, 514f, 515f, 517f
- Translational coupling, 512
- Translation-dependent regulation of mRNA and protein stability, 563–567
- Translation initiation factors, 529–535, 530f, 531f, 532f, 534f, 536f
- Translesion DNA synthesis, 325t, 333–338, 334f, 335f, 337f
- Translesion polymerase, 325
- Translocation of DNA, 237–238
- Translocation, ribosome, 522, 537, 541–543, 542f
- Transposable elements  
in *Arabidopsis* transformation studies, 813  
classes/types, 395, 395f, 409t  
description, 377, 393  
discovery of, 393, 408b  
DNA transposons
- autonomous and nonautonomous transposons, 396
- cut and paste, mechanism, 397–401, 398f, 400f, 409, 412
- genetic organization, 395–396, 395f
- replicative transposition  
mechanism, 401–403, 402f
- target-site duplications, 395f, 396
- examples, 406–416  
*Hermes*, 400f, 401, 408b  
LINEs, 414–416, 415f  
phage Mu, 405f, 411, 412f, 413b  
table of, 409t  
*Tc1/mariner*, 411–414  
*Tn7*, 399–400, 400f  
*Tn10*, 400, 400f, 409–410, 409f, 410f  
Ty elements, 414, 414f, 415f
- genetic organization, 395f
- genome occurrence and distribution, 393–394, 394f
- as mutagens, 393
- P-elements in *Drosophila*, 824, 824f, 825f
- poly-A retrotransposons  
genetic organization, 395f, 396–397  
reverse splicing mechanism, 405–406, 407f
- regulation, 406–416  
copy number control, 408, 409–410  
target-site choice, 408–409
- as repetitive DNA sequences, 207–208
- ubiquity of, 393, 406
- virus-like retrotransposons  
genetic organization, 395f, 396  
mechanism of transposition, 403, 404f
- Transposases, 399–401  
catalytic domain, 404–405, 405f  
description, 395  
structure, 404–405, 405f
- Transposition, 393–416  
description, 207–208  
mechanisms  
cut-and-paste, 397–401, 398f, 399f, 400f  
overview, 393f  
replicative, 401–403, 402f  
reverse-splicing, 405–406, 407f  
of virus-like retrotransposons and retroviruses, 403, 404f  
overview, 393–394, 393f  
target immunity, 409, 411, 413b
- Transpositional recombination, 377.  
See also Transposition
- Transposition target immunity, 409, 411, 413b

- Transposons, 313. *See also* Transposable elements  
 composite, 409–410  
 description, 393  
 discovery of, 408b  
 gene order change, 341  
 insertional mutagenesis, 805–806, 806f  
 silencing by RNAi, 724
- Transpososome, 397, 401, 405
- trans-splicing*, 482–483, 482f
- Transversion mutation, 575
- Transversions, 314, 314f
- TRCF, 445
- Triose phosphate isomerase, 131f
- Triplet repeat expansion, 316b
- Tri-snRNP particle, 476
- Trithorax complex (TRC), 765b
- Triticum aestivum*  
 gene density, 203t  
 genome size, 203t
- tRNA. *See* Transfer RNA
- tRNA-binding sites, ribosome, 525–527, 525f–528f
- tRNA<sup>Gln</sup>, 517, 518f
- tRNA<sup>Phe</sup>, 577f
- tRNA<sup>Ser</sup>, 587
- tRNA synthetases. *See* Aminoacyl-tRNA synthetases
- tRNA<sup>Trp</sup>, 585, 587
- tRNA<sup>Tyr</sup>, 584–585
- Trombone model, 285f, 286
- Troponin T, 484, 484f
- trp* operon, 707b–708b
- Trypanosomes, RNA editing in, 502f
- Tsix*, 729–730
- Tuberculosis, 565b
- Tubulin, 741b–742b
- TUDOR domains, 244
- Tup1 protein, 682, 682f
- TUTase (terminal uridylyl transferase), 503
- twist* gene (*Drosophila*), 750, 750f, 786
- Twist number (*Tw*), 94, 94f, 103
- Twist protein, 751
- Twist protein, 93–95
- Two-hybrid assay, 664b
- Ty elements, 414, 414f, 415f
- Tyrosine, 518, 518f
- Tyrosine recombinases  
 description, 380, 381f, 381t, 383–385, 385f  
 λ integrase, 387–389, 388f  
 XerCD, 392b
- Tyrosyl-tRNA synthetase, 518
- U**
- U1, 473, 474, 474f, 475f, 476, 482, 483, 486, 491, 491f
- U1A protein, 141, 141f
- U1 small nuclear RNA (snRNA), 141, 141f
- U2, 473, 474f, 475f, 476, 478f, 486
- U2AF (U2 auxilliary factor), 474, 482, 492
- U4, 475f, 476, 478f
- U5, 475f, 476, 478f, 483
- U6, 473, 474f, 475f, 476, 478f
- U11, 483, 483f
- U12, 483, 483f
- UAA codon, 500, 544, 577, 586, 588
- UAG codon, 544, 577, 584–585, 586, 588
- UBF, 462f, 463
- Ubiquitin, 336
- Ubiquitination, of sliding clamp, 336
- Ubx* (*Ultrabithorax*) (*Drosophila*), 762–763, 763f, 766–769, 766f–768f
- UCE (upstream control element), 462–463, 462f
- UGA codon, 544, 577, 585, 587, 588
- U:G mismatch, 325
- Ultraviolet light, DNA damage from, 322, 322f, 326f
- UmuC, 334, 335
- UmuD, 334, 335
- Unfolding, protein, 134, 135b
- UP-element, 435f, 436, 437, 438f
- Upstream activator sequences (UASs), 449
- Upstream ORFs (uORFs), 533b, 559, 560f, 561
- Uracil  
 from cytosine deamination, 320–321, 320f  
 description, 33  
 insertion, guide RNA-mediated, 501–503, 502f  
 pairing with adenine, 107, 108f  
 structure, 107, 108f
- Uracilglycosylase, 326, 327f
- Urey, Harold, 596–597
- Uridine, 514, 514f
- UvrA, 328–329, 329f
- UvrABC, 445
- UvrB, 328–329, 329f
- UvrC, 328–329, 329f
- UvrD, 316, 328–329, 329f
- V**
- Valence, 52
- Valine, 518, 518f, 528
- Valyl-tRNA synthetase, 518
- van der Waals forces  
 description, 52, 56–57, 56f, 56t, 57f  
 hydrophobic, 60–62, 61f
- van der Waals radius, 56–57, 56t, 57f
- Vand, Vladimir, 24
- Variegation, 689–690, 689f
- Vector, plasmid, 154–155, 155f
- Vernalization, 815
- Vesicles, lipid, 605, 605f, 606f
- Vif (viral infectivity factor), 503b
- VIGS (virally induced gene silencing), 814
- Viroids, hammerhead ribozyme in, 116–117
- Virulence, blocking, 635b–636b
- Virulence genes, bacterial, 110, 111f
- Virus  
 gene density in, 204  
 nucleic acids as genetic material, 23, 24f  
 spacer sequences from, 710, 710f
- VISTA (program), 171
- V(D)J recombination, 332b, 416–420, 417f–419f
- vnd* locus in *Drosophila*, 169f
- von Tschermark-Seysenegg, Erich, 6
- W**
- Wallace, Alfred R., 5
- Water  
 hydrogen bonds in, 59, 59f, 125, 126f  
 ice, 126f  
 protein structure, effect on, 125–126  
 structure of, 55, 55f
- Watson–Crick base pairing, 81–82, 81f
- Watson, James D., 24, 168, 596
- Weiss, Samuel B., 36
- Western blotting, 177
- White, John, 816
- white* mutation, in *Drosophila*, 689–690, 689f, 821–822, 824
- Whole-genome tiling arrays, 169–171, 170f
- Wieschaus, Eric, 749b
- Wild-type gene, defined, 10
- Wilkins, Maurice, 24
- Wobble concept, 575–577, 575t, 576f
- Wright, Sewall, 15
- Writhing, 93–95, 96–97, 230b
- Writhing number (*Wr*), 94, 94f, 103
- X**
- Xanthine, 321
- X chromosome, genes on, 10
- XerCD recombinase, 391f, 392b
- Xeroderma pigmentosum, 329, 330b
- X-gal, 626f
- Xic* (X-inactivation center), 729
- X-inactivation, 728–730, 729f, 730f
- Xisf, 467
- Xis protein, 389
- Xist*, 728, 729, 729f, 730, 730f
- XP genes/proteins, 329, 330b
- X-ray diffraction pattern of DNA, 24, 24f, 86, 88
- X-rays  
 DNA damage from, 322  
 mutation induction by, 16
- XRCC4, 332, 333f
- Xrn2, 461
- Y**
- Yanofsky, Charles, 37, 807
- Y-arc, 291b–292b
- Yeast  
 combinatorial control of the mating-type genes, 680–681, 680f

- Yeast (*Continued*)  
 evolution of regulatory circuit, 683b  
*GAL* genes, 681, 682f  
*Gcn4*, 129f, 137, 558–559, 560f, 561  
 genetic and physical maps compared,  
   373f  
*HO* gene, 675–676, 677f  
 mediator, 454, 454f  
 promoters, 671  
 regulatory elements of, 658f  
 RNA polymerases, 431, 432f, 455, 457f  
 SAGA complex, 452  
 silencing in, 688–689, 688f
- two-hybrid assay, 664b  
 Ty elements, 414, 414f, 415f  
 Yeast artificial chromosomes (YACs), 154  
 Yeast two-hybrid assay, 182  
*yellow (y)* locus (*Drosophila*),  
   759b–760b  
 Y family of DNA polymerase, 334, 335,  
   336b
- Z**  
 Zamecnik, Paul C., 34, 509  
 Z DNA, 87f, 89, 89f, 90t
- Zea mays*  
 crossing over, 11, 11f  
 gene density, 203t  
 genome size, 203t  
 transposons, 394f, 408b  
 Zebrafish, Her1 and Her7 proteins of,  
   788–789, 789f  
 Zigzag model, nucleosome, 232, 233,  
   233f, 234f  
 Zinc cluster domain, 662  
 Zinc finger, 139–140, 139f, 662,  
   662f, 741  
 Zinc-finger nucleases, 173

---

### The Genetic Code

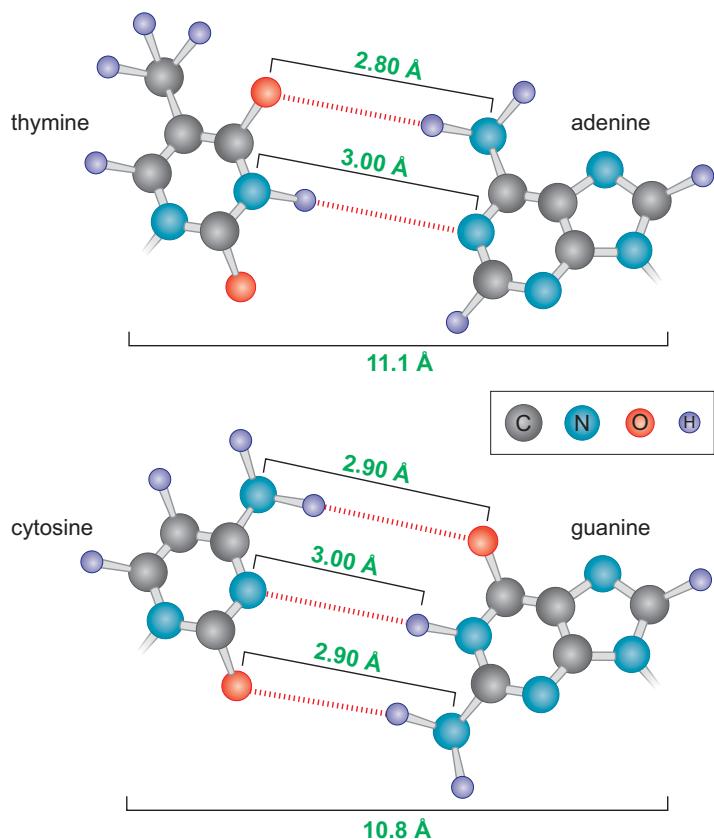
---

		second position								
		U	C	A	G					
first position (5' end)	U	UUU UUC UUA UUG	Phe  Leu	UCU UCC UCA UCG	Ser	UAU UAC UAA* UAG*	Tyr  stop	UGU UGC UGA* UGG	Cys  stop  Trp	U C A G
	C	CUU CUC CUA CUG	Leu	CCU CCC CCA CCG	Pro	CAU CAC CAA CAG	His  Gln	CGU CGC CGA CGG	Arg	U C A G
	A	AUU AUC AUA AUG†	Ile  Met	ACU ACC ACA ACG	Thr	AAU AAC AAA AAG	Asn  Lys	AGU AGC AGA AGG	Ser  Arg	U C A G
	G	GUU GUC GUA GUG	Val	GCU GCC GCA GCG	Ala	GAU GAC GAA GAG	Asp  Glu	GGU GGC GGA GGG	Gly	U C A G
		third position (3' end)								

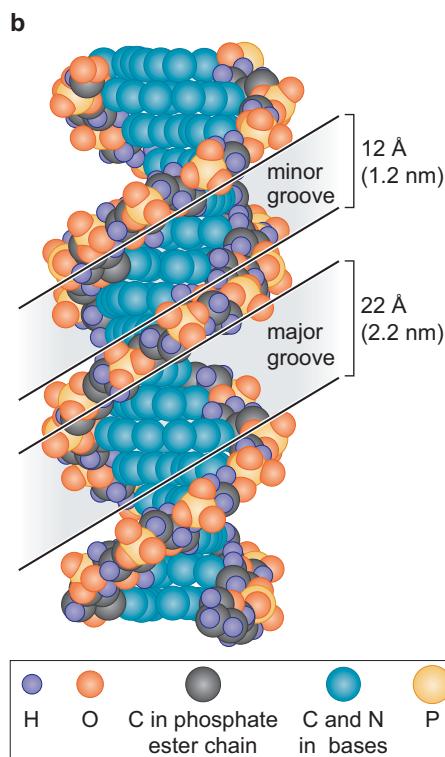
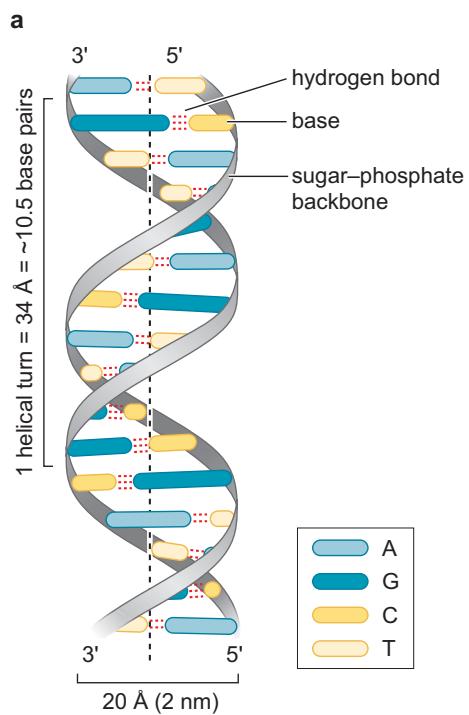
\* Chain-terminating or "nonsense" codons

† Also used in bacteria to specify the initiator formyl-Met-tRNA<sup>fMet</sup>

---



The position and length of the hydrogen bonds between the base pairs.



- 
- (a) Schematic model of the double helix.  
 (b) Space-filling model of the double helix.

---

**Abbreviations for Amino Acids**

Amino Acid	Three-Letter Abbreviation	One-Letter Symbol
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Asparagine or aspartic acid	Asx	B
Cysteine	Cys	C
Glutamine	Gln	Q
Glutamic acid	Glu	E
Glutamine or glutamic acid	Glx	Z
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

---