

MOLECULAR BIOLOGY

Principles and Practice

Michael M.

COX

Jennifer A.

DOUDNA

Michael

O'DONNELL

This page intentionally left blank

Molecular Biology

This page intentionally left blank

Molecular Biology

Principles and Practice

Michael M. Cox

University of Wisconsin-Madison

Jennifer A. Doudna

University of California, Berkeley

Michael O'Donnell

The Rockefeller University



W. H. Freeman and Company • New York

Publisher: Kate Ahr Parker

Developmental Editors: Erica Pantages Frost, Erica Ann Champion, Betsy Dilernia

Associate Director of Marketing: Debbie Clare

Media and Supplements Editors: J. D. Bullard, Patrick Shriner, Marni Rolfes

Photo Editor: Cecilia Varas

Photo Researcher: Elyse Rieder

Art Director: Diana Blume

Project Editor: Jane O'Neill

Manuscript Editors: Linda Strange, Brook Soltvedt

Illustrations: H. Adam Steinberg, Dragonfly Media Group

Senior Illustration Coordinator: Bill Page

Production Coordinator: Susan Wein

Composition: MPS Limited, a Macmillan Company

Printing and Binding: Quad Graphics

Front cover image: Livet, J., T.A. Weissman, H.N. Kang, R.W. Draft, J. Lu, R.A. Bennis, J.R. Sanes, and J.W. Lichtman. 2007. Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature* 450:56–62, Fig. 4. Courtesy of Jeff Lichtman. *Back cover illustration:* H. Adam Steinberg.

Throughout the text, a number of illustrations have been adapted from *Lehninger Principles of Biochemistry*, Fifth Edition, by David L. Nelson and Michael M. Cox.

Library of Congress Control Number: 2010943173

ISBN-13: 978-0-7167-7998-8

ISBN-10: 0-7167-7998-6

© 2012 by W. H. Freeman and Company

All rights reserved

Printed in the United States of America

First Printing

W. H. Freeman and Company

41 Madison Avenue

New York, NY 10010

Hounds mills, Basingstoke RG21 6XS, England

www.whfreeman.com

To our students, for the inspiration they provide every day
and to our mentors, in gratitude for their guidance:

Tom Cech

Fred Grieman

Bill Jencks

Arthur Kornberg

Bob Lehman

Sharon Panasenko

David Sheppard

Jack Szostak

Hal White

Charles Williams

About the Authors



Michael M. Cox was born in Wilmington, Delaware. After graduating from the University of Delaware, he went to Brandeis University to do his doctoral work with William P. Jencks, and then to Stanford for postdoctoral study with I. Robert Lehman. He is currently Professor of Biochemistry at the University of Wisconsin–Madison. His research focuses on recombinational DNA repair processes. Cox has received awards for both teaching and research, including the 1989 Eli Lilly Award in Biological Chemistry from the American Chemical Society and two major teaching awards from the University of Wisconsin. He has coauthored four editions of *Lehninger Principles of Biochemistry*.



Jennifer A. Doudna grew up on the Big Island of Hawaii and became interested in chemistry and biochemistry in high school. She received her B.A. in biochemistry from Pomona College and her Ph.D. from Harvard University, working in the laboratory of Jack Szostak, with whom she also did postdoctoral research. She then went to the University of Colorado as a Lucille P. Markey scholar and postdoctoral fellow with Thomas Cech. Doudna is currently Professor of Molecular and Cell Biology and Professor of Chemistry at the University of California, Berkeley, and an Investigator of the Howard Hughes Medical Institute. She is a member of the National Academy of Sciences, the American Academy of Arts and Sciences, and the Institute of Medicine. She is also a Fellow of the American Association for the Advancement of Science.



Michael O'Donnell grew up in a neighborhood on the banks of the Columbia River outside Vancouver, Washington. He had several inspirational teachers at Hudson Bay High School who led him into science. He received his B.A. in biochemistry from the University of Portland and his Ph.D. from the University of Michigan, where he worked under Charles Williams, Jr., on electron transfer in the flavoprotein thioredoxin reductase. He performed postdoctoral work on *E. coli* replication with Arthur Kornberg and then on herpes simplex virus replication with I. Robert Lehman in the Biochemistry Department at Stanford University. O'Donnell is currently Professor of Biochemistry and Structural Biology at The Rockefeller University and an Investigator of the Howard Hughes Medical Institute. He is a member of the National Academy of Sciences.

Contents in Brief

I Foundations

- 1** Studying the Molecules of Life 1
- 2** DNA: The Repository of Biological Information 23
- 3** Chemical Basis of Information Molecules 61
- 4** Protein Structure 95
- 5** Protein Function 135

II Nucleic Acid Structure and Methods

- 6** DNA and RNA Structure 175
- 7** Studying Genes 215
- 8** Genomes, Transcriptomes, and Proteomes 259
- 9** Topology: Functional Deformations of DNA 297
- 10** Nucleosomes, Chromatin, and Chromosome Structure 331

III Information Transfer

- 11** DNA Replication 363
- 12** DNA Mutation and Repair 409
- 13** Recombinational DNA Repair and Homologous Recombination 445
- 14** Site-Specific Recombination and Transposition 481
- 15** DNA-Dependent Synthesis of RNA 515
- 16** RNA Processing 547
- 17** The Genetic Code 585
- 18** Protein Synthesis 615

IV Regulation

- 19** Regulating the Flow of Information 667
- 20** The Regulation of Gene Expression in Bacteria 697
- 21** The Transcriptional Regulation of Gene Expression in Eukaryotes 733
- 22** The Posttranscriptional Regulation of Gene Expression in Eukaryotes 767

Appendix: Model Organisms A-1
Glossary G-1
Solutions to Problems S-1
Index I-1

Contents

I Foundations

1 Studying the Molecules of Life 1

Moment of Discovery Jack Szostak, on his discovery of self-dividing vesicles that mimic growing cells 1

1.1 The Evolution of Life on Earth 2

What Is Life? 2

Evolution Underpins Molecular Biology 4

Life on Earth Probably Began with RNA 5

HIGHLIGHT 1-1 Evolution Observing Evolution in the Laboratory 7

The Last Universal Common Ancestor Is the Root of the Tree of Life 8

Evolution by Natural Selection Requires Variation and Competition 9

1.2 How Scientists Do Science 12

Science Is a Path to Understanding the Natural Universe 12

The Scientific Method Underlies Scientific Progress 13

The Scientific Method Is a Versatile Instrument of Discovery 14

Scientists Work within a Community of Scholars 16

How We Know 19

Adenine Could Be Synthesized with Prebiotic Chemistry 19

Clay Had a Role in Prebiotic Evolution 20

Darwin's World Helped Him Connect the Dots 21

2 DNA: The Repository of Biological Information 23

Moment of Discovery James Berger, on his discovery of the structure and mechanism of topoisomerase II 23

2.1 Mendelian Genetics 25

Mendel's First Law: Allele Pairs Segregate during Gamete Formation 25

Mendel's Second Law: Different Genes Assort Independently during Gamete Formation 28

There Are Exceptions to Mendel's Laws 29

2.2 Cytogenetics: Chromosome Movements during Mitosis and Meiosis 31

Cells Contain Chromosomes and Other Internal Structures 32

Mitosis: Cells Evenly Divide Chromosomes between New Cells 32

Meiosis: Chromosome Number Is Halved during Gamete Formation 34

2.3 The Chromosome Theory of Inheritance 38

Sex-Linked Genes in the Fruit Fly Reveal That Genes Are on Chromosomes 38

Linked Genes Do Not Segregate Independently 40

Recombination Unlinks Alleles 40

Recombination Frequency Can Be Used to Map Genes along Chromosomes 41

2.4 Molecular Genetics 44

DNA Is the Chemical of Heredity 44

Genes Encode Polypeptides and Functional RNAs 46

The Central Dogma: Information Flows from DNA to RNA to Protein 47

Mutations in DNA Give Rise to Phenotypic Change 50

HIGHLIGHT 2-1 Medicine The Molecular Biology of Sickle-Cell Anemia, a Recessive Genetic Disease of Hemoglobin 52

How We Know 56

Chromosome Pairs Segregate during Gamete Formation in a Way That Mirrors the Mendelian Behavior of Genes 56

Corn Crosses Uncover the Molecular Mechanism of Crossing Over 57

Hershey and Chase Settle the Matter: DNA Is the Genetic Material 58

3 Chemical Basis of Information Molecules 61

Moment of Discovery Judith Klinman, on her discovery of hydrogen tunneling in enzyme-catalyzed reactions 61

3.1 Chemical Building Blocks of Nucleic Acids and Proteins 62

- Nucleic Acids Are Long Chains of Nucleotides 62
- Proteins Are Long Polymers of Amino Acids 64
- Chemical Composition Helps Determine Nucleic Acid and Protein Structure 65
- Chemical Composition Can Be Altered by Postsynthetic Changes 65

3.2 Chemical Bonds 68

- Electrons Are Shared in Covalent Bonds and Transferred in Ionic Bonds 68
- Chemical Bonds Are Explainable in Quantum Mechanical Terms 70
- Both the Making and Breaking of Chemical Bonds Involve Energy Transfer 72
- Electron Distribution between Bonded Atoms Determines Molecular Behavior 72

3.3 Weak Chemical Interactions 73

- Van der Waals Forces Are Nonspecific Contacts between Atoms 74
- Hydrophobic Interactions Bring Together Nonpolar Molecules 75
- Hydrogen Bonds Are a Special Kind of Noncovalent Bond 76
- Combined Effects of Weak Chemical Interactions Stabilize Macromolecular Structures 76
- Weak Chemical Bonds Also Facilitate Macromolecular Interactions 77

3.4 Stereochemistry 78

- Three-Dimensional Atomic Arrangements Define Molecules 78
- Biological Molecules and Processes Selectively Use One Stereoisomer 79
- Proteins and Nucleic Acids Are Chiral 79

HIGHLIGHT 3-1 Medicine The Behavior of a Peptide Made of D-Amino Acids 81

3.5 The Role of pH and Ionization 81

- The Hydrogen Ion Concentration of a Solution Is Measured by pH 81
- Buffers Prevent Dramatic Changes in pH 82

The Henderson-Hasselbalch Equation Estimates the pH of a Buffered Solution 83

3.6 Chemical Reactions in Biology 84

- The Mechanism and Speed of Chemical Transformation Define Chemical Reactions 84
- Biological Systems Follow the Laws of Thermodynamics 86
- Catalysts Increase the Rates of Biological Reactions 87
- Energy Is Stored and Released by Making and Breaking Phosphodiester Bonds 88

HIGHLIGHT 3-2 Evolution ATP: The Critical Molecule of Energy Exchange in All Cells 89

How We Know 91

- Single Hydrogen Atoms Are Speed Bumps in Enzyme-Catalyzed Reactions 91
- Peptide Bonds Are (Mostly) Flat 92

4 Protein Structure 95

Moment of Discovery Steve Mayo, on his discovery of the first successful method for computational protein design 95

4.1 Primary Structure 97

- Amino Acids Are Categorized by Chemical Properties 97
- Amino Acids Are Connected in a Polypeptide Chain 99

HIGHLIGHT 4-1 A Closer Look Purification of Proteins by Column Chromatography and SDS-PAGE 100

- Evolutionary Relationships Can Be Determined from Primary Sequence Comparisons 102

4.2 Secondary Structure 103

- The α Helix Is a Common Form of Secondary Protein Structure 104
- The β Sheet Is Composed of Long, Extended Strands of Amino Acids 105
- Reverse Turns Allow Secondary Structures to Fold 106

4.3 Tertiary and Quaternary Structures 107

- Tertiary and Quaternary Structures Can Be Represented in Different Ways 107
- Domains Are Independent Folding Units within the Protein 107
- Supersecondary Structure Elements Are Building Blocks of Domains 109

Quaternary Structures Range from Simple to Complex 112	The Rates of Enzyme-Catalyzed Reactions Can Be Quantified 149
HIGHLIGHT 4-2 A Closer Look Protein Structure Databases 114	DNA Ligase Activity Illustrates Some Principles of Catalysis 151
Protein Structures Help Explain Protein Evolution 114	
4.4 Protein Folding 115	HIGHLIGHT 5-1 A Closer Look Reversible and Irreversible Inhibition 152
Predicting Protein Folding Is a Goal of Computational Biology 115	
Polypeptides Fold through a Molten Globule Intermediate 116	
HIGHLIGHT 4-3 Medicine Prion-Based Misfolding Diseases 118	5.3 Motor Proteins 156
Chaperones and Chaperonins Can Facilitate Protein Folding 119	Helicases Abound in DNA and RNA Metabolism 157
Protein Isomerases Assist in the Folding of Some Proteins 121	Helicase Mechanisms Have Characteristic Molecular Parameters 158
4.5 Determining the Atomic Structure of Proteins 121	5.4 The Regulation of Protein Function 161
Most Protein Structures Are Solved by X-Ray Crystallography 121	Modulator Binding Causes Conformational Changes in Allosteric Enzymes 161
Smaller Protein Structures Can Be Determined by NMR 125	Allosteric Enzymes Have Distinctive Binding and/or Kinetic Properties 161
How We Know 129	Enzyme Activity Can Be Affected by Autoinhibition 163
Sequence Comparisons Yield an Evolutionary Roadmap from Bird Influenza to a Deadly Human Pandemic 129	Some Proteins Are Regulated by Reversible Covalent Modification 163
We Can Tell That a Protein Binds ATP by Looking at Its Sequence 130	Phosphoryl Groups Affect the Structure and Catalytic Activity of Proteins 165
Disulfide Bonds Act as Molecular Cross-Braces to Stabilize a Protein 131	HIGHLIGHT 5-2 Medicine HIV Protease: Rational Drug Design Using Protein Structure 166
5 Protein Function 135	Some Proteins Are Regulated by Proteolytic Cleavage 167
Moment of Discovery Tim Lohman, on his discovery of multiple DNA-binding modes for SSB 135	How We Know 169
5.1 Protein-Ligand Interactions 136	The Lactose Repressor Is One of the Great Sagas of Molecular Biology 169
Many Proteins Bind to Other Molecules Reversibly 136	The <i>lacI</i> Gene Encodes a Repressor 170
Protein-Ligand Interactions Can Be Quantified 137	The Lactose Repressor Is Found 171
DNA-Binding Proteins Guide Genome Structure and Function 138	II Nucleic Acid Structure and Methods
5.2 Enzymes: The Reaction Catalysts of Biological Systems 144	6 DNA and RNA Structure 175
Enzymes Catalyze Specific Biological Reactions 145	Moment of Discovery Jamie Cate, on determining the molecular structure of the bacterial ribosome 175
Enzymes Increase the Rate of a Reaction by Lowering the Activation Energy 147	6.1 The Structure and Properties of Nucleotides 177
	Nucleotides Comprise Characteristic Bases, Sugars, and Phosphates 177
	Phosphodiester Bonds Link the Nucleotide Units in Nucleic Acids 178
	Nucleotide Bases Affect the Three-Dimensional Structure of Nucleic Acids 180
	Nucleotides Play Additional Roles in Cells 181

6.2 DNA Structure 185

- DNA Molecules Have Distinctive Base Compositions 185
- DNA Is Usually a Right-Handed Double Helix 185
- DNA Adopts Different Helical Forms 187
- Certain DNA Sequences Adopt Unusual Structures 190

HIGHLIGHT 6-1 Technology DNA Computing 191**HIGHLIGHT 6-2 Technology The Design of a DNA Octahedron 194****6.3 RNA Structure 194**

- RNAs Have Helical Secondary Structures 196
- RNAs Form Various Stable Three-Dimensional Structures 197

HIGHLIGHT 6-3 Medicine RNA Structure Governing HIV Gene Expression 199**6.4 Chemical and Thermodynamic Properties of Nucleic Acids 200**

- Double-Helical DNA and RNA Can Be Denatured 200
- Nucleic Acids from Different Species Can Form Hybrids 202
- Nucleotides and Nucleic Acids Undergo Uncatalyzed Chemical Transformations 203
- Base Methylation in DNA Plays an Important Role in Regulating Gene Expression 205
- RNA Molecules Are Often Site-Specifically Modified In Vivo 206
- The Chemical Synthesis of DNA and RNA Has Been Automated 206

How We Know 209

- DNA Is a Double Helix 209
- DNA Helices Have Unique Geometries That Depend on Their Sequence 210
- Ribosomal RNA Sequence Comparisons Provided the First Hints of the Structural Richness of RNA 211

7 Studying Genes 215**Moment of Discovery** Norman Arnheim, on the 1980 discovery of interspersed CA repeats in genomic DNA 215**7.1 Isolating Genes for Study (Cloning) 216**

- Genes Are Cloned by Splicing Them into Cloning Vectors 217

Cloning Vectors Allow Amplification of Inserted DNA Segments 219

DNA Libraries Provide Specialized Catalogs of Genetic Information 224

7.2 Working with Genes and Their Products 226

- Gene Sequences Can Be Amplified with the Polymerase Chain Reaction 226
- The Sanger Method Identifies Nucleotide Sequences in Cloned Genes 228

HIGHLIGHT 7-1 Technology A Potent Weapon in Forensic Medicine 230

Cloned Genes Can Be Expressed to Amplify Protein Production 233

HIGHLIGHT 7-2 Technology DNA Sequencing: Ever Faster and Cheaper 234

- Many Different Systems Are Used to Express Recombinant Proteins 236
- Alteration of Cloned Genes Produces Altered Proteins 239
- Terminal Tags Provide Handles for Affinity Purification 240

7.3 Understanding the Functions of Genes and Their Products 242

- Protein Fusions and Immunofluorescence Can Localize Proteins in Cells 242
- Proteins Can Be Detected in Cellular Extracts with the Aid of Western Blots 243
- Protein-Protein Interactions Can Help Elucidate Protein Function 244
- DNA Microarrays Reveal Cellular Protein Expression Patterns and Other Information 248

How We Know 251

- New Enzymes Take Molecular Biologists from Cloning to Genetically Modified Organisms 251
- A Dreamy Night Ride on a California Byway Gives Rise to the Polymerase Chain Reaction 252
- Coelenterates Show Biologists the Light 253

8 Genomes, Transcriptomes, and Proteomes 259**Moment of Discovery** Joe DeRisi, on his discovery of the SARS virus 259**8.1 Genomes and Genomics 260**

- Many Genomes Have Been Sequenced in Their Entirety 260
- Annotation Provides a Description of the Genome 263

HIGHLIGHT 8-1 Evolution Getting to Know the Neanderthals 264

Genome Databases Provide Information about Every Type of Organism 266

HIGHLIGHT 8-2 Technology Sampling Biodiversity with Metagenomics 268

The Human Genome Contains Many Types of Sequences 269

Genome Sequencing Informs Us about Our Humanity 271

Genome Comparisons Help Locate Genes Involved in Disease 274

8.2 Transcriptomes and Proteomes 277

Special Cellular Functions Are Revealed in a Cell's Transcriptome 277

High-Throughput DNA Sequencing Is Used in Transcriptome Analysis 278

The Proteins Generated by a Cell Constitute Its Proteome 278

Electrophoresis and Mass Spectrometry Support Proteomics Research 280

Computational Approaches Help Elucidate Protein Function 281

Experimental Approaches Reveal Protein Interaction Networks 282

8.3 Our Genetic History 282

All Living Things Have a Common Ancestor 283

Genome Comparisons Provide Clues to Our Evolutionary Past 283

HIGHLIGHT 8-3 Evolution Phylogenetics Solves a Crime 284

The Human Journey Began in Africa 287

Human Migrations Are Recorded in Haplotypes 288

How We Know 292

Haemophilus influenzae Ushers in the Era of Genome Sequences 292

9 Topology: Functional Deformations of DNA 297

Moment of Discovery Carlos Bustamante, on discovering the elasticity of DNA 297

9.1 The Problem: Large DNAs in Small Packages 298

Chromosome Function Relies on Specialized Genomic Sequences 298

Chromosomes Are Longer Than the Cellular or Viral Packages Containing Them 300

HIGHLIGHT 9-1 Medicine The Dark Side of Antibiotics 303

9.2 DNA Supercoiling 305

Most Cellular DNA Is Underwound 307

DNA Underwinding Is Defined by the Topological Linking Number 308

DNA Compaction Requires a Special Form of Supercoiling 310

9.3 The Enzymes That Promote DNA Compaction 312

Topoisomerases Catalyze Changes in the Linking Number of DNA 312

The Two Bacterial Type II Topoisomerases Have Distinct Functions 313

Eukaryotic Topoisomerases Have Specialized Functions in DNA Metabolism 313

SMC Proteins Facilitate the Condensation of Chromatin 316

HIGHLIGHT 9-2 Medicine Curing Disease by Inhibiting Topoisomerases 318

How We Know 323

The Discovery of Supercoiled DNA Goes through Twists and Turns 323

The First DNA Topoisomerase Unravels Some Mysteries 324

DNA Gyrase Passes the Strand Test 325

10 Nucleosomes, Chromatin, and Chromosome Structure 331

Moment of Discovery Jonathan Widom, on discovering the code for genome-wide nucleosome organization 331

10.1 Nucleosomes: The Basic Units of DNA Condensation 332

Histone Octamers Organize DNA into Repeating Units 332

DNA Wraps Nearly Twice around a Single Histone Octamer 334

Histone Tails Mediate Internucleosome Connections That Regulate the Accessibility of DNA 336

10.2 Higher-Order Chromosome Structure 338

Histone H1 Binds the Nucleosome to Form the Chromatosome 338

Chromosomes Condense into a Compact Chromatin Filament 339

Higher-Order Chromosome Structure Involves Loops and Coils 341
 Bacterial DNA, Like Eukaryotic DNA, Is Highly Organized 341

10.3 The Regulation of Chromosome Structure 343

Nucleosomes Are Intrinsically Dynamic 344
 ATP-Driven Chromatin Remodeling Complexes Can Reposition Nucleosomes 344
 Variant Histone Subunits Alter DNA-Binding Affinity 346
 Nucleosome Assembly Requires Chaperones 348
 Modifications of Histone Tails Alter DNA Accessibility 348

HIGHLIGHT 10-1 A Closer Look The Use of a Histone Variant in X Chromosome Inactivation 350

Histone Modifications and Remodeling Complexes May Read a Histone Code 351
 Histone Modifying Enzymes Maintain Epigenetic States through Cell Division 352

HIGHLIGHT 10-2 Medicine Defects in Epigenetic Maintenance Proteins Associated with Cancer 355

How We Know 359

Kornberg Wrapped His Mind around the Histone Octamer 359
 A Transcription Factor Can Acetylate Histones 360

III Information Transfer

11 DNA Replication 363

Moment of Discovery Robert Lehman, on discovering DNA ligase 363

11.1 DNA Transactions during Replication 364

DNA Replication Is Semiconservative 364
 Replication Is Initiated at Origins and Proceeds Bidirectionally 367
 Replication Is Semidiscontinuous 368

11.2 The Chemistry of DNA Polymerases 369

DNA Polymerases Elongate DNA in the 5'→3' Direction 369
 Most DNA Polymerases Contain DNA Exonuclease Activity 371

Five *E. coli* DNA Polymerases Function in DNA Replication and Repair 372
 DNA Polymerase Structure Reveals the Basis for Its Accuracy 373
 Processivity Increases the Efficiency of DNA Polymerase Activity 375

11.3 Mechanics of the DNA Replication Fork 377

DNA Polymerase III Is the Replicative Polymerase in *E. coli* 377
 A DNA Sliding Clamp Increases the Speed and Processivity of the Chromosomal Replicase 379
 Many Different Proteins Advance a Replication Fork 380
 Helicase Activity Is Stimulated by Its Connection to the DNA Polymerase 384
 DNA Loops Repeatedly Grow and Collapse on the Lagging Strand 384
 Okazaki Fragments Require Removal of RNA and Ligase-Mediated Joining of DNA 386
 The Replication Fork Is More Complex in Eukaryotes Than in Bacteria 387

11.4 Initiation of DNA Replication 390

Assembly of the Replication Fork Follows an Ordered Sequence of Events 391
 Replication Initiation in *E. coli* Is Controlled at Multiple Steps 393
 Eukaryotic Origins "Fire" Only Once per Cell Cycle 394

11.5 Termination of DNA Replication 395

E. coli Chromosome Replication Terminates Opposite the Origin 395

HIGHLIGHT 11-1 Technology Two-Dimensional Gel Analysis of Replication Origins 396

Telomerase Solves the End Replication Problem in Eukaryotes 398
 Proteins Bind Telomeres to Protect the Ends of Chromosomes 399
 Telomere Length Is Associated with Immortality and Cancer 401

How We Know 403

DNA Polymerase Uses a Template and a Proofreader: Nature's Spell Check 403
 Polymerase Processivity Depends on a Circular Protein That Slides along DNA 404
 Replication Requires an Origin 405

12 DNA Mutation and Repair 409

Moment of Discovery Myron Goodman, on his discovery of DNA polymerase V 409

12.1 Types of DNA Mutations 410

- A Point Mutation Can Alter One Amino Acid 411
- Small Insertion and Deletion Mutations Change Protein Length 412
- Some Mutations Are Very Large and Form Abnormal Chromosomes 414

12.2 DNA Alterations That Lead to Mutations 416

- Spontaneous DNA Damage by Water Can Cause Point Mutations 416
- Oxidative Damage and Alkylating Agents Can Create Point Mutations and Strand Breaks 418
- The Ames Test Identifies DNA-Damaging Chemicals 419
- DNA-Damaging Agents Are Used in Cancer Chemotherapy 421
- Solar Radiation Causes Interbase Cross-Links and Strand Breaks 421
- Errant Replication and Recombination Lead to DNA Damage 424

12.3 Mechanisms of DNA Repair 424

- Mismatch Repair Fixes Misplaced-Nucleotide Replication Errors 425

HIGHLIGHT 12-1 Medicine Mismatch Repair and Colon Cancer 428

- Direct Repair Corrects a Damaged Nucleotide Base in One Step 429
- Base Excision Repairs Subtle Alterations in Nucleotide Bases 430
- Nucleotide Excision Repair Removes Bulky Damaged Bases 433
- Recombination Repairs Lesions That Break DNA 435
- Specialized Translesion DNA Polymerases Extend DNA Past a Lesion 435

HIGHLIGHT 12-2 Medicine Nucleotide Excision Repair and Xeroderma Pigmentosum 436

How We Know 439

- Mismatch Repair in *E. coli* Requires DNA Methylation 439
- UV Lights Up the Pathway to DNA Damage Repair 440
- Translesion DNA Polymerases Produce DNA Mutations 441

13 Recombinational DNA Repair and Homologous Recombination 445

Moment of Discovery Lorraine Symington, on discovering how DNA ends are processed to initiate DNA recombination 445

13.1 Recombination as a DNA Repair Process 447

- Double-Strand Breaks Are Repaired by Recombination 447
- Collapsed Replication Forks Are Reconstructed by Double-Strand Break Repair 449
- A Stalled Replication Fork Requires Fork Regression 450
- Single-Stranded DNA Regions Are Filled In by Gap Repair 452

13.2 Enzymatic Machines in Bacterial Recombinational DNA Repair 453

- RecBCD and RecFOR Initiate Recombinational Repair 454
- RecA Is the Bacterial Recombinase 456
- RecA Protein Is Subject to Regulation 457
- Multiple Enzymes Process DNA Intermediates Created by RecA 458
- Repair of the Replication Fork in Bacteria Can Lead to Dimeric Chromosomes 460

HIGHLIGHT 13-1 Evolution A Tough Organism in a Tough Environment: *Deinococcus radiodurans* 462

13.3 Homologous Recombination in Eukaryotes 464

- Meiotic Recombination Is Initiated at Double-Strand Breaks 465

HIGHLIGHT 13-2 Medicine Why Proper Chromosomal Segregation Matters 466

- Meiotic Recombination Is Completed by a Classic DSBR Pathway 468
- Meiotic Recombination Contributes to Genetic Diversity 469
- Recombination during Mitosis Is Also Initiated at Double-Strand Breaks 469
- Programmed Gene Conversion Events Can Affect Gene Function and Regulation 470
- Some Introns Move via Homologous Recombination 472

13.4 Nonhomologous End Joining 472

- Nonhomologous End Joining Repairs Double-Strand Breaks 473
- Nonhomologous End Joining Is Promoted by a Set of Conserved Enzymes 473

How We Know 476

A Motivated Graduate Student Inspires the Discovery of Recombination Genes in Bacteria 476

A Biochemical Masterpiece Catches a Recombination Protein in the Act 477

14 Site-Specific Recombination and Transposition 481

Moment of Discovery *Wei Yang, on researching the structure and molecular mechanisms of γδ resolvase* 481

14.1 Mechanisms of Site-Specific Recombination 482

Precise DNA Rearrangements Are Promoted by Site-Specific Recombinases 483

Site-Specific Recombination Complements Replication 485

Site-Specific Recombination Can Be a Stage in a Viral Infection Cycle 485

Gene Expression Can be Regulated by Site-Specific Recombination 487

Site-Specific Recombination Reactions Can Be Guided by Auxiliary Proteins 488

14.2 Mechanisms of Transposition 489

HIGHLIGHT 14-1 Technology Using Site-Specific Recombination to Trace Neurons 490

Transposition Takes Place by Three Major Pathways 492

Bacteria Have Three Common Classes of Transposons 496

Retrotransposons Are Especially Common in Eukaryotes 497

HIGHLIGHT 14-2 Evolution Awakening Sleeping Beauty 499

Retrotransposons and Retroviruses Are Closely Related 500

A Retrovirus Causes AIDS 501

HIGHLIGHT 14-3 Medicine Fighting AIDS with HIV Reverse Transcriptase Inhibitors 502

14.3 The Evolutionary Interplay of Transposons and Their Hosts 503

Viruses, Transposons, and Introns Have an Interwoven Evolutionary History 503

A Hybrid Recombination Process Assembles Immunoglobulin Genes 505

How We Know 509

Bacteriophage λ Provided the First Example of Site-Specific Recombination 509

If You Leave Out the Polyvinyl Alcohol, Transposition Gets Stuck 510

15 DNA-Dependent Synthesis of RNA 515

Moment of Discovery *Robert Tjian, on discovering the first specific eukaryotic transcription factor* 515

15.1 RNA Polymerases and Transcription Basics 516

RNA Polymerases Differ in Details but Share Many Features 516

HIGHLIGHT 15-1 A Closer Look The ABCs of RNA: Complexity of the Transcriptome 517

Transcription Initiation, Elongation, and Termination Occur in Discrete Steps 521

DNA-Dependent RNA Polymerases Can Be Specifically Inhibited 521

Transcriptional Regulation Is a Central Mechanism in the Control of Gene Expression 522

15.2 Transcription in Bacteria 523

Promoter Sequences Alter the Strength and Frequency of Transcription 523

Sigma Factors Specify Polymerase Binding to Particular Promoters 525

Structural Changes Lead to Formation of the Transcription-Competent Open Complex 526

Initiation Is Primer-Independent and Produces Short, Abortive Transcripts 527

Transcription Elongation Is Continuous until Termination 529

Specific Sequences in the Template Strand Cause Transcription to Stop 531

15.3 Transcription in Eukaryotes 532

Eukaryotic Polymerases Recognize Characteristic Promoters 532

HIGHLIGHT 15-2 Medicine Using Transcription Factors to Reprogram Cells 533

Pol II Transcription Parallels Bacterial RNA Transcription 535

Transcription Factors Play Specific Roles in the Transcription Process 536

Transcription Initiation In Vivo Requires the Mediator Complex 538

Termination Mechanisms Vary among RNA Polymerases 539

Transcription Is Coupled to DNA Repair, RNA Processing, and mRNA Transport 539

How We Know 541

RNA Polymerase Is Recruited to Promoter Sequences 541
RNA Polymerases Are Both Fast and Slow 542

16 RNA Processing 547

Moment of Discovery *Melissa Jurica, on determining the first electron microscopic structures of spliceosomes* 547

16.1 Messenger RNA Capping and Polyadenylation 549

Eukaryotic mRNAs Are Capped at the 5' End 549
Eukaryotic mRNAs Have a Distinctive 3'-End Structure 550

HIGHLIGHT 16-1 Evolution Eukaryotic mRNA with Unusual 3' Tails 552

mRNA Capping, Polyadenylation, and Splicing Are Coordinately Regulated during Transcription 552

16.2 Pre-mRNA Splicing 554

Eukaryotic mRNAs Are Synthesized as Precursors Containing Introns 554

A Gene Can Give Rise to Multiple Products by Alternative RNA Splicing 555

The Spliceosome Catalyzes Most Pre-mRNA Splicing 557

Some Introns Can Self-Splice without Protein or Spliceosome Assistance 559

Exons from Different RNA Molecules Can Be Fused by *Trans*-Splicing 563

HIGHLIGHT 16-2 Evolution The Origin of Introns 564

16.3 RNA Editing 565

RNA Editing Can Involve the Insertion or Deletion of Bases 566

RNA Editing by Substitution Involves Deamination of A or C Residues 566

16.4 RNA Transport and Degradation 568

Different Kinds of RNA Use Different Nuclear Export Pathways 569

mRNA Transport from the Nucleus to the Cytoplasm Is Coupled to Pre-mRNA Splicing 569

Some mRNAs Are Localized to Specific Regions of the Cytoplasm 570

Cellular mRNAs Are Degraded at Different Rates 571

Processing Bodies Are the Sites of mRNA Storage and Degradation in Eukaryotic Cells 571

16.5 Processing of Non-Protein-Coding RNAs 572

Maturation of tRNAs Involves Site-Specific Cleavage and Chemical Modification 572
Maturation of rRNA Involves Site-Specific Cleavage and Chemical Modification 573
Small Regulatory RNAs Are Derived from Larger Precursor Transcripts 575

16.6 RNA Catalysis and the RNA World Hypothesis 576

Ribozyme Diversity Correlates with Function 576

HIGHLIGHT 16-3 Evolution A Viral Ribozyme Derived from the Human Genome? 577

Could RNA Have Formed the Basis for Early Life on Earth? 578

How We Know 579

Studying Autoimmunity Led to the Discovery of snRNPs 579

RNA Molecules Are Fine-Tuned for Stability or Function 580

Ribozyme Form Explains Function 581

17 The Genetic Code 585

Moment of Discovery *Steve Benner, on discovering that borate minerals stabilize ribose* 585

17.1 Deciphering the Genetic Code: tRNA as Adaptor 586

All tRNAs Have a Similar Structure 587
The Genetic Code Is Degenerate 588
Wobble Enables One tRNA to Recognize Two or More Codons 589
Translation Is Started and Stopped by Specific Codons 590
The Genetic Code Resists Single-Base Substitution Mutations 591
Some Mutations Are Suppressed by Special tRNAs 592

17.2 The Rules of the Code 593

The Genetic Code Is Nonoverlapping 594
There Are No Gaps in the Genetic Code 594
The Genetic Code Is Read in Triplets 595
Protein Synthesis Is Linear 596

17.3 Cracking the Code 596

Random Synthetic RNA Polymers Direct Protein Synthesis in Cell Extracts 597

RNA Polymers of Defined Sequence Complete the Code 598

The Genetic Code Is Validated in Living Cells 601

17.4 Exceptions Proving the Rules 601

Evolution of the Translation Machinery Is a Mystery 601

Mitochondrial tRNAs Deviate from the Universal Genetic Code 602

HIGHLIGHT 17-1 Evolution The Translation Machinery 604

Initiation and Termination Rules Have Exceptions 604

How We Know 608

Transfer RNA Connects mRNA and Protein 608

Proteins Are Synthesized from the N-Terminus to the C-Terminus 609

The Genetic Code In Vivo Matches the Genetic Code In Vitro 610

18 Protein Synthesis 615**Moment of Discovery** Harry Noller, on discovering the functional importance of ribosomal RNA 615**18.1 The Ribosome 616**

The Ribosome Is an RNA-Protein Complex Composed of Two Subunits 616

Ribosomal Subunits Associate and Dissociate in Each Cycle of Translation 619

The Ribosome Is a Ribozyme 620

The Ribosome Structure Facilitates Peptide Bond Formation 621

HIGHLIGHT 18-1 Evolution Mitochondrial Ribosomes: A Window into Ribosome Evolution? 622**18.2 Activation of Amino Acids for Protein Synthesis 624**

Amino Acids Are Activated and Linked to Specific tRNAs 625

Aminoacyl-tRNA Synthetases Attach the Correct Amino Acids to Their tRNAs 625

The Structure of tRNA Allows Accurate Recognition by tRNA Synthetases 625

Proofreading Ensures the Fidelity of Aminoacyl-tRNA Synthetases 627

18.3 Initiation of Protein Synthesis 629

Base Pairing Recruits the Small Ribosomal Subunit to Bacterial mRNAs 629

HIGHLIGHT 18-2 Technology Genetic Incorporation of Unnatural Amino Acids into Proteins 630

Eukaryotic mRNAs Recruit the Small Ribosomal Subunit Indirectly 631

A Specific Amino Acid Initiates Protein Synthesis 631

Initiation in Bacterial Cells Requires Three Initiation Factors 633

Initiation in Eukaryotic Cells Requires Additional Initiation Factors 634

Some mRNAs Use 5' End-Independent Mechanisms of Initiation 635

18.4 Elongation of the Polypeptide Chain 638

Peptide Bonds Are Formed in the Translation Elongation Stage 638

Substrate Positioning and the Incoming tRNA Contribute to Peptide Bond Formation 638

The GTPase EF-G Drives Translocation by Displacing the A-Site tRNA 640

GTP Binding and Hydrolysis Regulate Successive Elongation Cycles 640

18.5 Termination of Protein Synthesis and Recycling of the Synthesis Machinery 642

Completion of a Polypeptide Chain Is Signaled by an mRNA Stop Codon 642

Ribosome Recycling Factor Prepares Ribosomes for New Rounds of Translation 644

Fast and Accurate Protein Synthesis Requires Energy 644

Antibiotics and Toxins Frequently Target the Protein Synthesis Cycle 644

HIGHLIGHT 18-3 Medicine Toxins That Target the Ribosome 647**18.6 Translation-Coupled Removal of Defective mRNA 647**

Ribosomes Stalled on Truncated mRNAs Are Rescued by tmRNA 647

Eukaryotes Have Other Mechanisms to Detect Defective mRNAs 651

18.7 Protein Folding, Covalent Modification, and Targeting 654

Some Proteins Fold Spontaneously, and Others Need Help from Molecular Chaperones 654

Covalent Modifications Are Common in Newly Synthesized Proteins	654
Proteins Are Targeted to Correct Locations during or after Synthesis	655
Posttranslational Modification of Many Eukaryotic Proteins Begins in the Endoplasmic Reticulum	655
Glycosylation Plays a Key Role in Eukaryotic Protein Targeting	657
Signal Sequences for Nuclear Transport Are Not Removed	657
Bacteria Also Use Signal Sequences for Protein Targeting	658

How We Know 661

The Ribosome Is a Ribozyme	661
Ribosomes Check the Accuracy of Codon-Anticodon Pairing, but Not the Identity of the Amino Acid	662

IV Regulation

19 Regulating the Flow of Information 667

Moment of Discovery Lin He, on discovering that microRNA overexpression accelerates tumor development 667

19.1 Regulation of Transcription Initiation 669

Activators and Repressors Control RNA Polymerase Function at a Promoter	669
Transcription Factors Can Function by DNA Looping	670
Regulators Often Work Together for Signal Integration	673
Gene Expression Is Regulated through Feedback Loops	673
Related Sets of Genes Are Often Regulated Together	675
Eukaryotic Promoters Use More Regulators Than Bacterial Promoters	676
Multiple Regulators Provide Combinatorial Control	677
Regulation by Nucleosomes Is Specific to Eukaryotes	678

19.2 The Structural Basis of Transcriptional Regulation 678

Transcription Factors Interact with DNA and Proteins through Structural Motifs	679
Transcription Activators Have Separate DNA-Binding and Regulatory Domains	682

19.3 Posttranscriptional Regulation of Gene Expression 684

Some Regulatory Mechanisms Act on the Nascent RNA Transcript	684
Small RNAs Sometimes Affect mRNA Stability	685
Some Genes Are Regulated at the Level of Translation	685
Some Covalent Modifications Regulate Protein Function	686
Gene Expression Can Be Regulated by Intracellular Localization	686

HIGHLIGHT 19-1 Medicine Insulin Regulation: Control by Phosphorylation 688

Protein Degradation by Ubiquitination Modulates Gene Expression	690
---	-----

How We Know 693

Plasmids Have the Answer to Enhancer Action 693

20 The Regulation of Gene Expression in Bacteria 697

Moment of Discovery Bonnie Bassler, on her discovery of interspecies quorum sensing 697

20.1 Transcriptional Regulation 698

The lac Operon Is Subject to Negative Regulation	698
The lac Operon Also Undergoes Positive Regulation	703

HIGHLIGHT 20-1 Technology Classical Techniques in the Analysis of Gene Regulation 704

CRP Functions with Activators or Repressors to Control Gene Transcription	707
Transcription Attenuation Often Controls Amino Acid Biosynthesis	708
The SOS Response Leads to Coordinated Transcription of Many Genes	710

20.2 Beyond Transcription: Control of Other Steps in the Gene Expression Pathway 712

RNA Sequences or Structures Can Control Gene Expression Levels	712
Translation of Ribosomal Proteins Is Coordinated with rRNA Synthesis	716

HIGHLIGHT 20-2 A Closer Look T-Box Riboswitches 718

20.3 Control of Gene Expression in Bacteriophages 720

Bacteriophage Propagation Can Take One of Two Forms	721
---	-----

Differential Activation of Promoters Regulates Bacteriophage λ Infection 722
 The λ Repressor Functions as Both an Activator and a Repressor 723
 More Regulation Levels Are Invoked during the Bacteriophage λ Life Cycle 725

How We Know 728

TRAPped RNA Inhibits Expression of Tryptophan Biosynthetic Genes in *Bacillus subtilis* 728
 Autoinducer Analysis Reveals Possibilities for Blocking Cholera Infection 729

21 The Transcriptional Regulation of Gene Expression in Eukaryotes 733

Moment of Discovery *Tracy Johnson, on discovering that pre-mRNA splicing requires specific histone acetylation* 733

21.1 Basic Mechanisms of Eukaryotic Transcriptional Activation 734

Eukaryotic Transcription Is Regulated by Chromatin Structure 735
 Positive Regulation of Eukaryotic Promoters Involves Multiple Protein Activators 736

HIGHLIGHT 21-1 A Closer Look The Intertwining of Transcription and mRNA Splicing 738

Transcription Activators and Coactivators Help Assemble General Transcription Factors 740

21.2 Combinatorial Control of Gene Expression 743

Combinatorial Control of the Yeast *GAL* Genes Involves Positive and Negative Regulation 743

HIGHLIGHT 21-2 Technology Discovering and Analyzing DNA-Binding Proteins 744

Yeast Mating-Type Switches Result from Combinatorial Control of Transcription 746
 Combinatorial Mixtures of Heterodimers Regulate Transcription 747
 Differentiation Requires Extensive Use of Combinatorial Control 749

21.3 Transcriptional Regulation Mechanisms Unique to Eukaryotes 751

Insulators Separate Adjacent Genes in a Chromosome 751
 Some Activators Assemble into Enhanceosomes 752

Gene Silencing Can Inactivate Large Regions of Chromosomes 752

Imprinting Enables Selective Gene Expression from One Allele Only 753

HIGHLIGHT 21-3 A Closer Look Gene Silencing by Small RNAs 754

Dosage Compensation Balances Gene Expression from Sex Chromosomes 755

Steroid Hormones Bind Nuclear Receptors That Regulate Gene Expression 756

Nonsteroid Hormones Control Gene Expression by Triggering Protein Phosphorylation 758

How We Know 762

Transcription Factors Bind Thousands of Sites in the Fruit Fly Genome 762

Muscle Tissue Differentiation Reveals Surprising Plasticity in the Basal Transcription Machinery 763

22 The Posttranscriptional Regulation of Gene Expression in Eukaryotes 767

Moment of Discovery *Judith Kimble, on discovering that noncoding regions of mRNA regulate cell fate* 767

22.1 Posttranscriptional Control inside the Nucleus 768

Alternative Splicing Controls Sex Determination in Fruit Flies 769
 Multiple mRNA Cleavage Sites Allow the Production of Multiple Proteins 771
 Nuclear Transport Regulates Which mRNAs Are Selected for Translation 772

22.2 Translational Control in the Cytoplasm 773

Initiation Can Be Down-Regulated by Phosphorylation of eIF2 773
 The 3'UTR of Some mRNAs Controls Translational Efficiency 774
 Upstream Open Reading Frames Control the Translation of *GCN4* mRNA 777
 Translational Efficiency Can Be Controlled by mRNA Degradation Rates 777

22.3 The Large-Scale Regulation of Groups of Genes 779

Some Sets of Genes Are Regulated by Pre-mRNA Splicing in the Nucleus 779
 5'UTRs and 3'UTRs Coordinate the Translation of Multiple mRNAs 779

Conserved AU-Rich Elements in 3'UTRs Control Global mRNA Stability for Some Genes 779	Bacterium, <i>Escherichia coli</i> A-6
HIGHLIGHT 22-1 Evolution Regulation of Splicing in Response to Stress 781	Early Studies of <i>E. coli</i> as a Model Organism A-6
22.4 RNA Interference 782	Life Cycle A-6
MicroRNAs Encoded in Eukaryotic Genomes Target mRNAs for Gene Silencing 783	Genetic Techniques A-7
Short Interfering RNAs Target mRNAs for Degradation 784	<i>E. coli</i> as a Model Organism Today A-7
RNAi Pathways Regulate Viral Gene Expression 786	Budding Yeast, <i>Saccharomyces cerevisiae</i> A-8
RNAi Provides a Useful Tool for Molecular Biologists 787	Early Studies of Yeast as a Model Organism A-8
HIGHLIGHT 22-2 Medicine Viral Takeover Using a Cell Type-Specific miRNA 788	Life Cycle A-8
22.5 Putting It All Together: Gene Regulation in Development 790	Genetic Techniques A-8
Development Depends on Asymmetric Cell Divisions and Cell-Cell Signaling 790	Yeast as a Model Organism Today A-9
Early Development Is Mediated by Maternal Genes 793	Bread Mold, <i>Neurospora crassa</i> A-10
Segmentation Genes Specify the Development of Body Segments and Tissues 795	Early Studies of <i>Neurospora</i> as a Model Organism A-10
Homeotic Genes Control the Development of Organs and Appendages 796	Life Cycle A-10
Stem Cells Have Developmental Potential That Can Be Controlled 798	Genetic Techniques A-11
22.6 Finale: Molecular Biology, Developmental Biology, and Evolution 801	<i>Neurospora</i> as a Model Organism Today A-11
The Interface of Evolutionary Biology and Developmental Biology Defines a New Field 801	Nematode, <i>Caenorhabditis elegans</i> A-12
Small Genetic Differences Can Produce Dramatic Phenotypic Changes 801	Early Studies of <i>C. elegans</i> as a Model Organism A-12
How We Know 804	Life Cycle A-12
A Natural Collaboration Reveals a Binding Protein for a 3'UTR 804	Genetic Techniques A-13
Little RNAs Play a Big Role in Controlling Gene Expression 805	<i>C. elegans</i> as a Model Organism Today A-13
Everything Old Is New Again: Beauty at the Turn of a Developmental Switch 806	Mustard Weed, <i>Arabidopsis thaliana</i> A-14
Appendix: Model Organisms A-1	Early Studies of <i>Arabidopsis</i> as a Model Organism A-14
A Few Organisms Are Models for Understanding Common Life Processes A-1	Life Cycle A-14
Three Approaches Are Used to Study Human Disease A-2	Genetic Techniques A-15
	<i>Arabidopsis</i> as a Model Organism Today A-15
	Fruit Fly, <i>Drosophila melanogaster</i> A-16
	Early Studies of <i>Drosophila</i> as a Model Organism A-16
	Life Cycle A-16
	Genetic Techniques A-17
	<i>Drosophila</i> as a Model Organism Today A-17
	House Mouse, <i>Mus musculus</i> A-18
	Early Studies of the Mouse as a Model Organism A-18
	Life Cycle A-18
	Genetic Techniques A-18
	The Mouse as a Model Organism Today A-19
	Glossary G-1
	Solutions to Problems S-1
	Index I-1

Preface

As teachers, we know that undergraduate science education is evolving. Simply conveying facts does not produce a scientifically literate student, a long-held perception now reinforced by numerous studies. Students of science need more: a better window on what science is and how it is done, a clear presentation of key concepts that rises above the recitation of details, an articulation of the philosophical underpinnings of the scientific discipline at hand, exercises that demand analysis of real data, and an appreciation for the contributions of science to the well-being of humans throughout the world. As undergraduate science educators rise to these challenges, we are faced with both higher numbers of students and declining resources. How can we all do more with less?

Textbooks are an important part of the equation. A good textbook must now be more than a guide to the information that defines a discipline. For instructors, a textbook must organize information, incorporate assessment tools, and provide resources to help bring a discipline to life. For students, a textbook must relate science to everyday experience, highlight the key concepts, and show each student the process that generated those key concepts.

This book had its genesis at a meeting of the authors in Napa Valley in January, 2006. From the outset, we set ambitious goals designed to address the key challenges we face as teachers:

Students see science as a set of facts rather than an active human endeavor.

Molecular biology has a wealth of great stories to tell. We wanted to convey the excitement that drives modern molecular biology, the creativity at the bench, and the genuine wonder that takes hold as the workings of a new biological process are revealed. This theme is set in the first chapter, dedicated in large measure to an introduction to the scientific process. Every chapter then begins with a *Moment of Discovery*, highlighting a researcher's own description of a memorable moment in their career. After Chapter 1, every chapter ends with a *How We Know* section with stories relating the often circuitous path to a new insight. Additional anecdotes—scientists in action—are woven into the text and the accompanying Highlights. As students turn these pages, the laboratories and the people behind the discoveries will never be far away.

20

The Regulation of Gene Expression in Bacteria



Bonnie Bassler [Source: Paul Fetters Photography.]

Moment of Discovery

Science for me is all about those moments of clarity, when years of struggling to figure something out finally pay off with an incredible insight about how nature works. I am fascinated by how bacterial cells communicate with each other in the process known as quorum sensing. Quorum-sensing bacteria make, release, and detect chemical signal molecules that increase in concentration in proportion to increasing cell population numbers. Cells respond to these chemicals with synchronous population-wide changes in behavior; community behavior allows bacteria to carry out tasks that could never be accomplished if a single bacterium acted alone. We suspect that the evolution of cell-cell communication in bacteria is one of the first steps in the development of multicellular organisms.

Vibrio harveyi is a bioluminescent gram-negative marine bacterium that regulates light production in response to two distinct chemical "words" or autoinducers. As a new professor, I wanted to answer a question that had baffled the field for several years: Why does *V. harveyi* need two chemical signals for communication, when one should be sufficient? The identity of one autoinducer, AI-1 (autoinducer-1), had been determined, but the other, AI-2, remained an enigma. Our lab cloned the gene responsible for synthesizing AI-2, and sequenced the gene. At that time

◀ MOMENT OF DISCOVERY

Scientific breakthroughs represent the exhilarating culmination of a lot of hard work. Each chapter opens with a description of a significant breakthrough in molecular biology, as told by the scientist who made the discovery. The scientists featured in the Moments of Discovery are: Norm Arnheim, Bonnie Bassler, Steve Benner, James Berger, Carlos Bustamante, Jamie Cate, Joe DeRisi, Myron Goodman, Lin He, Tracy Johnson, Melissa Jurica, Judith Kimble, Judith Klinman, Robert Lehman, Tim Lohman, Steve Mayo, Harry Noller, Lorraine Symington, Jack Szostak, Robert Tjian, Jonathan Widom, and Wei Yang.

20.1 Transcriptional Regulation 698

20.2 Beyond Transcription: Control of Other Steps in the Gene Expression Pathway 712

20.3 Control of Gene Expression in Bacteriophages 720

How We Know

A Dreamy Night Ride on a California Byway Gives Rise to the Polymerase Chain Reaction

Brock, T.D., and H. Freese. 1969. *Thermus aquaticus* gen. n. and sp. n., a nonsporulating extreme thermophile. *J. Bacteriol.* 98:289–297.

Salki, R.K., S. Scherf, F. Falona, K.B. Mullis, G.T. Horn, H.A. Erlich, and N. Arnheim. 1985. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle-cell anemia. *Science* 230:1350–1354.

Large advances are sometimes driven by inspiration. In the spring of 1983, Kary Mullis was an employee of the Cetus Corporation in northern California. Hired in 1979 to synthesize oligonucleotides, Mullis discovered that as oligonucleotide synthesis became increasingly automated, he had more and more time to contemplate other projects. He became interested in methods to detect small sequence differences in human DNA, but initially did not make much progress. The idea for the polymerase chain reaction occurred to Mullis one night, in April 1983, as he drove with a friend up the coast. As he described it, he stopped the car and started drawing—DNA molecules hybridizing and lengthening, a chain reaction in which the products of one cycle became the templates for the next.

The first experiment was carried out a few months later, and the first report of the polymerase chain reaction came out in a 1985 paper in *Science* describing a new procedure for detecting the hemoglobin mutation



FIGURE 2 Hot springs in Yellowstone National Park, one of which is shown here, were the source of the bacterium *Thermus aquaticus*. [Source: FoxTV/Dreamstime.]

◀ HOW WE KNOW

Each chapter ends with a How We Know section that combines fascinating stories of research and researchers with experimental data for students to analyze.

Students often view science as a completed story with an oversimplified script.

Data can take a researcher in unexpected directions. An experiment designed to test one hypothesis can end up testing something quite different. The analysis of real data is a fundamental skill to be honed by every student of science. We have tried to address this need aggressively. Each chapter in this text features a challenging set of problems, including at least one requiring the analysis of data from the literature. Many of these are linked to the tales of discovery found in the How We Know sections. At the same time, each chapter ends with a section on Unanswered Questions, providing just a sampling of the endless challenges that remain for those with the motivation to tackle them.

Unanswered Questions

The study of RNA processing reactions has been a long-standing and active area of research, yet much remains to be deciphered.

1. **Why do introns exist?** We don't yet know why there are introns and whether introns are ancient or more recent acquisitions in genes. Some introns have been found to encode regulatory RNA molecules that function in the processing of rRNA and in the control of gene expression levels (see Chapter 22). Whether these regulatory RNAs are a cause or a result of the presence of introns is not known.

Although the origin of introns may remain uncertain, further insights about their roles in the continuing evolution of genomes will be exciting and may shed light on diseases that result from inaccurate intron removal and processing.

2. **How does alternative splicing work?** Experimental methods, including microarray technology and genome-wide transcript sequencing, have revealed an abundance of alternative splicing in mammalian cells. However, the mechanics of such molecular gymnastics have yet to be determined. Future research will focus on how splicing is regulated, the frequency with which genes are alternatively spliced, and the roles of splicing regulation in disease.

◀ UNANSWERED QUESTIONS

A short section at the end of each chapter describes important areas still open to discovery, showing students that even well-covered subjects, such as nucleic acid structure and DNA replication, are far from fully explored.

Data Analysis Problem

Zhang, B., M. Gallegos, A. Puoti, E. Durkin, S. Fields, J. Kimble, and M.P. Wickens. 1997. A conserved RNA-binding protein that regulates sexual fates in the *C. elegans* hermaphrodite germ line. *Nature* 390:477–484.

11. The collaborative effort by Judith Kimble, Marvin Wickens, and colleagues, as described in their 1997 paper (see How We Know), resulted in discovery of the FBF proteins that bind the 3'UTR of the mRNA from the *fem-3* developmental regulatory gene of nematodes. Compare the three-hybrid strategy used in this study (see How We Know, Figure 1) with the three-hybrid method presented in Chapter 7.

(a) How was the three-hybrid method modified in the Kimble and Wickens study?

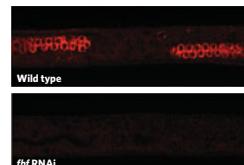


FIGURE 1

Six different RNA sequences were screened for FBF-1 binding. The normal PME sequence (UCUUG) gave a positive response. The other five sequences were two 5-nucleotide sequences, an iron response element, a segment containing 30 consecutive A residues, and a 573-nucleotide RNA sequence derived from HIV.

(b) Of these controls, which would make the best case for specific binding of the PME by FBF-1?

(c) Why might the other controls be useful?

The investigators used immunofluorescence to detect expression of the FBF-1 protein in wild-type nematodes, as shown in the upper panel of Figure 1. All cells illuminated are in the germ line. The dark spots are cell nuclei.

(d) What conclusion can you draw from the protein expression pattern in the upper panel?

The lower panel in Figure 1 shows results for an animal treated with RNAi directed at the gene for FBF-1 (*fbf*). The expression of FBF-1 is essentially abolished.

(e) Given the function of FBF-1 in the germ line, what is the likely effect of this RNAi treatment on the germ line of the treated animals?

END-OF-CHAPTER PROBLEMS ▶

Extensive problem sets at the end of each chapter give students the opportunity to think about and work with the chapter's key ideas. Each problem set concludes with a Data Analysis Problem, giving students the critical experience of interpreting real research data. Solutions to the problems can be found at the back of the book.

Students get lost in the details. Presenting the major concepts clearly, in the text as well as in the illustrations, is just as important as teaching students how science is done. We have worked to use straightforward language and a conversational writing style to draw students in to the material. We have collaborated closely with our illustrator, Adam Steinberg, to create clean, focused figures. Featured Key Conventions highlight the implicit but often unstated conventions used when sequences and structures are displayed, and in naming biological molecules.

KEY CONVENTION

When an amino acid sequence is given, it is written and read from the N-terminus to the C-terminus, left to right.

◀ KEY CONVENTIONS

In brief paragraphs, the Key Conventions clearly lay out for students some fundamental principles often glossed over.

ILLUSTRATIONS ▶

Good figures should speak for themselves. We have worked to make our figures simple and to keep the figure legends as brief as possible. The illustrations in the text are the product of close collaboration with our colleague Adam Steinberg. Together with the talented artists at Dragonfly Media Group, Adam has helped to hone and implement our vision.

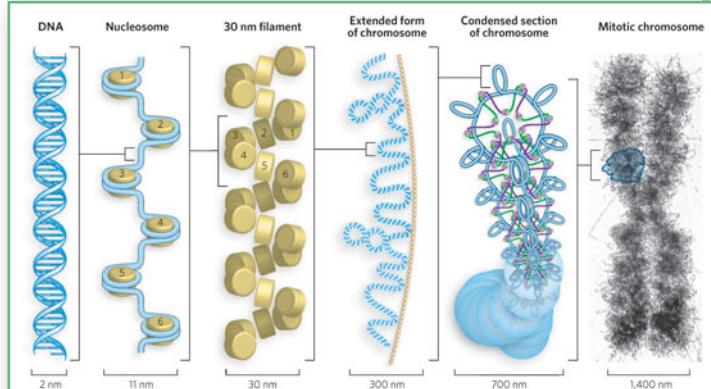


FIGURE 10-16 Higher-order DNA compaction in a eukaryotic chromosome. This model shows the levels of organization that could provide the observed degree of DNA compaction in the chromosomes of eukaryotes. First the DNA is wrapped around histone octamers, then H1 stimulates formation of the 30 nm filament. Further levels of organization are not well understood but seem to involve

further coiling and loops in the form of rosettes, which also coil into thicker structures. Overall, progressive levels of organization take the form of coils upon coils upon coils. It should be noted that in cells, the higher-order structures (above the 30 nm filament) are unlikely to be as uniform as depicted here. [Source: Photo from G. F. Bahar/Biological Photo Service.]

Students see evolution as an abstract theory. Every time a molecular biologist studies a developmental pathway in nematodes, identifies key parts of an enzyme active site by determining what parts are conserved between species, or searches for the gene underlying a human genetic disease, he or she is relying on evolutionary theory. *Evolution is a foundational concept, upon which every discipline in the biological sciences is built.* In this text, evolution is a theme that pervades every chapter, beginning with a major section in Chapter 1 and continuing as the topic of many Highlights and chapter segments.

HIGHLIGHT 1-1 EVOLUTION

Observing Evolution in the Laboratory

The bacterium *Deinococcus radiodurans* has a remarkable capacity to survive the effects of ionizing radiation (IR or γ rays). A human being would be killed by exposure to 2 Gy (1 Gy = 100 rads) of IR, but cultures of *Deinococcus* routinely survive 5,000 Gy with no lethality. *Deinococcus* is a desert dweller, and this characteristic reflects its adaptation to the effects of desiccation. After months or years of dry conditions, the bacterium can reconstitute its genome quickly when conditions favorable for growth return. That same extraordinary capacity for DNA repair is put to use after exposure to IR.

How long does it take for a bacterium to evolve extreme resistance to IR? A recent study demonstrates that, as *Escherichia coli*, the common laboratory bacterium, can acquire this resistance by directed evolution. Twenty cycles of exposure to enough IR to kill more than 99% of the cells, each cycle followed by the outgrowth of survivors, produce an *E. coli* population with a radiation resistance approaching that of *Deinococcus*. The entire selection process can be achieved in less than a month. Complete genomic sequencing of cells isolated from the evolved populations typically reveals 40 to 80 mutations. The answer to survival varies from cell to cell, with different cells displaying different arrays of mutations, even when they come from the same evolved population. In just a single, small bacterial culture, evolution takes many paths, and a variety of solutions are found that lead to a single trait.

This is just one of many experiments demonstrating that dramatic changes in microorganisms can be readily generated and observed in the labora-

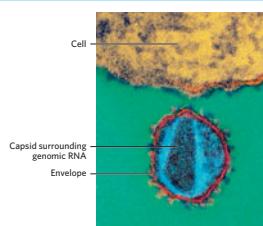


FIGURE 1 HIV is a retrovirus. Like other retroviruses, it has an RNA genome condensed within a proteinaceous capsid. The capsid is surrounded by a spherical lipid envelope derived from its host cell's cytoplasmic (plasma) membrane. Its relationship to other retroviruses is not just structural, but is embedded in definable ways in its chromosome. [Source: Hans Gelderblom/Getty Images.]

determined. Scientists did not have to characterize this novel and very dangerous virus from scratch. Its small genome held all the clues that science needed for a rapid understanding of its infection cycle and the development of effective treatments. As a retrovirus, HIV had a clear evolutionary relationship to other viruses that were already known and understood (Figure 1). Molecular biologists relied on its evolutionary connectedness with these other viruses to understand how HIV might be countered. The

◀ HIGHLIGHTS

These discussions are designed to enhance student understanding and appreciation of the relevance of each chapter's material. There are four categories of Highlights:

- Medicine explores diseases that arise from defects in biochemical pathways, and how concepts uncovered in molecular biology have contributed to drug therapies and other treatments.
- Technology focuses on cutting-edge molecular biology methods.
- Evolution reveals the role of molecular biology research in understanding key biological processes and the connections among organisms.
- A Closer Look examines a wide variety of additional, intriguing topics.

Experimental Techniques

As researchers, we know that it is critical to understand the benefits and limitations of experimental techniques. We strive to give students a sense of how an experiment is designed and what makes the use of a particular technique or model organism appropriate. For your reference, the techniques covered in this book are:

Ames test 420	Electrophoretic mobility shift assay (EMSA) 704
Chemical protection footprinting 704	Electroporation 221
Chemical modification interference 704	Epitope tagging 242
Chemical synthesis of nucleic acids 206	Haplotype analysis 271
ChIP-Chip 345	Immunoprecipitation 244
ChIP-Seq 345	Linkage analysis 274
Chromatography	Localization of GFP fusion proteins 242
Affinity chromatography 100	Mass spectrometry 280
Using terminal tags 240	Northern blotting 203
Using tandem affinity purification (TAP) tags 246	Nuclear magnetic resonance (NMR) 125
Column chromatography 100	Optical trapping 344
Gel-exclusion chromatography 100	Photolithography 248
Ion-exchange chromatography 100	Phylogenetic analysis 272
Thin-layer chromatography 580	Phylogenetic profiling 281
Colony blot hybridization 202	Polymerase chain reaction (PCR) 226
Detecting A=T-rich segments of DNA by denaturation analysis 201	Real-time PCR (quantitative PCR, qPCR) 228
DNA cloning 217	Protein chips 282
DNA cloning with artificial chromosomes (BACs, YACs) 222	Protein localization via indirect immunofluorescence 242
DNA footprinting 704	Recombinant protein expression 233
DNA genotyping (DNA fingerprinting, DNA profiling, STR analysis) 230	Reverse transcriptase PCR (RT-PCR) 228
DNA library creation (cDNA, genomic) 224	RNA interference (RNAi) 782
DNA microarrays 248	RNA-Seq 278
DNA sequencing	Selection and screening 221
Automated Sanger sequencing 232	Site-directed mutagenesis 239
Genome sequencing techniques 261	Somatic cell nuclear transfer (SCNT) 533
Next generation sequencing 234	Southern blotting 203
Pyrosequencing 234	Transformation 221
Reversible terminator sequencing 235	Western blotting 243
Sanger sequencing 228	X-ray crystallography 121
Electrophoresis	Yeast two-hybrid analysis 246
Agarose gel electrophoresis 203	Yeast three-hybrid analysis 247
Isoelectric focusing 280	
Pulsed field gel electrophoresis (PFGE) 223	
Sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) 101	
Two-dimensional gel electrophoresis 280	

Media and Supplements

A full package of media resources and supplements provides instructors and students with innovative tools to support a variety of teaching and learning approaches.

eBook

The online version of the textbook combines the contents of the printed book with electronic study tools, including instant navigation to any section or page of the book, bookmarks, highlighting, note-taking, instant search for any term, pop-up key term definitions, and a spoken glossary. Instructor features include the ability to add notes or files to any page and to share these notes with students. The eBook is available for purchase through your bookstore or online at www.ebooks.bfwpub.com.

Companion Website: www.whfreeman.com/cox

This free companion website includes study tools for students and valuable instructor resources (also available on the Instructor Resource DVD), including:

- Fully optimized JPEG files of every figure, photo, and table in the text, with enhanced color, higher resolution, and enlarged fonts. Figures are available labeled and unlabeled.
- PowerPoint files of figures from each chapter.

Acknowledgments

This is a first edition text, representing our best effort to synthesize this complex and ever-shifting field and to contribute to the broadening requirements of 21st-century education in molecular biology. We welcome your comments and suggestions. We thank our many colleagues whose input has helped shape this book:

Steven Ackerman, *University of Massachusetts Boston*
Byron J. Adams, *Brigham Young University*
Ravi Allada, *Northwestern University*
David K. Asch, *Youngstown State University*
Karen Beemon, *Johns Hopkins University*
Craig Berezowsky, *University of British Columbia*
Sandra Berry-Lowe, *University of Colorado, Colorado Springs*
Prakash H. Bhuta, *Eastern Washington University*
Judith L. Campbell, *California Institute of Technology*
Mitchell Chernin, *Bucknell University*
Paul Cliften, *Utah State University*
Melanie Cocco, *University of California, Irvine*
Claire Crommiller, *University of Virginia*
Kelly J. Cude, *Western Washington University*
Sumana Datta, *Texas A&M University*
Elizabeth A. De Stasio, *Lawrence University*
Jeff DeJong, *University of Texas, Dallas*
Michele L. Engel, *Claremont McKenna College*
Jeffrey Fillingham, *Ryerson University*
Samuel Galewsky, *Millikin University*
Thomas Geoghegan, *University of Louisville*
Amy Hark, *Muhlenberg College*
Daniel P. Herman, *University of Wisconsin, Eau Claire*
David C. Higgs, *University of Wisconsin, Parkside*
Manju M. Hingorani, *Wesleyan University*
Mitchell M. Holland, *Pennsylvania State University*
Margaret Hollingsworth, *SUNY Buffalo*
Barbara Chadwick Hoopes, *Colgate University*
Constance Jeffery, *University of Illinois at Chicago*
Melissa S. Jurica, *University of California, Santa Cruz*
Anuj Kumar, *University of Michigan*
Justin P. Kumar, *Indiana University*

Hong Li, *Florida State University*
Carol S. Lin, *Columbia University*
Curtis M. Loer, *University of San Diego*
Susan T. Lovett, *Brandeis University*
Mitch McVey, *Tufts University*
Scott Moye-Rowley, *University of Iowa*
Katsu Murakami, *Pennsylvania State University*
Frank Naya, *Boston University*
James B. Olesen, *Ball State University*
Robert Osuna, *SUNY Albany*
Anthony J. Otsuka, *Illinois State University*
Donna L. Pattison, *University of Houston*
Veronica Pereira, *University of Toronto, Mississauga*
Marie C. Pizzorno, *Bucknell University*
David H. Price, *University of Iowa*
Gerry A. Prody, *Western Washington University*
Charles W. Putnum, *University of Arizona*
Claire A. Rinehart, *Western Kentucky University*
Phillip E. Ryals, *University of West Florida*
David Samols, *Case Western Reserve University*
Lillie L. Searles, *University of North Carolina-Chapel Hill*
Konstantin Severinov, *Rutgers University*
David Shub, *SUNY Albany*
Rey A. Sia, *SUNY Brockport*
Gary R. Skuse, *Rochester Institute of Technology*
Paul Keith Small, *Eureka College*
Claus Tittiger, *University of Nevada, Reno*
Akif Uzman, *University of Houston-Downtown*
Lori L. Wallrath, *University of Iowa*
Q. Tian Wang, *University of California, Berkeley*
Joanna Wysocka-Diller, *Auburn University*

This book would not have been possible without the support of our publishers at W. H. Freeman. A book of this sort is an undertaking measured not just in hours, but in sleepless nights, almost-met deadlines, conference calls, and occasional levity. It is an enterprise where teachers sometimes become students. The needed guidance has been provided by an exceptionally talented team of editors and copy editors. Kate Ahr Parker has overseen the effort from the beginning. Few human beings are as gifted in the art of articulating urgency with grace. Erica Frost, Erica Champion, and Betsy Dilernia have been our development editors. Guided by their capable hands, first draft chapters have been created, reworked, broken up and sometimes merged. They provided encouragement and pointed out deficiencies. They have been our partners throughout, scrutinizing every word we produced. As the project progressed, the work was honed and the chapters integrated with the help of Brook Soltvedt and Linda Strange. Both Brook and Linda are long-time veterans of the *Lehninger Biochemistry* series, and their expertise added immeasurably to the final product in your hands. In the end, they have managed the impressive feat of merging three voices into one. We are extraordinarily grateful to all of the editors for their dedication to this project. We are fortunate to have had the benefit of their insights and expertise.

The artwork for this book was a labor of love handled by Adam Steinberg and the artists at Dragonfly Media. Adam is also a *Lehninger Principles of Biochemistry* veteran, and his experience and skill is evident on almost every page of this book. He worked hand in hand with the authors to create illustrations that convey concepts concisely and in a unified style.

Our thanks also go to the consummate professionals who ensured the high quality production of the book: art director Diana Blume; project editor, Jane O'Neill; production coordinator Susan Wein; senior illustration coordinator Bill Page; photo editor Cecilia Varas; photo researcher Elyse Rieder; and media and supplements editors, J. D. Bullard, Patrick Shriner, and Marni Rolfes. We greatly appreciated their flexibility and creativity working with complex material and ever-shifting schedules.

We express our appreciation for our colleagues, friends, and families for their patience and support.

Last but certainly not least, we are grateful to the Moments of Discovery authors, who shared some of their favorite scientific career moments with us. Each of them provided valuable time and effort for this project, and helped us add a personal touch to every chapter.



From left to right:
Adam Steinberg,
Mike Cox, Betsy
Dilernia, Kate Ahr
Parker, Erica
Pantages Frost,
Jennifer Doudna,
Erica Champion, and
Mike O'Donnell.

Michael M. Cox
Jennifer A. Doudna
Michael O'Donnell
December 2010

This page intentionally left blank

Studying the Molecules of Life



Jack Szostak [Source: Courtesy of Jussi Puikkonen.]

Moment of Discovery

A big question in the origin of life concerns how primitive cells might have evolved. My own approach to this question involved lots of discussions with Irene Chen and others in my lab about how lipid vesicles containing RNA, which might mimic a simple self-replicating life form, could be capable of dividing. In other words, as the amount of genetic material increased through replication, *how would the increased RNA content affect the physical properties of the vesicle?* We envisioned that osmotic pressure might make vesicles

grow by extracting lipids from neighboring vesicles, ultimately leading to division by rupture and resealing. This idea seemed pretty far out, though, until Irene began doing experiments with vesicles containing lipids bearing fluorescent dyes. We could encapsulate RNA inside the vesicles and watch the vesicles change in size (or not) under different conditions by following the level of fluorescence as a function of vesicle surface area. Irene found that empty vesicles or vesicles “swollen” with RNA were stable over time, but when she mixed them together, the swollen vesicles started to grow by stealing lipid molecules from neighboring empty vesicles! So the system worked exactly as we had imagined, demonstrating that vesicle growth and division is a process that can occur without requiring any catalytic function.

More recently, we found that vesicles loaded with RNA can also take up nucleotides from the surrounding environment, disproving an old idea that it would be hard for primitive cells to survive by taking up small molecules, including negatively charged nucleotides, from the environment. It’s been very exciting to find that each potential roadblock to primitive cellular replication that we’ve explored so far can be overcome, often without requiring specialized catalysts or input energy.

—**Jack Szostak**, on his discovery of self-dividing vesicles that mimic growing cells

1.1 The Evolution of Life on Earth 2

1.2 How Scientists Do Science 12

Born in the second half of the twentieth century, molecular biology has only recently come of age. Broadly speaking, **molecular biology** is the study of essential cellular macromolecules, including DNA, RNA, and proteins, and the biological pathways between them. Over the decades, molecular biology has become firmly associated with the structure, function, and regulation of information pathways at the molecular level. All of the processes required to reliably pass genetic information from one generation to another and from DNA to RNA to protein are included in this area of study. Of the requirements for life, it is the information in our genetic material that links all organisms to each other and documents their intertwined history. The biological information pathways that maintain, use, and transmit that information are the focus of this book.

Molecular biology may have a relatively short history, but its impact on the human experience is already considerable. Medicine, modern agriculture, forensic science, and many other endeavors rely on technologies developed by molecular biologists. Our current understanding of information pathways has given rise to diagnostic tests for genetic diseases, forensic DNA analysis, crops with improved yields and resistance to disease, new cancer therapies, an unprecedented ability to track pandemics, new wastewater treatment methods, new approaches to the generation of energy, and much more. Many of these advances are chronicled throughout this textbook.

This first chapter introduces three of the most important themes that link the book's topics. First, there are two key requirements for life: **biological information**, the genetic instructions that shape every living cell and virus, and **catalysis**, a capacity to accelerate critical molecular processes. Molecular biology deals with both, and much of the discipline focuses on the interplay between information-containing polymers (nucleic acids and proteins) and the enzymes that catalyze and regulate their synthesis, modification, function, and degradation.

The second theme is **evolution**. Many of the processes we will consider can be traced back billions of years, and a few can be traced to the last universal common ancestor. Genetic information is a kind of molecular clock that can help define ancestral relationships among species. Shared information pathways connect humans to every other living organism on Earth and to all the organisms that came before.

Third, we look at molecular biology as a scientific endeavor. Any scientific discipline is a construct not only of the knowledge it has generated but also of the human processes behind that knowledge. Molecular biology has both an inspirational history and a promising future, to

be forged by contributors as yet unnamed. Breakthroughs rely on more than technology and ideas: they require an understanding of the scientific process and are informed by the struggles of the past.

1.1 The Evolution of Life on Earth

All organisms on Earth are connected by an evolutionary journey spanning more than 3 billion years. The diversity of life we see around us is the sum of a limitless number of **mutations**, changes in genetic information that are usually subtle but sometimes dramatic. When Charles Darwin proposed that natural selection acts on variation in populations, he had no knowledge of the mechanisms that give rise to that variation. Such mechanisms lie at the heart of modern molecular biology.

What Is Life?

Almost anyone can distinguish a living organism from an inanimate object. However, a rigorous scientific description of life is harder to achieve. Life differs from nonlife in identifiable ways, as summarized in **Figure 1.1**. Organisms move, reproduce, grow, and alter their environment in ways that inanimate objects cannot. But such characteristics alone provide an unsatisfying definition of life, particularly when a few of them may be shared by inanimate substances. In 1994, the United States National Aeronautics and Space Administration (NASA) convened a panel to consider the question, "What is life?" A simple definition resulted: *Life is a chemical system capable of Darwinian evolution*. The importance of evolutionary theory to all biological sciences gains full expression in this concise statement.

Every living system we know about has several requirements for its existence. Two of these—raw materials and energy—are supplied by a home planet endowed with an abundance of both. Molecules in Earth's life forms are made up largely of the elements carbon, nitrogen, oxygen, and hydrogen. These are the smallest and most abundant atoms that can make, respectively, four, three, two, and one covalent bonds with other atoms. The molecules formed by these elements tend to be quite stable and can be very complex. The energy required for life is derived from the sun. Plants and photosynthetic microorganisms collect and store the energy derived from sunlight in the chemical bonds of complex biomolecules.

A third requirement for a living system is an envelope, creating a barrier between the living and inanimate worlds and establishing a means of selective

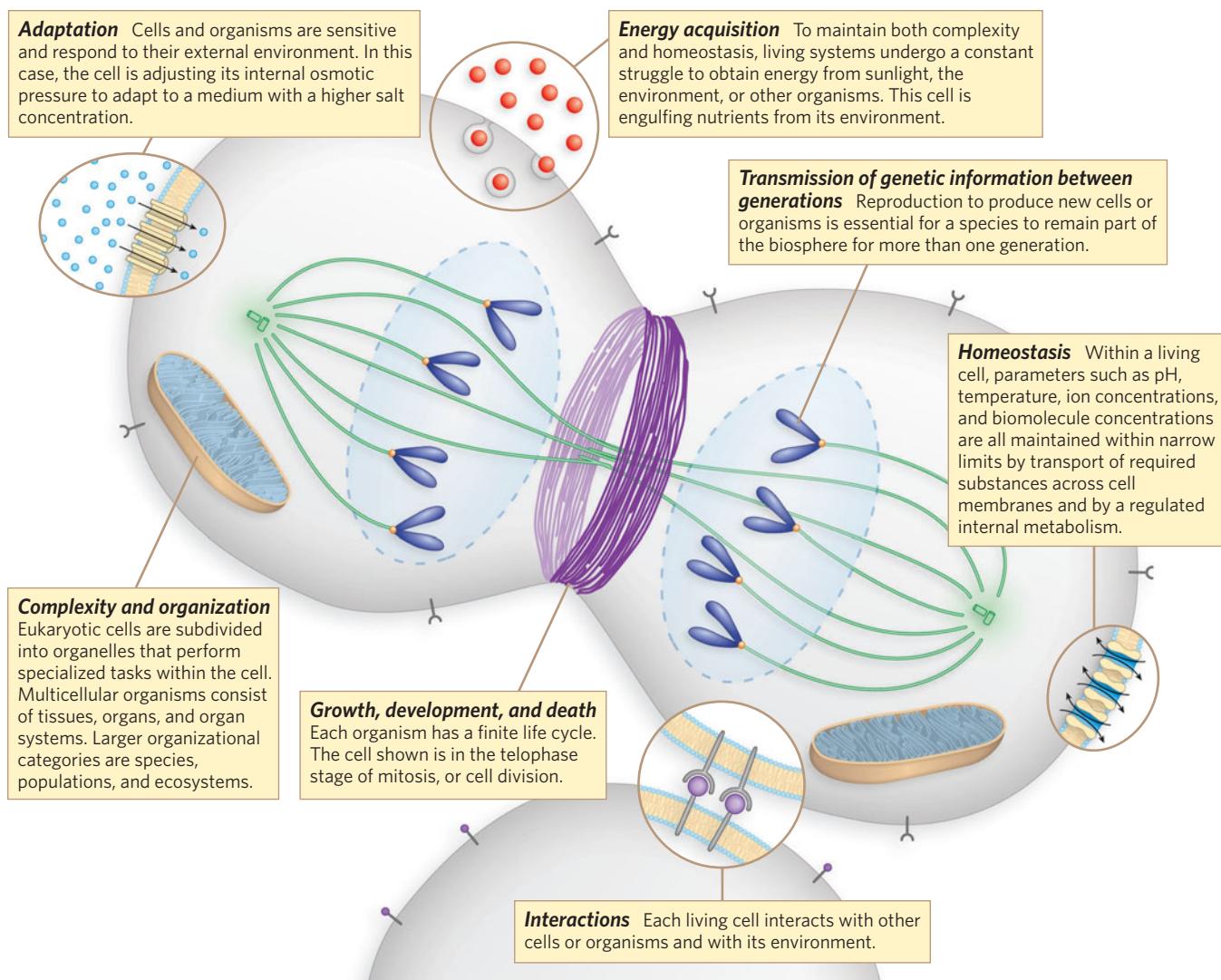


FIGURE 1-1 Characteristics of living systems. Each characteristic distinguishes living organisms from inanimate matter.

interaction for a cell and its environment. The work of Jack Szostak, chronicled in this chapter's Moment of Discovery, may be replicating some key evolutionary moments that led to enveloped living systems (Figure 1-2).

The final two requirements—catalysis and biological information—are particularly important, truly distinguishing a living organism from an inanimate object. These requirements are the domain of molecular biology. The energy transactions that support homeostasis and enable the transmission of genetic information from one generation to the next are initiated by powerful catalysts called **enzymes**. Enzymes are highly specific, and each enzyme accelerates only one or a small number of chemical reactions. Most

enzymes are proteins, although a few catalytic RNA molecules play important roles in cells. The catalysts that a particular organism possesses define which reactions can occur in that organism. Enzymes determine what a cell takes in for nourishment, how fast the cell grows, how it discards wastes, how it constructs its cellular membranes, how it responds to other cells, and how it reproduces.

The presence of enzymes in a cell depends on the faithful transmission of the genetic information that encodes them from one generation to the next. Enzymes, as well as the myriad other proteins and RNA molecules that regulate their synthesis and function, are the actual molecular targets of evolution. When a

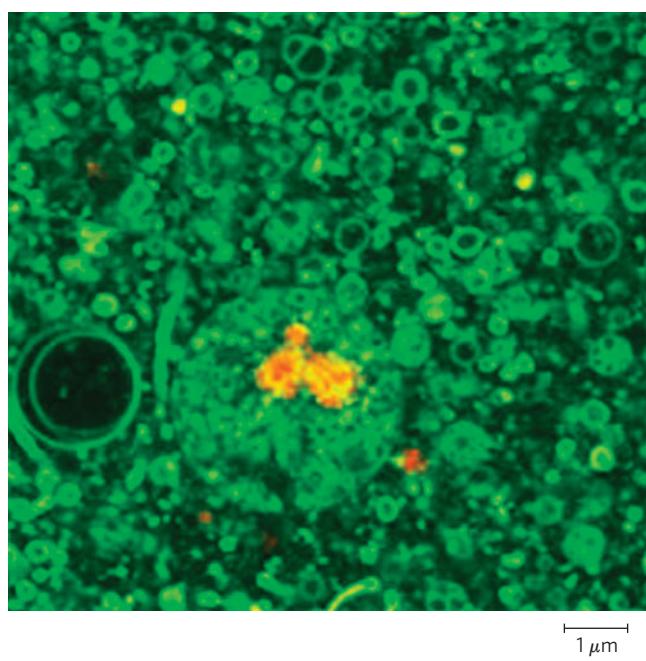


FIGURE 1-2 RNA-containing vesicles undergoing division. Division is driven by the uptake of lipids by RNA-containing vesicles (red). [Source: Courtesy of J. W. Szostak; first published in S. Graham, *Scientific American*, Oct. 4, 2003.]

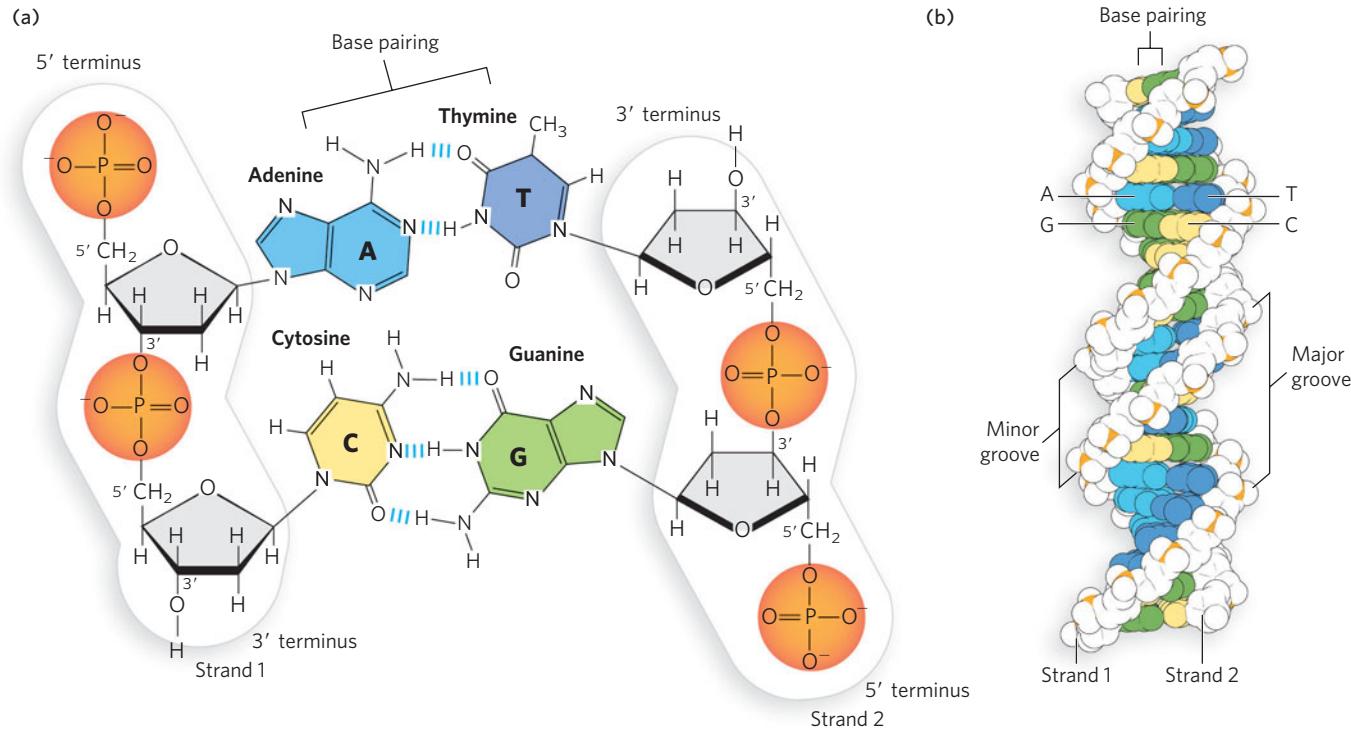


FIGURE 1-3 DNA structure. Because of its structural properties, DNA is well suited for long-term information storage. (a) The G≡C and A=T base pairs are similarly sized, allowing them to stack in any sequence. Complementary base pairing facilitates replication and transmission from one generation to the next. (b) The double-helical structure and

cell acquires a new function, it generally reflects the presence of a new enzyme or set of enzymes, or an alteration in the regulation or function of an existing enzyme or process. The new functions arise through changes in genes that are shaped by evolutionary processes. In the biosphere of today, DNA is the standard macromolecule for the long-term storage and transmission of biological information. It is exquisitely adapted to that function (Figure 1-3). However, as we'll see, there were probably stages in the evolution of life when DNA did not serve as the primary genetic library in living systems.

Evolution Underpins Molecular Biology

In 1973, the geneticist Theodosius Dobzhansky published an article in the *American Biology Teacher* entitled “Nothing in Biology Makes Sense Except in the Light of Evolution.” This sentiment has special meaning in molecular biology, because the pathways and processes of molecular biology give rise to the genetic variation on which natural selection acts (Figure 1-4). They also inform the ongoing investigations into how life arose on Earth.

base stacking confer stability. Major and minor helical grooves in the structure provide access to genetic information for a wide range of DNA-binding proteins. The uniform structure of the DNA backbone allows the synthesis of very long polymers.

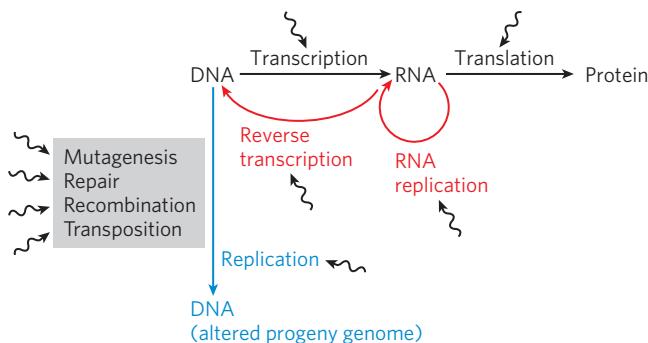


FIGURE 1-4 Pathways of biological information flow. In

almost all living systems, information is stored in DNA, then transcribed into RNA, which is processed and translated into protein. DNA is replicated to prepare for cell division, and various mechanisms of DNA repair exist to protect against mutations. The transfer and maintenance of genetic information are regulated at each of these stages. RNA viruses and retroviruses are exceptions; they store their genetic information in RNA and can use RNA replication and reverse transcription to maintain their genomes. The short, wavy arrows represent points of regulation.

Mutations can be as simple as a change in a single base pair of DNA or RNA or as substantial as the inversion, deletion, or insertion of large segments of nucleic acid. As we will be discussing in detail, errors can arise during replication (Chapter 11), and DNA damage can lead to permanent mutation when repair systems (Chapter 12) go awry. Larger chromosomal changes can arise from recombination (Chapter 13) or transposition (Chapter 14). Some mutations affect genes directly; others affect the ways in which the DNA is transcribed into RNA or RNA is processed or translated (Chapters 15–18). Relatively minor changes in genes involved in regulatory processes (Chapters 19–22) can give rise to dramatic changes in the organism; this realization has created a new field, essentially a modern merger of the fields of evolutionary and developmental biology, dubbed “evo-devo.” All the processes that contribute to information transfer are highly, but not perfectly, accurate, and the slow accumulation of alterations is inevitable. Many organisms even have mechanisms to speed up the pace of mutational change, which they draw upon in times of stress.

An understanding of these processes has also given us insights into the origins of life and the process of evolution. Continuing explorations of RNA structure (Chapter 6) and metabolism (Chapters 15 and 16) have informed new theories of prebiotic evolution. The genetic code (Chapter 17) provides a particularly vivid look at the shared history of every organism on Earth.

Molecular biology has provided the enzymes that make most of the methods of biotechnology possible (Chapter 7). These increasingly powerful methods for studying the genes of many different organisms allow us to trace their evolution. Through modern genomics (Chapter 8), molecular biology is opening a window onto evolution that Charles Darwin would marvel at.

The interrelationship of molecular biology and evolution is of more than academic interest. Human beings exist in a world where every organism continues to evolve. Microorganisms, with their short life cycles, evolve most rapidly (Highlight 1-1). Of special concern are human pathogens, as well as the microorganisms, fungi, insects, and other organisms that affect our food crops, livestock, and water supply. Molecular biology provides essential tools for use in tracking pandemics, investigating new microbial pathogens, identifying the genes underlying human genetic diseases, solving crimes, tracing the origin of diseases, treating cancer, and engineering microorganisms for new purposes in bioremediation and bioenergy. All of these efforts rely heavily on the concepts of evolutionary biology.

Life on Earth Probably Began with RNA

About 4.6 billion years ago, the sun and Earth and the other planets and asteroids of our solar system were formed. Within the first billion years of our planet’s existence, life appeared on its surface. How did this happen, and how likely is it that this has happened on other, similar worlds? Modern geologists, paleontologists, and molecular biologists are slowly piecing together the history of life on Earth from the rich trove of clues in the geologic, fossil, and genomic records. A plausible sequence emerges, providing a wide range of hypotheses that can be tested using modern chemical and physical methods.

The first few hundred million years were a time of prebiotic chemistry (Figure 1-5). No life was present, but chemical reactions were happening everywhere. A modestly reducing atmosphere containing primarily water, methane, ammonia, hydrogen, nitrogen, and carbon dioxide, in reactions driven by the constant stream of energy coming from the sun, was slowly yielding more complex molecules such as simple sugars, amino acids, and nucleotide bases. The accumulation of organic material was supplemented by a multitude of collisions between the early Earth and meteors laden with organic materials. Prebiotic chemistry is being studied by a large community of researchers, and a small sampling of their work is presented in How We Know at the end of this chapter.

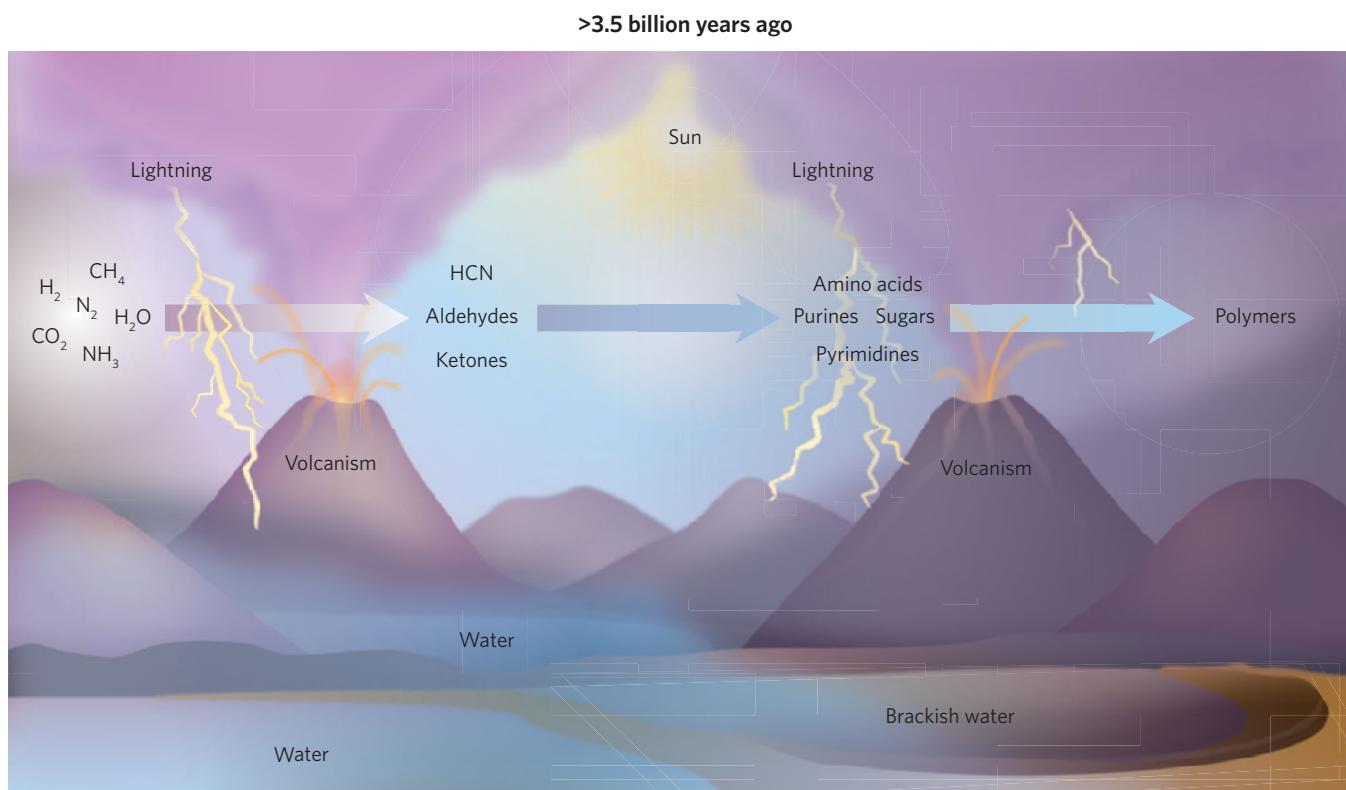


FIGURE 1-5 Prebiotic chemistry. Over hundreds of millions of years, and with constant energy input from solar radiation, volcanism, and other sources, the molecular constituents of Earth's early atmosphere were converted to

a range of more complex organic molecules and polymers. The resulting tarry substance may have coated the planet's surface and turned bodies of water into concentrated and complex solutions.

Over a period of millions of years, the accumulation of reaction products yielded a soup containing molecules and polymers. As they grew increasingly complex, particular polymers acquired the capacity to duplicate themselves. The first self-replicating polymer, possessing two of the key requirements for life—catalysis and biological information—might be considered the first life form.

We do not know what this first “living” polymer was. However, modern molecular biology has given us many reasons to think that RNA either was the first self-replicator or arose as a much-improved descendant of that first self-replicator. RNA has a capacity for both catalysis and information storage that has made it indispensable for life, from its beginnings to the present time.

The **RNA world hypothesis** was first proposed as a stage in evolution by molecular biologists Carl Woese, Francis Crick, and Leslie Orgel, in separate papers published during the late 1960s. The hypothesis describes a living system (or set of living systems) based on RNA. In this system, a variety of RNA enzymes could catalyze all of the reactions needed to synthesize the molecules required for life from simpler molecules

available in the environment. The RNA enzymes would include replicators to duplicate all of the RNA catalysts. The “RNA organism,” out of equilibrium with its surroundings, would have to be defined by a boundary. The experiments of Szostak and colleagues show one way in which lipid-enclosed RNA systems can arise (see How We Know).

Four more-recent lines of evidence have added much breadth and depth to the plausibility of the proposal. The first was the discovery by Thomas Cech and Sidney Altman, in the early 1980s, of **catalytic RNAs**, or **ribozymes**—enzymes that are made of RNA instead of protein. Thus we learned that some extant RNA molecules catalyze reactions and thus possess both of the key conditions for life—biological information and catalysis. These ribozymes catalyze a relatively narrow range of reactions, such as the cleavage and ligation of other RNA molecules.

The second line of supportive research demonstrated that artificially constructed RNA molecules can catalyze almost any imaginable reaction needed in a living system—certainly a range of reactions much broader than those attributable to existing ribozymes

HIGHLIGHT 1-1 EVOLUTION

Observing Evolution in the Laboratory

The bacterium *Deinococcus radiodurans* has a remarkable capacity to survive the effects of ionizing radiation (IR, or γ rays). A human being would be killed by exposure to 2 Gy (1 Gy (gray) = 100 rads) of IR, but cultures of *Deinococcus* routinely survive 5,000 Gy with no lethality. *Deinococcus* is a desert dweller, and this characteristic reflects its adaptation to the effects of desiccation. After months or years of dry conditions, the bacterium can reconstitute its genome quickly when conditions favorable for growth return. That same extraordinary capacity for DNA repair is put to use after exposure to IR.

How long does it take for a bacterium to evolve extreme resistance to IR? A recent study demonstrates that, *Escherichia coli*, the common laboratory bacterium, can acquire this resistance by directed evolution. Twenty cycles of exposure to enough IR to kill more than 99% of the cells, each cycle followed by the outgrowth of survivors, produce an *E. coli* population with a radiation resistance approaching that of *Deinococcus*. The entire selection process can be achieved in less than a month. Complete genomic sequencing of cells isolated from the evolved populations typically reveals 40 to 80 mutations. The answer to survival varies from cell to cell, with different cells displaying different arrays of mutations, even when they come from the same evolved population. In just a single, small bacterial culture, evolution takes many paths, and a variety of solutions are found that lead to a single trait.

This is just one of many experiments demonstrating that dramatic changes in microorganisms can be readily generated and observed in the laboratory within short periods of time. The work is inspired by our experience with human pathogens. When AIDS appeared as a new threat to human health in the early 1980s, the power of evolutionary theory was quickly on display. The causative agent, HIV, was soon isolated and its genomic sequence

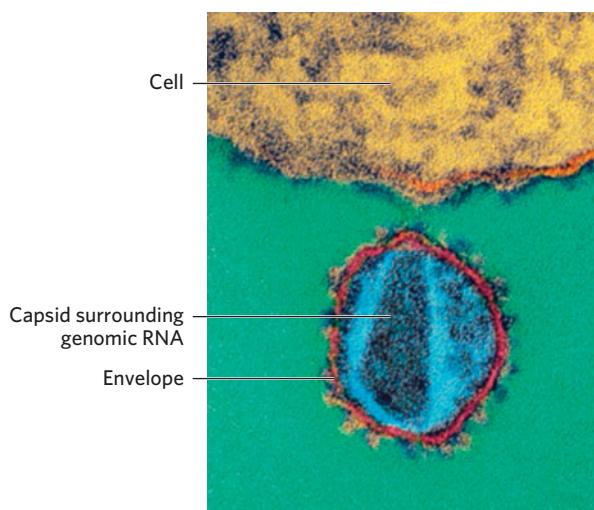


FIGURE 1 HIV is a retrovirus. Like other retroviruses, it has an RNA genome condensed within a proteinaceous capsid. The capsid is surrounded by a spherical lipid envelope derived from its host cell's cytoplasmic (plasma) membrane. Its relationship to other retroviruses is not just structural, but is embedded in definable ways in its chromosome. [Source: Hans Gelderblom/Getty Images.]

determined. Scientists did not have to characterize this novel and very dangerous virus from scratch. Its small genome held all the clues that science needed for a rapid understanding of its infection cycle and the development of effective treatments. As a retrovirus, HIV had a clear evolutionary relationship to other viruses that were already known and understood (Figure 1). Molecular biologists relied on its evolutionary connectedness with these other viruses to understand how HIV might be countered. The key enzymes that would be viable drug targets were immediately evident. One result was the development of treatments at an unprecedented rate, ranging from AZT to protease inhibitors that have saved or lengthened millions of lives (see Highlights 5-2 and 14-3).

today. Early RNA molecules could clearly have catalyzed all of the reactions required to set up a primordial cellular metabolism.

The third and fourth discoveries have further broadened our perspectives on RNA function. We now know that in ribosomes, the large ribonucleoprotein

complexes that translate RNA into protein, the RNA is the active component with the capacity to catalyze protein synthesis (Figure 1-6; see also Chapter 18, Moment of Discovery). Finally, and most recently, RNA sequences capable of simple forms of self-replication have been discovered.

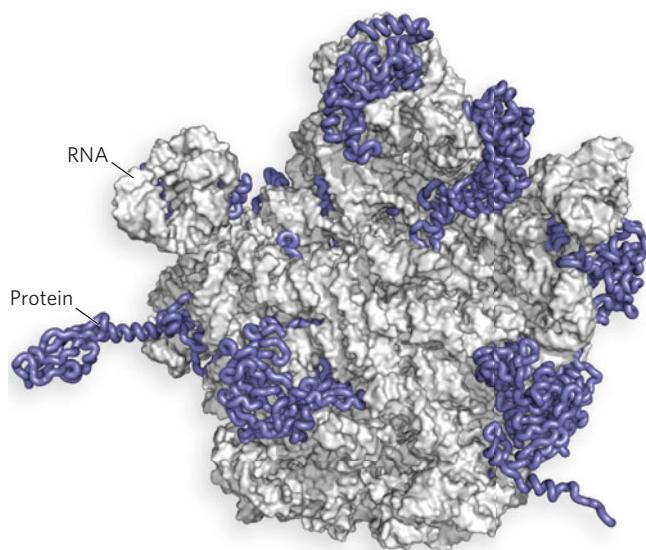


FIGURE 1-6 The 50S subunit of a bacterial ribosome. The gray parts of the subunit are RNA and the blue parts are protein. The structure is a huge ribozyme that evolved for the synthesis of protein.

Ongoing research thus makes it possible to visualize a highly plausible sequence of events unfolding on the pathway from prebiotic soup to living systems. An RNA world came into being and gradually became more complex. An RNA capable of reliable self-replication may have been the first living entity. Self-replicators would have diversified to synthesize other ribozymes, leading to an RNA-based metabolism capable of providing a greater supply of needed RNA precursors. As the RNA molecules in that world increased in size and structural complexity, a need for stabilization and auxiliary functions arose. Peptides were synthesized to neutralize the negative charges of the phosphates in the RNA backbone, to stabilize RNA structure in other ways, and to augment early metabolism. As more peptides were synthesized, some with catalytic activities arose. Proteins gradually supplanted RNA as catalysts, because the greater catalytic potential of proteins yielded an advantage. The protein world emerged, but not without leaving important vestiges of the RNA world (ribosomes and some other RNA catalysts), as we find them today.

The Last Universal Common Ancestor Is the Root of the Tree of Life

Countless nascent life forms probably arose from the primordial soup, along with many biological advances that improved their fitness. Successful combinations of RNA catalysts gave way to systems based on protein catalysts. A systematized genetic code and improve-

ments in catalytic efficiency appeared. Additional changes facilitated cellular metabolism and reproduction. Protein synthesis was systematized through the evolution of an efficient genetic code. RNA became more specialized for information storage and transmission. Cell membranes appeared, eventually including mechanisms to selectively transport materials into and out of the cell as needed. And some processes became regulated. In this way, a variety of primitive cells may have evolved. One, sometimes called **LUCA (last universal common ancestor)**, won the struggle and ultimately gave rise to all life on Earth (Figure 1-7).

LUCA is a special source of fascination for molecular biologists. Although LUCA probably lived more than 3 billion years ago, our speculation about what this cell was like is informed by experiment. One approach is to determine the minimum protein and genetic requirements for life. Attempts to create a minimal life form, either by reconstituting basic components or by taking bacteria and stripping them of all unnecessary parts, are underway in laboratories around the world. These experiments are not only defining properties that must have been present in LUCA, they are also setting the stage for the laboratory generation of engineered cells.

Another approach to understanding LUCA is to survey all types of living systems on Earth to determine which genes or characteristics are universal. The only genes that are truly universal in living systems are those encoding the cellular machinery for protein synthesis and some components of RNA transcription. LUCA must have had essentially the same genetic code found in all organisms today. To support protein synthesis and RNA synthesis, a simple metabolism must have been present that allowed for the uptake of chemical energy

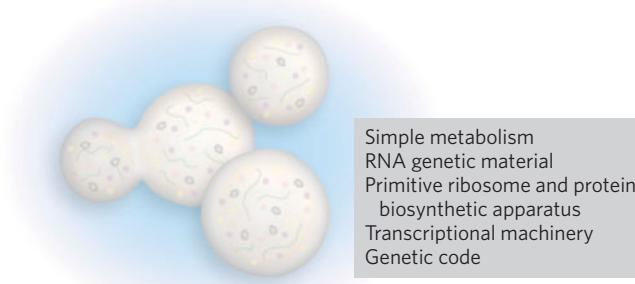


FIGURE 1-7 A possible last universal common ancestor.

LUCA probably had a simple metabolism and a form of transcriptional machinery to replicate its RNA genome. A primitive ribosome and protein-biosynthetic apparatus would have used the same universal genetic code found in all modern organisms.

and its use to synthesize amino acids, nucleotides, and whatever lipids existed in the cell membrane from precursors available in the environment. The study of LUCA is described in more detail in Chapter 8.

The appearance of LUCA signaled the beginning of biological evolution on Earth. New types of cells gradually appeared, and new environments were exploited. The first cells were probably **chemoheterotrophs**, capable of taking up organic molecules from their surroundings and converting them to the molecules needed to support protein and RNA synthesis. DNA, perhaps appearing first in viruses, gradually supplanted RNA as the most stable platform for the long-term storage and transmission of genetic information. DNA replication and systems for the segregation of replicated DNA chromosomes into daughter cells evolved.

The early single-celled organisms derived from LUCA diversified to inhabit all niches in the ecosystem of this early Earth. The diversification gradually generated the three major groups of organisms that we recognize today: **bacteria**, **archaea**, and **eukaryotes** (Figure 1-8).

Many additional events helped shape the life we see around us. Notably, photosynthesis appeared about 2.5 billion years ago, as evidenced by the sudden rise in the concentration of atmospheric oxygen documented in the geologic record. As cells engulfed other cells, some endosymbiotic relationships developed and

became permanent. The engulfed cells became organelles within their hosts more than 1 billion years ago, and we see these organelles today as chloroplasts and mitochondria. Loose clusters of unicellular organisms led to cell specialization, and more permanent assemblies produced multicellular organisms. Diversification of body plans became more rapid about 600 million years ago, eventually generating all the major types of organisms we observe today.

Evolution by Natural Selection Requires Variation and Competition

Charles Darwin (1809–1882) was one of the most influential thinkers in history, and his name is forever associated with the concept of evolution. In his famous book *On the Origin of Species*, published in 1859, Darwin developed several general observations and ideas, laying out the evidence he had collected during and after his now famous voyage on the *Beagle*. He documented the variation among individuals in a population and the inheritance of the variations by offspring. He noted that individuals in a population compete for resources. He argued that those individuals best adapted to exploit the prevailing resources are the ones most likely to survive and reproduce. These ideas together constitute a mechanism for evolution that can be described by the term **natural selection**.

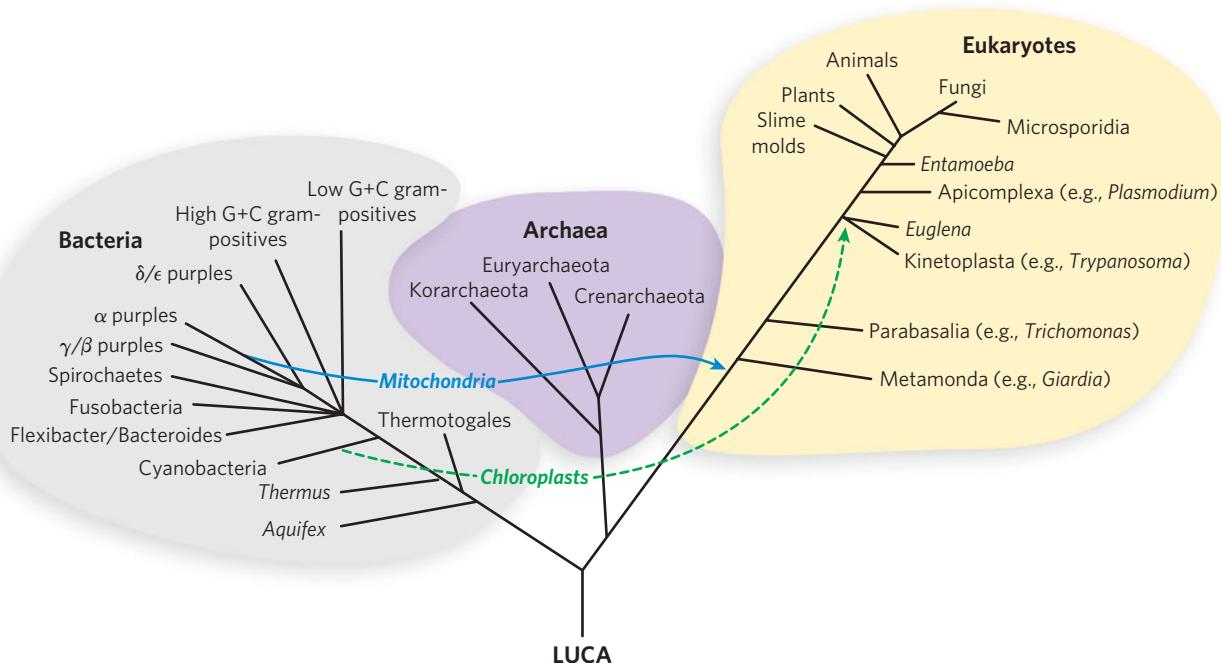
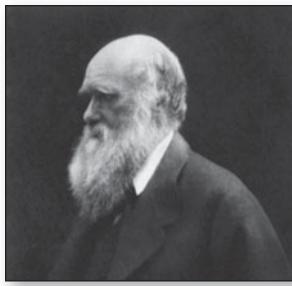


FIGURE 1-8 The universal tree of life. A current version of the tree is shown, with branches for the three main groups of known organisms: bacteria, archaea, and eukaryotes. [Source: Adapted from J. R. Brown, "Universal tree of life," in *Encyclopedia of Life Sciences*, Wiley InterScience (online), 2005.]



Charles Robert Darwin, 1809-1882

predecessors and contemporaries (see How We Know). Darwin's study of finches and other organisms introduced the idea of branching evolution (**Figure 1-9**), which ultimately led to the idea that all life on Earth has a common ancestor. For natural selection to work, evolution must be gradual, with no discontinuities. All of these ideas coalesced, in *The Origin of Species*, into an internally consistent and compelling story describing the development of life in its many forms. Darwin's definition of the mechanism by which all of this occurred—natural selection—was the crowning achievement.

Natural selection depends on two characteristics of a population: variation and competition (**Figure 1-10**). However, the source of a population's variation eluded Darwin. The genetic program that exists in every organism was unknown to him, as were the mechanisms by which it was handed down from one generation to the next. Darwin was unaware that the work that would eventually reveal these mechanisms had been begun by one of his contemporaries, Gregor Mendel (see Chapter 2).

Darwin's ideas have been expanded and developed into a modern synthesis of the theory of evolution, a

The Origin of Species had a tremendous and immediate influence on scientific thought, derived in part from the huge volume of work it described and in part from the story it told. Darwin contributed a detailed body of evidence to support a range of interconnected ideas—some his own, some borrowed from his

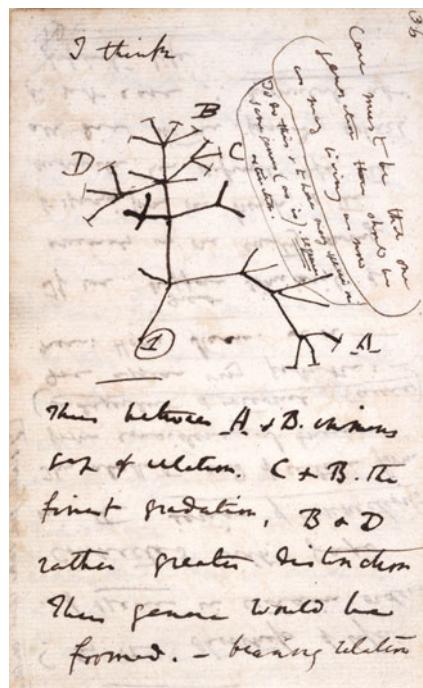


FIGURE 1-9 An evolutionary tree as sketched by Darwin in his 1837 notebook. [Source: Reproduced by kind permission of the Syndics of Cambridge University Library.]

direct outgrowth of the development of genetics in the early twentieth century. The rediscovery of Mendelian genetics, along with the concept of the gene developed by influential geneticists such as Thomas Hunt Morgan, J. B. S. Haldane, and Theodosius Dobzhansky, provided the necessary mechanism of inheritance. The concept of changes in the genetic program—mutations—was introduced, explaining the needed source of variation. By the 1940s, the theory of evolution could be stated in



FIGURE 1-10 Variation and competition. On the plains of Africa, predation eliminates the weakest individuals from a population. [Source: Gary Dublanko/Alamy.]

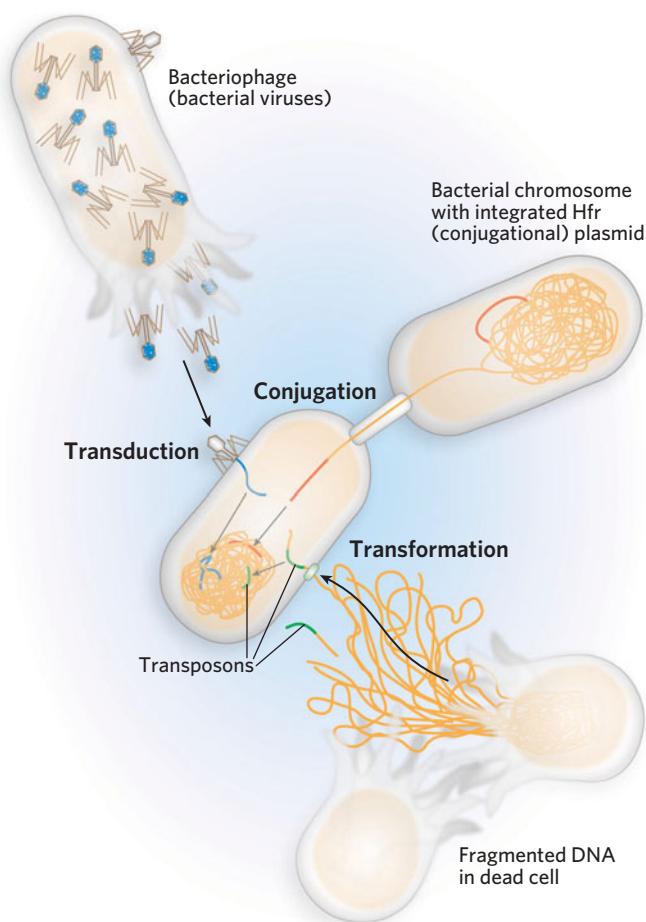


FIGURE 1-11 Horizontal gene transfer. Genetic material is transferred between organisms, especially bacteria, by several mechanisms. DNA may be taken up from the environment by transformation, transferred by viruses through transduction, or passed purposefully from one bacterium to another by conjugation. In all cases, the new DNA may be incorporated into the chromosome by recombination. The movement of genetic elements called transposons can augment the effects of all these processes, if transposons are part of the introduced DNA. These processes are described in more detail in Chapter 14.

more detail and was bolstered by more evidence on mechanisms than were conceivable in Darwin's day. Populations, as we now know, contain inherent genetic variation generated by random mutation and genetic recombination. The frequency of different forms of genes in a population changes from generation to generation as a result of several processes. Random genetic drift—changes in the frequency of a particular form of a gene—can occur, especially in small, isolated populations. Gene flow between different species can take place in a process called **horizontal gene transfer** (Figure 1-11) (see Chapters 8 and 14).

Darwin's theory of natural selection provides a mechanism that responds directly to the environment. Most of the genetic changes that do not kill an organism produce only small changes in protein function or expression, resulting in a small change in the whole organism. As a result, evolutionary change is usually gradual. Sufficient diversification leads to new species and, with time, to new genera and phyla.

On the most practical level, our connectivity through the tree of life has a critical effect on the study of molecular biology: we can learn about ourselves by studying other organisms (as described in detail in the Model Organisms Appendix). Even the simplest organisms have much to teach us about the inner workings of our own cells. As we'll see throughout this book, the processes involved in the flow of biological information, though common to all organisms, are often much more complex in eukaryotes than in bacteria. Much of our understanding of these processes is due to groundbreaking research on bacteria or yeast, followed by further research on more complex model organisms such as worms, insects, or mice (Figure 1-12). In this way, the elucidation of gene functions in yeast can lead to cures for human disease. Discoveries made in bacteria can generate improvements in human agriculture. Fruit flies instruct us about the intricacies of human cognition and the complexities of fetal development. Pandemics of the future can be predicted and tamed by studying the pathogens of the past. Each investigation into the molecular biology of an organism is made more valuable by the fact that all species are related through a shared evolutionary history.

As Darwin remarked in *The Origin of Species*, "There is grandeur in this view of life, with its several powers,



FIGURE 1-12 Similarities among organisms during development. As an example here, a human embryo (left) is compared with a mouse embryo (right). Although the adult forms differ greatly in appearance, the embryos reveal similarities in body plan and development. These similarities, and many more that exist on a molecular level, allow us to learn about ourselves through the analysis of model organisms. [Source: (left) © Last Refuge, Ltd./Phototake; (right) © Michael F. McElwaine.]

having been originally breathed into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved.”

SECTION 1.1 SUMMARY

- Living systems have definable characteristics and requirements. Catalysis and biological information are particularly important requirements for any life form.
- The first molecule that fulfilled the requirements of catalysis and biological information may have been a self-replicating RNA, according to the RNA world hypothesis.
- LUCA, the last universal common ancestor of all life now present on Earth, can be studied by identifying the common characteristics of living organisms and defining the minimal complement of genes necessary to support a living cell.
- Evolution by natural selection is a result of genetic variation within a population and competition between individuals for limited resources. Darwin’s theory of evolution by natural selection has been strengthened by modern studies that reveal the sources of genetic variation, mutation, and recombination.
- A common evolutionary heritage links all organisms, allowing the study of model organisms to aid in our understanding of ourselves.

1.2 How Scientists Do Science

Science provides a story about the natural universe around us. It is a story that inspires wonder and, at the same time, has enormous practical implications for every aspect of human existence. As is the case for any scientific discipline, success in molecular biology is defined to a large extent by the contributions a scientist makes to the larger scientific story. That growing story provides a context and community within which scientists frame every experiment.

The history and philosophical underpinnings of science have brought about guidelines and rules for new experimentation at every level of this intertwined enterprise. The scientific community is highly interactive, with ongoing discussions supplying both constraints and insights that can stimulate progress. The collective discussion is carried out in informal conversations, at scientific meetings, and, most importantly,

in the peer-reviewed scientific literature. A successful contribution is usually one that eventually appears in this literature.

Molecular biology is a scientific enterprise, giving rise both to information, as conveyed in the chapters of this text, and to future advances. An understanding of the scientific enterprise can help in our assimilation of existing information and accelerate success in any scientific effort.

Science Is a Path to Understanding the Natural Universe

Science is both a body of knowledge and a process for generating that knowledge. In the scientific community, scholars attempt to answer questions about the natural universe. Since the dawn of humanity, people have employed many approaches to understanding the world around them, guiding human development for millennia. The modern version of the scientific method, relying on careful observation and experimentation, has been widely applied for only about four centuries.

The scientific approach to discovery has at least three characteristics. First, science focuses only on the natural universe. The realm of science is thus limited to what we can observe and measure. Second, science relies on ideas that can be tested by experiments and on observations that can be reproduced. Finally, the experiments are carried out within an ever-expanding web of scientific theories that provide guidance and insight along the way.

Science makes one philosophical assumption: that forces and phenomena existing in the universe are not subject to capricious or arbitrary change. They can be understood by applying a systematic process of inquiry—the scientific method. The French biochemist Jacques Monod referred to this basic underlying assumption as the **postulate of objectivity**. Stated more simply, science cannot succeed in a universe that plays tricks on us. This assumption is made every time a scientist performs an experiment. If an experiment is repeated and the results are different, there is always a reason; something must have been different during the second experiment. When faced with seemingly contradictory results, every scientist is trained to figure out why. Often, inconsistencies have a trivial cause, such as a degraded or inactive reagent; but sometimes, they lead to great insights.

In modern science, an initial idea is called a **hypothesis**—a proposal that provides a reasonable explanation for one or more observations but is not yet substantiated by sufficient experimental tests to stand

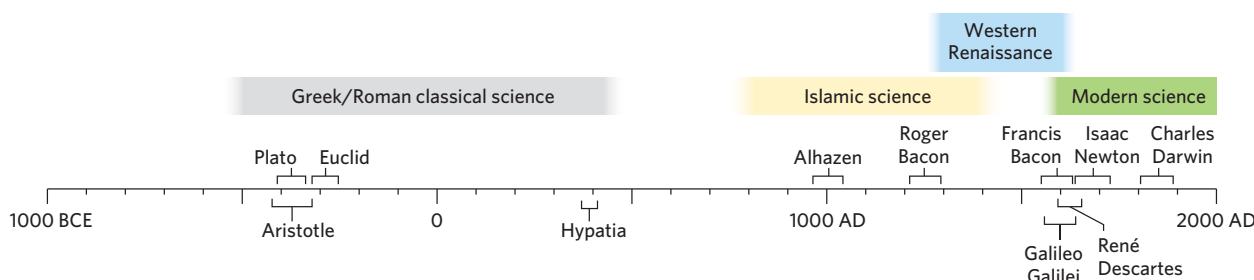


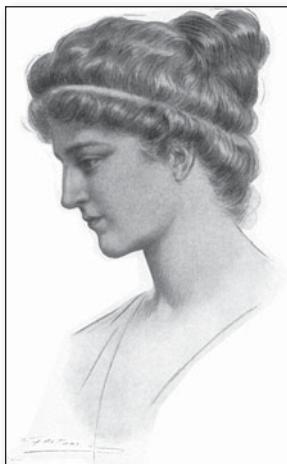
FIGURE 1-13 A timeline of development of the scientific method. The life spans of a few of the major contributors are shown.

up to rigorous critical examination. A **scientific theory** is much more than a hypothesis; it provides an explanation for a body of experimental observations and is thus a firm basis for further inquiry. When a scientific theory has been repeatedly tested and validated by many types of experiments, it can be considered a fact.

Scientific hypotheses and theories can also be defined by their presence in the peer-reviewed scientific literature. Papers are accepted or rejected for inclusion in that literature based on professional reviews by other working scientists. According to the Publishers Association of the United Kingdom, there are more than 16,000 peer-reviewed scientific journals worldwide, publishing some 1.4 million papers each year. This continuing rich harvest of information is ultimately the foundation of scientific progress.

The Scientific Method Underlies Scientific Progress

The first recorded efforts to systematize scientific inquiry have been credited to the classical Greeks, who were clearly influenced by other civilizations in Babylon, Assyria, Egypt, Persia, and elsewhere (Figure 1-13). Key thinkers replaced trial and error, chance discoveries, and appeals to the supernatural with a more formalized system of reasoning. Euclid, Aristotle, Plato, and others made significant advances by using inductive reasoning, making a broad conclusion from specific facts (e.g., my stove is hot, my brother's stove is hot, so all stoves are hot), and deductive reasoning, forming a conclusion from a premise (e.g., premise: all flying organisms are birds; observation: a crow flies; conclusion: a crow is a bird). Experiments played a minor role. The process worked reasonably well for mathematics and engineering, less well for most other pursuits. Scientific progress declined in the West with the decline of the Roman Empire, symbolized by the assassination in AD 415 of the last director of the great library of Alexandria in Egypt, the accomplished Greek mathematician and philosopher Hypatia.



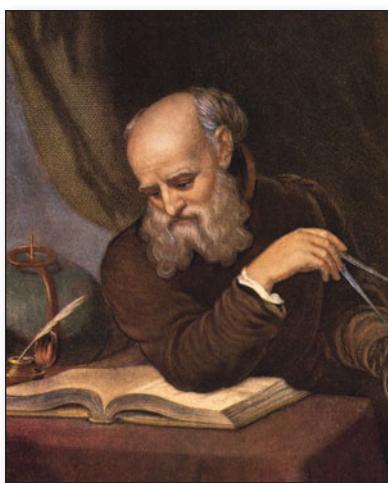
Hypatia, 370–415

A reawakening began with the emergence of Islamic science during the Middle Ages, building on Greek and Roman ideas with rational arguments combined more liberally with observations and experiments. The physicist Abū 'Alī al-Hasan ibn al-Hasan ibn al-Haytham (965–1040), known as Alhazen in the West, is regarded as the father of modern optics. Alhazen was particularly influential in refocusing the scientific process on reproducible experimentation. Latin translations of Arabic and Greek texts filtered into Europe, helping to stimulate the Renaissance. The Franciscan friar Roger Bacon (1214–1294) was inspired by these texts. He was among the first to describe a method of inquiry involving observation, hypothesis, and experiment, while also stressing independent repetition to verify results.

By the beginning of the seventeenth century, the flaws in Aristotelian philosophy had become clear to many scholars. René Descartes (1596–1650), Galileo Galilei (1564–1642), and Francis Bacon (1561–1626) led a revolution in scientific thinking. Frustrated with the impediments inherent in the Aristotelian system (e.g., if one encounters a bat, the premise that all flying organisms are birds is falsified), they laid the groundwork for the modern scientific method.



Alhazen, 965–1040



**Galileo Galilei,
1564–1642** [Source:
© Pictorial Press Ltd./
Alamy.]

In 1619, Descartes set down four rules for inquiry, which we paraphrase here: (1) never accept anything as true unless that truth can be clearly demonstrated; (2) reduce a problem to its parts; (3) begin with the simplest part and gradually work to the more complex; and (4) be thorough. Bacon divided inquiry into a somewhat similar series of steps, first gathering relevant observations, and then deriving ever more complex conclusions from them. Galileo combined careful observation and rational argument to develop new fields of inquiry.

The **scientific method** had matured by the beginning of the nineteenth century. In its most basic form, it begins with a question about nature. Relevant observations are made, and a hypothesis is crafted to explain them. Assumptions that underlie the hypothesis are defined, and each assumption is carefully tested by experimentation. Experimental results lead to the acceptance, rejection, or modification of the hypothesis. Additional experiments can lead to the acceptance of a new theory or fact about the natural world.

In modern molecular biology, this now classical version of the scientific method underlies huge numbers of advances. One example is the discovery of the DNA-synthesizing enzyme DNA polymerase by Arthur Kornberg and his coworkers in the 1950s (see Chapter 11). With the elucidation of DNA structure by James Watson and Francis Crick, Kornberg and others hypothesized that an enzyme must exist to synthesize this polymer. They assumed that the enzyme could be purified and studied, and its need for a template determined. To test these assumptions, they lysed cells, fractionated the extracts, and developed assays to measure the activity of the hypothetical enzyme in the presence of all plausible substrates and precursors. The effort paid off in 1956, with an assay that detected the incor-

poration of nucleotides into DNA polymers. DNA polymerase was subsequently purified, named, and thoroughly characterized by the same team, substantiating the hypothesis: the existence of a DNA polymerase became fact.

The Scientific Method Is a Versatile Instrument of Discovery

The path to scientific discovery is rarely as rigid, linear, and empirical as implied in the preceding paragraphs. Molecular biology is certainly replete with examples of the purposeful application of the scientific method, but is also full of surprises, excitement, and occasional messiness. Major discoveries are often characterized by hard work, extraordinary perseverance (a surprising number of discoveries seem to be consummated in the middle of the night), and more than a little innovative thinking. Many such examples are chronicled throughout this book.

To understand how science is done, we need to expand the concept of the scientific method to include the contribution of scientific context. Scientific inquiries are not carried out in a vacuum. The discovery of DNA polymerase, described above, was much more than “hypothesize the existence of this enzyme and then go get it.” The search was carried out in the context of the ideas and facts that prevailed at the time. These included information about the structure of DNA, the chemistry and thermodynamics of enzyme catalysis, cellular physiology and metabolism, the properties of proteins, and the recent finding of a related enzyme, ribonucleotide phosphorylase, by Severo Ochoa. The entire discovery process was guided by the interconnected web of theories and information within which the DNA polymerase hypothesis was developed and tested.

To understand the scientific method, it is also instructive to examine other examples of how important ideas were conceived and tested. The examples below are meant to demonstrate the variety of ways in which science is advanced, and they underscore the fact that science is a very human experience, despite the rigor of the discipline.

Hypothesis and Discovery This classical path to scientific discovery led Kornberg and colleagues to DNA polymerase. It was also used by Jack Szostak in his exploration of prebiotic chemistry (see Moment of Discovery at the opening of this chapter), as well as by countless other scientists. Bob Lehman describes the quite similar path to his discovery of DNA ligase, an enzyme that joins segments of DNA (see Chapter 11,

Moment of Discovery). These discoveries started with hypotheses and ideas reasonably based on the knowledge of the time, and the hypotheses were carefully tested. In addition, the discoveries were made within a broader context of constraints imposed by theories related to the geologic history of our planet, basic chemistry, nucleic acid chemistry, structural biology, the properties of enzymes, and many other areas of knowledge.

Model Building and Calculation As described in Chapter 6, Watson and Crick relied on intuition and important experimental clues from other researchers to build a DNA model structure so clearly fitting the data that its acceptance was at once shocking and immediate. A similar approach can be seen in the work of Steve Mayo in his efforts to deduce the three-dimensional structure of proteins from their amino acid sequence (see Chapter 4, Moment of Discovery). In this variant of the scientific method, a researcher draws on experiments performed by others and mines the unrealized implications of that existing body of knowledge to bring about a significant advance in understanding. Insights obtained in this way are almost always subjected to further experimental examination to confirm them. In both of the cases described here, later solution of the three-dimensional structures of the macromolecules in question confirmed the results of the original investigations.

Hypothesis and Deduction Soon after the structure of DNA was determined, it became apparent that the information in DNA was converted into RNA, and then into protein. However, how amino acids could bind to RNA molecules to guide their assembly into proteins remained unclear. Then, in 1955, Crick wrote an influential note to some colleagues that laid out his “adaptor hypothesis,” proposing that there was an as yet undiscovered molecule that linked each amino acid to sequence information in the RNA. This insight, never published but widely disseminated, was not an endpoint. The hypothesis was subjected to experimental confirmation, and the adaptor-transfer RNA (tRNA)—was discovered within a year by Paul Zamecnik and Mahlon Hoagland (see Chapter 17, How We Know).

Exploration and Observation Darwin’s voyage on the *Beagle* was not designed to answer the questions that his work eventually addressed, and no hypotheses were offered when the ship set sail. Thousands of individual observations (and context provided by the work of many predecessors and contemporaries) simply gelled

in a great mind to create a transformational theory. Molecular biologists routinely embark on similar voyages, with computers and new technologies as vessels of discovery. Rather than focusing on one or a few enzymes and reactions, we can now approach issues related to entire systems. Cells, organisms, and even ecosystems are being explored. Many of these technologies are described in Chapter 8. The new technologies enable us to explore broadly, asking: what’s out there that we have missed? These efforts provide a rich context for new ideas and hypotheses.

Innovation The polymerase chain reaction, often abbreviated PCR, is a convenient process for amplifying a DNA sample. PCR is now so integral to biotechnology and forensic science that it is hard to imagine doing molecular biology without it. PCR was invented by Kary Mullis in a flash of inspiration that came to him while driving along a northern California highway in 1983. Reflecting on one problem, Mullis stumbled on a solution to an even greater one: a method to produce almost unlimited copies of a DNA sequence in a test tube. (This story is presented in more detail in Chapter 7, How We Know.) Although this advance was not precipitated by orderly application of the conventional “question, hypothesis, experiment” approach, the inspiration was a logical and probably inevitable outgrowth of the growing body of knowledge that constituted the molecular biology of that moment.

Serendipity Accidental discoveries are, by definition, unplanned and unexpected, and they can change the way we think about living systems. Such discoveries are not preceded by a carefully orchestrated quest organized around a posed question. They just happen—and they happen often. One example involves the discovery of RNA catalysis. In the early 1980s, Thomas Cech was investigating the mechanism by which segments of RNA (introns) are removed from a messenger RNA in a reaction known as RNA splicing (see Chapter 16). Searching for the protein enzymes that catalyzed the process, Cech and his coworkers were startled when the splicing reaction still occurred in a necessary control trial, in which the protein extract was left out of the reaction mixture. After long and careful experimentation to eliminate any possible source of protein contamination, they were able to report the paradigm-shifting discovery of an RNA-catalyzed reaction.

These many pathways vary in detail, but they have characteristics in common that identify them as part of a modern and adaptable scientific method.

They all focus exclusively on questions and properties related to the natural universe. All the ideas, insights, and experimental facts that arise from these endeavors rely on reproducible observation and/or experiment. Ultimately, the information from the various approaches fits into the web of scientific knowledge. The experimental efforts and results can be used by other scientists anywhere in the world to build new hypotheses and make new discoveries. An attempt to construct a flow chart for a realistic application of the scientific method is shown in [Figure 1-14](#).

Scientists Work within a Community of Scholars

Any individual who uses the scientific method to answer questions about the natural universe is a scientist—from a high school student examining pond water with a microscope to an investigator using high-tech equipment at a major research institute. A specific academic degree is not a prerequisite for making a scientific contribution. However, success in science is strongly correlated with training and experience.

Scientific training is much more than lessons about experimental methods; it is an education in how to think about and approach scientific problems. Progress often demands a willingness to give up a favorite hypothesis. In his book *Advice to a Young Scientist* (1979), the Nobel Prize winner Peter Medawar expressed it thus: “I cannot give any scientist of any age better advice than this: the intensity of the conviction that a hypothesis is true has no bearing on whether it is true or not.”

If any idea about the natural universe does not align with reproducible observations, a scientist must challenge it. Except for the postulate of objectivity, no assumption is inviolate. The ideas that a scientist accepts must be based on evidence that can be observed, measured, and reproduced. The standard of reproducibility, coupled with the continual advance of new experimentation, provides a robust system for weeding out the inevitable errors and guiding insightful reinterpretations.

The many decisions along any path to discovery, such as determining the number and types of tests to use in examining a hypothesis, are guided by training and experience. There is no prescribed formula for these decisions, but they are much less subjective than they might at first appear. The context of scientific theory and information plays a major role in ensuring rigor. When a DNA-synthesizing enzyme was first purified from bacterial crude extracts by Kornberg and his colleagues, specific tests had to be performed to

establish that the enzyme was involved in normal chromosomal replication. These tests were dictated by the broader scientific context in which the discovery was made. The experiments were, in part, conceived by the need to fit this new enzyme into the existing web of scientific information.

When a result contradicts the original hypothesis, a scientist must decide whether to reject or revise. This decision is similarly rooted in both the context of scientific information known at the time and the experience of the investigator. When Bob Lehman and colleagues detected the action of DNA ligase (in the bacterium *Escherichia coli*) for the first time, they anticipated that a high-energy source, probably ATP, would function as a necessary cofactor for the reaction. This expectation was rooted in a great deal of precedent from similar reactions. However, when ATP failed to stimulate the ligation reaction, the DNA ligase hypothesis was not abandoned. Instead, the investigators questioned their assumptions about ATP. A search eventually identified another cofactor, NAD⁺, as the critical energy source for the *E. coli* DNA ligase, revealing a new chemical function for this important cellular molecule.

In science, the work is not done by individuals in isolation, and scientific decisions are influenced and affected by the scientific community at large. It is not enough to demonstrate a new idea to oneself. To integrate an advance into the web of scientific information, a scientist must make a case that is compelling to the broader community. A continual discussion within this community fosters progress and helps ensure the integrity of the information generated. When an idea is tested, the most compelling case is made when the idea passes multiple tests involving two or more different techniques. Collaboration among scientists often strengthens the case. Different individuals bring different perspectives and expertise to bear on a problem. The cross-talk can correct errors, provide novel insights, and make connections that one individual might miss. The case becomes even more compelling when tests are replicated in multiple laboratories. To gain entry to the scientific literature, the work must pass the inspection of fellow scientists. Peer reviews of scientific papers, and of grant applications, are sometimes challenging and even painful experiences for the authors, but they are critical for ensuring quality and rigor. A conscientious participation in this peer-review system is a shared duty of every working scientist.

Ultimately, this literature documents the scientific story about the natural universe. That story is ever expanding, never complete, and always ready for new contributors.

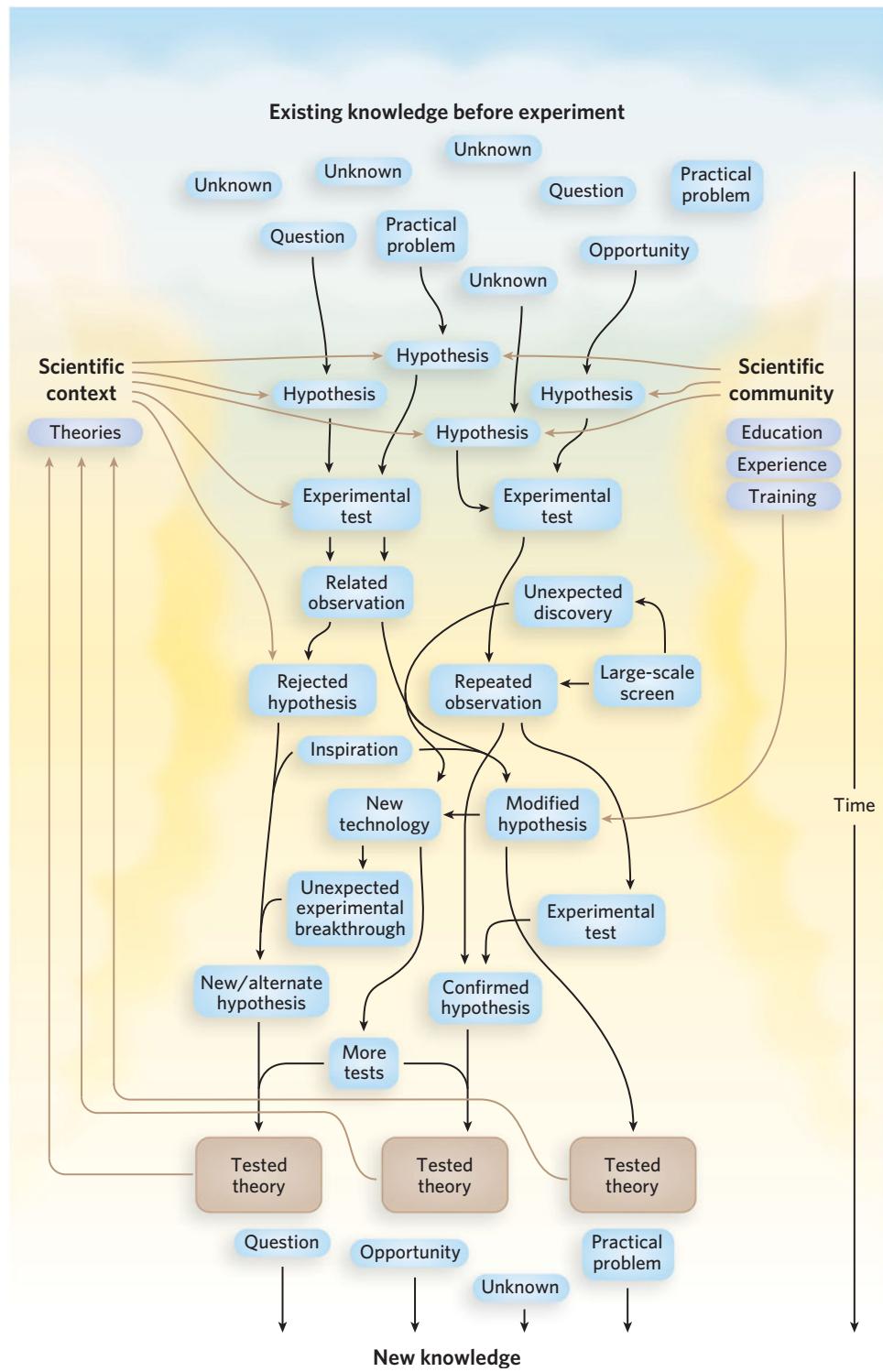


FIGURE 1-14 A flow chart of the scientific method. When scientists apply the scientific method, it is a nonlinear process with many and varied inputs.

SECTION 1.2 SUMMARY

- Science is both a body of knowledge and a process of inquiry by a community of scholars. Science is a path to knowledge about the natural universe, relying on reproducible observation and experiment. Science relies on one foundational assumption: the postulate of objectivity, that the universe is governed by immutable, and therefore discoverable, laws.
- The classic form of the modern scientific method poses questions, generates explanatory hypotheses, and tests the hypotheses with experiments. Hypotheses are accepted, rejected, or modified depending on the results of the experiments.
- The scientific method is malleable, and there are multiple paths to scientific knowledge. All pathways are linked by a reliance on reproducible observation or experiment and by their capacity to generate knowledge consistent with the broader context of scientific information.
- Scientists are individuals who apply the scientific method to answer questions about the natural universe. Their decisions are guided by training, their knowledge of the scientific context of their work, and continuing interactions with the worldwide scientific community.

Unanswered Questions

The questions that are relevant to this introductory chapter include some of the most fundamental and far-reaching issues in molecular biology. Many of them relate to discussions in chapters throughout the book.

- What is the minimal set of genes required in a living cell?** Efforts to define a cell at its most basic level are well underway. Scientists may create an engineered cell in the near future. The research has practical implications, because such a cell could be a living scaffold for the engineering of cells to bioremediate toxic waste or generate biofuels. The work should also tell us something about LUCA.
- Can we reconstruct the entire tree of life?** Developing a complete tree of life is an ongoing effort of evolutionary biology, making use of the tools of molecular biology and complementary information from the fossil and geologic records (see Chapter 8).
- What is the full set of mechanisms that drive evolution, and how much does each mechanism contribute?** Genetic variation in a species results from spontaneous DNA damage, replication errors, transposition, DNA repair processes gone awry, genetic recombination, and many other processes. Molecular biology provides the roadmap to a comprehensive listing and understanding.
- Can evolution be controlled?** When bacteria are subjected to stress, specific sets of genes turn on in programmed responses. One part of that response in *E. coli* involves the production of higher levels of mutagenesis. Can the development of resistance of bacterial pathogens to antibiotics—an increasing problem in medical treatments—be slowed by inhibiting bacterial stress responses to DNA damage? Some academic and biotechnology industry researchers around the world are betting that it can.

How We Know

Adenine Could Be Synthesized with Prebiotic Chemistry

Ferris, J.P., and L.E. Orgel. 1966. An unusual photochemical rearrangement in the synthesis of adenine from hydrogen cyanide. *J. Am. Chem. Soc.* 88:1074.

Oro, J., and A.P. Kimball. 1962. Synthesis of purines under possible primitive earth conditions. II. Purine intermediates from hydrogen cyanide. *Arch. Biochem. Biophys.* 96:293–313.

Few biomolecules are more important than adenine. It is the base component not only of the adenine nucleotides in RNA and DNA, but also of a wide range of enzyme cofactors, including ATP (adenosine triphosphate), an important energy source that drives many cellular reactions. Thus, over the course of evolution, adenine took on a larger role than other nucleotide bases in the biochemistry of living systems, presumably reflecting some particular aspect of the evolution of life. Experiments carried out by Juan Oro in the 1950s and early 1960s, and continued by Leslie Orgel and colleagues, demonstrated that there are good chemical reasons to think that adenine may have been present in much higher concentrations than other nucleotide bases in the prebiotic soup. This suggests that adenine's current place in the scheme of life

reflects the conditions when life was formed: adenine was there!

One of the most common reagents generated from the molecules thought to predominate in the atmosphere of the early Earth is hydrogen cyanide (HCN). HCN can be converted into a variety of organic molecules, one of the most abundant products being adenine. Adenine can be synthesized from HCN in several ways, but the scheme shown in **Figure 1** is the most plausible under the conditions believed to have existed on the early Earth. An HCN tetramer forms in the first step (Figure 1a), and this reacts with formamidine (also formed from HCN) to produce 4-amino-5-cyanoimidazole (AICN) (Figure 1b). Adenine and related molecules are formed from AICN or from its hydrolysis product, 4-aminoimidazole-5-carboxamide (AIC) (Figure 1c).

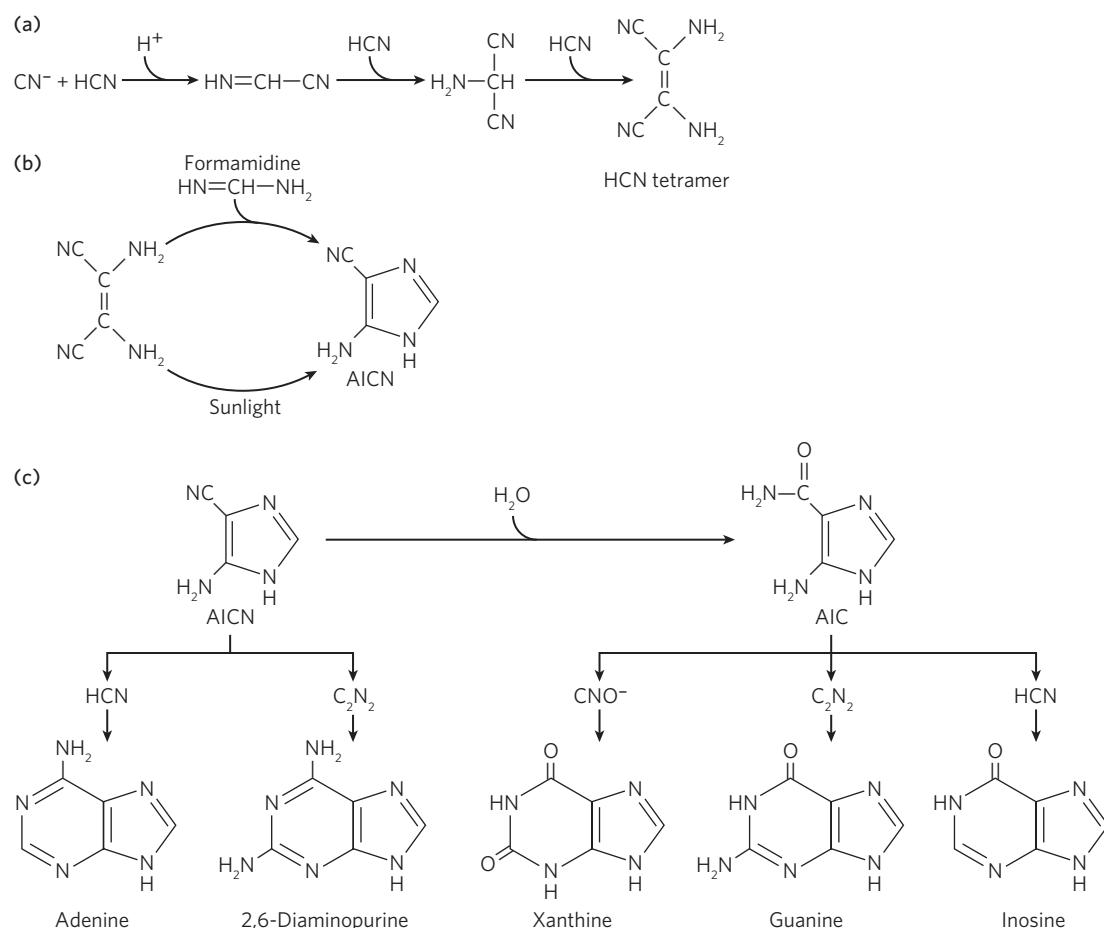


FIGURE 1 HCN can be chemically converted into adenine and related molecules. [Source: Adapted from L. E. Orgel, *Crit. Rev. Biochem. Mol. Biol.* 39:99, 2004.]

Clay Had a Role in Prebiotic Evolution

Hanczyc, M.M., S.M. Fujikawa, and J.W. Szostak. 2003. Experimental models of primitive cellular compartments: Encapsulation, growth, and division. *Science* 302:618–622.

Huang, W.H., and J.P. Ferris. 2003. Synthesis of 35–40 mers of RNA oligomers from unblocked monomers: A simple approach to the RNA world. *Chem. Commun.* 12:1458–1459.

Sodium montmorillonite, first mined at Fort Benton, Wyoming, and commonly known as bentonite, is a type of clay used commercially. It generally contains various mineral impurities (10% to 20%) that give it a layered aluminosilicate structure. The clay readily expands to permit large molecules to enter the interlayers. Products making use of montmorillonite clays include lubricating grease, paints, copy paper, dynamite, plaster, cat litter, matches, cement tiles, shoe polish, concrete, cleaning agents, wall boards, crayons, and bleaching agents. Some forms of montmorillonite are claimed to have health-giving properties and are used in edible preparations marketed by health spas and health food companies. The claims that montmorillonite clays are “living clays” may not be far off the mark.

Any nucleotides formed in the prebiotic soup would have had to be joined into polymers before the properties of RNA could be exploited by evolution. Wenhua Huang and James Ferris demonstrated that a phosphoramidite-activated nucleotide, based on the structure of 1-methyladenine, is readily polymerized in the presence of montmorillonite clay, producing polymers of up to 40 nucleotides (Figure 2). These clays have also been shown to facilitate a variety of other reactions leading to the production of polynucleotides and other complex organic molecules that may have been part of the prebiotic soup and key ingredients in the cauldron that generated the first living organisms.

Jack Szostak’s research group has long been interested in identifying reactions that might explain how early replicating polymers could have become enclosed in lipid envelopes (see this chapter’s Moment of Discovery). Szostak and his coworkers found that montmorillonite clay facilitates the conversion of lipids into vesicles. In their experiments, clay particles often became entrapped in the vesicles, bringing with them any RNA embedded in the particle surfaces. As more lipids were added, the vesicles would grow and bud off, without dilution of their contents. This system models many properties of a primordial cell.

We’ve included here just a small sampling of studies exploring the properties of this unusual clay material. A wide array of reactions are facilitated by this clay, many of them of interest to evolutionary biologists.

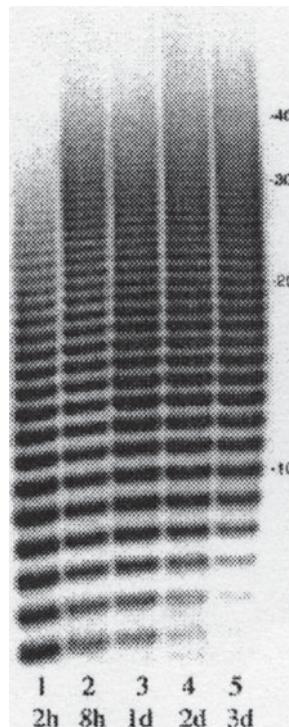


FIGURE 2 The polymerization of 1-methyladenine nucleotides is shown in this gel electrophoresis experiment by the increasing size of DNA fragments with time. The numbers on the right indicate the length of the polymer in nucleotides; the labels below the lane numbers indicate duration of polymerization in hours (h) and days (d). [Source: W. H. Huang and J. P. Ferris, *Chem. Commun.* 12:1458–1459, 2003, Fig. 2. Photo reproduced by permission of The Royal Society of Chemistry/Courtesy of James Ferris.]

Darwin's World Helped Him Connect the Dots

Mayr, E. 2000. Darwin's influence on modern thought. *Sci. Am.* 283(July):78–83.

Padian, K. 2008. Darwin's enduring legacy. *Nature* 451:632–634.

Darwin was not the first to come up with the idea of evolution. The sense that the world changes over time, and that the organisms inhabiting it also change, can be traced back to classical Greece. By the end of the eighteenth century, the study of geology had impressed on the scientists of the day that Earth is much older than previously thought and that the environment changes over time. The discovery of and growing interest in fossils had similarly shown that there were animals on Earth in past eras that were no longer present. They had either become extinct or undergone great changes in appearance.

One of the first influential efforts to explain all these observations came from the French biologist Jean-Baptiste Lamarck (1744–1829). Lamarck was the first scientist to use the term *biology*, coined in 1802 to encompass the sciences of botany and zoology. Lamarckian evolution is linear, not branched. He also proposed that changes in the environment triggered adaptive changes in individuals that could be passed on to succeeding generations, an idea known as the inheritance of acquired characteristics.

Lamarck's ideas were hotly debated for many decades, and a variety of competing ideas about the transmutation or transformation of animals and plants appeared in contemporary writings. The basic idea that organisms change, or evolve, over time was familiar, if not well accepted, before Darwin presented his

theory of evolution in *The Origin of Species*. However, until he set pen to paper, no plausible mechanism of evolution had been proposed.

Drawing on his experiences during the 1831–1836 voyage of H.M.S. *Beagle* (Figure 3), Darwin formulated his own new ideas over a period of years. Observations he had accumulated during a five-week stay in the Galápagos Islands had a particularly strong influence. The species on the islands differed from those on the mainland, consisting only of representatives of groups that could have managed a trip over a large expanse of ocean. Patterns of divergence were evident in certain groups, such as the multiple finch species on the islands. Darwin's observations indicated that a few individuals of one finch species had originally colonized the islands. That species had then diversified to exploit the various island environments.

Darwin was influenced by the works of Thomas Malthus (1766–1834), who had argued that human reproduction, if not controlled, would lead to a population that would outstrip the food supply. This would lead to an inevitable struggle to survive.

Darwin's great work was published in 1859, after he discovered that similar ideas had been developed by Alfred Russel Wallace (1823–1913). Wallace deserves credit as a codiscoverer of natural selection. However, Darwin conceived the idea earlier and developed it more fully.

FIGURE 3 The route taken and the lands visited during the 1831–1836 voyage of H.M.S. *Beagle*. [Source: Charles Darwin, *The Voyage of the Beagle*, 1839.]



Key Terms

molecular biology, p. 2	catalytic RNA, p. 6	eukaryotes, p. 9
biological information, p. 2	ribozyme, p. 6	natural selection, p. 9
catalysis, p. 2	LUCA (last universal common ancestor), p. 8	horizontal gene transfer, p. 11
evolution, p. 2	chemoheterotroph, p. 9	postulate of objectivity, p. 12
mutation, p. 2	bacteria, p. 9	hypothesis, p. 12
enzyme, p. 3	archaea, p. 9	scientific theory, p. 13
RNA world hypothesis, p. 6		scientific method, p. 14

Additional Reading

General

- Carroll, S.B.** 2005. *Endless Forms Most Beautiful: The New Science of Evo Devo and the Making of the Animal Kingdom*. New York: W. W. Norton & Company. An accessible exploration of evolutionary developmental biology by one of the leaders in the field.
- Wilson, E.O., ed.** 2005. *From So Simple a Beginning: Darwin's Four Great Books*. New York: W. W. Norton & Company. A one-volume compendium of the four great works of Charles Darwin: *Voyage of the H.M.S. Beagle* (1845), *The Origin of Species* (1859), *The Descent of Man* (1871), and *The Expression of Emotions in Man and Animals* (1872).

The Evolution of Life on Earth

- Dawkins, R.** 2004. *The Ancestor's Tale: A Pilgrimage to the Dawn of Evolution*. New York: Houghton Mifflin Company.
- Gould, S.J.** 2002. *The Structure of Evolutionary Theory*. Boston: Harvard University Press. The last contribution from one of the great popularizers of evolutionary biology.
- Hanczyc, M.M., and J.W. Szostak.** 2004. Replicating vesicles as models of primitive cell growth and division. *Curr. Opin. Chem. Biol.* 8:660–664. A summary of some current efforts to re-create a protocell in the laboratory.
- Harris, D.R., S.V. Pollock, E.A. Wood, R.J. Goiffon, A.J. Klingele, E.L. Cabot, W. Schackwitz, J. Martin, J. Eggington, T.J. Durfee, et al.** 2009. Directed evolution of radiation resistance in *Escherichia coli*. *J. Bacteriol.* 191:5240–5252. Bacteria can evolve very quickly when the environment demands it.
- Kirschner, M.W., and J.C. Gerhart.** 2005. *The Plausibility of Life: Resolving Darwin's Dilemma*. New Haven, CT: Yale University Press.
- Lazcano, A., and S.L. Miller.** 1996. The origin and early evolution of life: Prebiotic chemistry, the pre-RNA world, and time. *Cell* 85:793–798.
- Lincoln, T.A., and G.F. Joyce.** 2009. Self-sustained replication of an RNA enzyme. *Science* 323:1229–1232. Yes, RNA molecules are capable of self-replication.
- Mayr, E.** 2000. Darwin's influence on modern thought. *Sci. Am.* 283(July):78–83. A terrific explanation of how

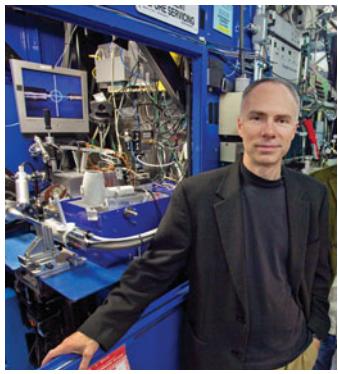
Darwin has affected the way in which all of us approach our world.

- Mayr, E.** 2001. *What Evolution Is*. New York: Basic Books.
- Orgel, L.E.** 2004. Prebiotic chemistry and the origin of the RNA world. *Crit. Rev. Biochem. Mol. Biol.* 39:99–123. An excellent summary of the field exploring prebiotic chemistry.
- Trevors, J.T.** 2003. Early assembly of cellular life. *Prog. Biophys. Mol. Biol.* 81:201–217.
- Woese, C.R.** 2004. A new biology for a new century. *Microbiol. Mol. Biol. Rev.* 68:173–186. A thoughtful appeal for some new approaches in biology.
- Zimmer, C.** 2001. *Evolution: The Triumph of an Idea*. New York: HarperCollins Books. The companion volume to the PBS (WGBH) series on evolution.
- Zuckerkandl, E., and L. Pauling.** 1965. Molecules as documents of evolutionary history. *J. Theor. Biol.* 8:357–366. One of the first articles to recognize that much of evolutionary history is documented in the genomes of existing organisms.

How Scientists Do Science

- Bryson, B.** 2004. *A Short History of Nearly Everything*. New York: Broadway Books. Easy to read, this book provides an excellent lay description of how scientists approach problems.
- Dennett, D.C.** 1995. *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. New York: Touchstone.
- Feynman, R.** 1999. *The Pleasure of Finding Things Out: The Best Short Works of Richard Feynman*. New York: Helix Books/Perseus Books.
- Kosso, P.** 2009. The large-scale structure of scientific method. *Sci. Educ.* 18:33–42.
- Medawar, P.** 1979. *Advice to a Young Scientist*. New York: HarperCollins.
- Sober, E.** 2008. *Evidence and Evolution: The Logic behind the Science*. Cambridge: Cambridge University Press.
- Working Group on Teaching Evolution, National Academy of Sciences.** 1998. *Teaching about Evolution and the Nature of Science*. Washington, DC: National Academy Press. www.nap.edu. This excellent resource is available as a free download.

DNA: The Repository of Biological Information



James Berger [Source: Courtesy of James Berger]

Moment of Discovery

The first time I had an “Aha!” moment in science was when I was a graduate student. The question that intrigued me was related to the mechanism proposed for topoisomerases, which are essential enzymes that coil or uncoil DNA during DNA synthesis in all cells. Topoisomerase II-type enzymes (called Topo II) pass DNA strands through each other by cutting and rejoicing DNA without marking or changing the genome in any way. In textbooks, the enzyme was shown as a sphere that bound to one seg-

ment of DNA, cut it and then split in half to pass a second DNA segment through the split. But what held the DNA ends together during the passage of the DNA duplex through the double-stranded break? There had to be something else going on.

Francis Crick once said that you can’t understand how an enzyme works unless you see its structure, and *I wanted to see the structure of Topo II*. I spent the next couple of years trying to crystallize the enzyme with no success, and eventually reached the point where I wondered if my project would ever work, and whether I had what it took be a scientist. I made one last preparation of the enzyme, and after working overnight in the lab, I put the purified enzyme on ice and went home to bed. When I came back the next day, the protein in the tube had turned white, and I was crushed, thinking it had precipitated into a useless aggregate. But when I looked at a sample under the microscope, I saw crystals growing in the tube! At that moment I knew I had a project. I spent the next nine months solving the molecular structure of the enzyme, and I’ll never forget the thrill of seeing the structure for the first time.

It was instantly clear how Topo II must work. The enzyme has two jaws, one of which grabs and cleaves the DNA duplex and holds it while the other jaw passes a different segment of DNA through the gap. I experienced the intense joy of discovering this fundamental mechanism of DNA metabolism, and of knowing that at that moment I was the first person in the world to have this understanding of the natural world.

—James Berger, on his discovery of the structure and mechanism of topoisomerase II

2.1 Mendelian Genetics	25
2.2 Cytogenetics: Chromosome Movements during Mitosis and Meiosis	31
2.3 The Chromosome Theory of Inheritance	38
2.4 Molecular Genetics	44

Genetics is the science of heredity and the variation of inherited characteristics. Today, we know that biological information is stored and transmitted from generation to generation by deoxyribonucleic acid, or DNA, but this understanding arose only gradually. DNA was not widely accepted as the chemical of heredity until the 1940s, and its structure was not determined until 1953, when James Watson and Francis Crick introduced the world to the DNA double helix. (The structure of DNA is described in Chapter 6.) Our knowledge of the beautiful double-helical DNA structure has transformed the way that science is performed, to the extent that it is tempting to think of the field of genetics in terms of pre- and post-DNA structure. But genetics has a wonderfully rich and varied history, every bit as exciting in the decades before the double helix as afterward.

The beginnings of modern genetics can be traced to the 1850s, when Gregor Mendel studied the inheritance of traits in the garden pea. He deduced that organisms contain particles of heredity (what we now call genes) that exist in pairs and that the paired particles split up when **gamete cells** (sex cells, the ovum and pollen in peas) are formed; pairs of hereditary particles are reformed on the union of two gametes during fertilization. Mendel was absolutely correct, but decades ahead of his time. His marvelous work went unnoticed for more than 30 years, until well after his death.

In contrast, a contemporary of Mendel's, Charles Darwin, was exceedingly famous in his lifetime. Darwin's theory of evolution started an awakening, one that continues to this day (see Chapter 1). For evolutionary theory to work, there must be diversity among individuals within a species, and variants more suited to the environment are selected and survive to produce offspring. Darwin's evolutionary theory, as wonderful as it is, completely lacks an explanation for how this diversity is produced. In fact, Darwin spent considerable time pondering this problem. He espoused the theory of pangenesis, first proposed by the ancient Greeks, in which genetic traits are shaped by life experience and transferred by "pangenes" to gamete cells, via the blood, enabling the traits to be inherited. In principle, the mistaken pangenesis theory is a variation of Jean-Baptiste Lamarck's theory of inheritance of acquired characteristics (see Chapter 1, How We Know).

Darwin's theory of evolution became widely known, but Mendel's work fell into obscurity. During the late 1800s, advances in microscopy pushed the optical limits, enabling scientists to visualize subcellular structures. Of particular interest to geneticists were chromosomes, structures found in the nuclei of cells. A rash of intense studies documented chromosome behavior during cell

division, fertilization, and the formation of gamete cells. New discoveries revealed that the number of chromosomes in **somatic cells** (all cells in a multicellular organism other than sex cells) is constant for a given species and that the chromosome number is halved to form gametes. When Mendel's work was rediscovered in 1900, his principles of heredity and particles of inheritance fit nicely with the behavior of chromosomes observed under the microscope.

Proof that genes reside on chromosomes soon followed, from a series of wonderful studies on fruit flies started in 1908 by Thomas Hunt Morgan. Central to Morgan's work were mutants, flies displaying atypical physical traits not found in the average fly. The variety of mutant flies accumulated by Morgan's lab during 15 years of study—generations of flies reared in milk bottles—was amazing, including flies with bodies of different shapes and sizes, a variety of wing patterns, legs of different sizes, and a whole spectrum of eye colors. These fly mutants simply appeared spontaneously over generations of growth in Morgan's lab. Here was the answer to the variation required to make Darwin's theory of evolution work. Spontaneous mutants are infrequent, but given the expanse of evolutionary time, sufficient numbers and types of mutants are produced for nature to select and mold new species.

Genes and mutations explain heredity and illuminate evolutionary theory. But what are genes made of, and how is the information within them translated into the physical traits of an organism? Chromosomes were known to consist of both DNA and protein—but which of these is the genetic material? Several elegant and now classic experiments identified DNA as the molecule of heredity and found that DNA contains a code to direct the synthesis of RNA and proteins. The structure of the DNA double helix intuited by Watson and Crick revealed an architecture more beautiful than anyone could have imagined. The DNA molecule consists of two long strands twisted about each other, each chain a series of repeating units called nucleotides.

Watson and Crick immediately realized that the cell must have a mechanism to untwist the two strands in order to duplicate the DNA molecule and pass the genetic information to the next generation. Indeed, as we shall see in Chapter 9, the cell contains a complex arsenal of enzymes devoted to untwisting and altering the topology of DNA. These enzymes, called topoisomerases, are important targets of anticancer drugs, and their mechanisms of action are still actively investigated by James Berger and many other researchers. By extension, it is also critical to have a thorough understanding of the DNA itself, and that is the subject of this and several other chapters in this textbook.

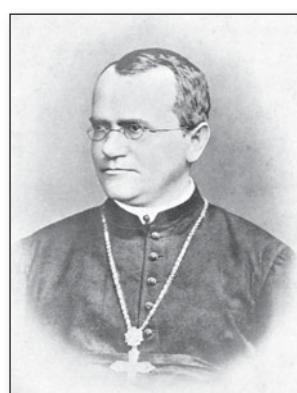
Despite the intrinsic beauty of the newly discovered double helix, the process by which a sequence of nucleotides could code for a sequence of amino acids in a protein remained a mystery. In a rapid series of advances over just 10 years following Watson and Crick's breakthrough, mRNA, tRNA, and rRNA were discovered and the workings of the directional flow of biological information, DNA→RNA→protein, were understood.

The field of molecular biology developed from these great discoveries. It seeks a detailed explanation of how biological information brings order to living processes. Molecular biology—the subject of this book—is a rapidly evolving scientific pursuit. The fundamental discoveries described in this chapter provided the groundwork for all subsequent studies in the field.

2.1 Mendelian Genetics

Gregor Mendel was a monk at the monastery of St. Thomas at Brünn (now Brno in the Czech Republic). Renowned for teaching the sciences, the monastery sent Mendel to the University of Vienna in 1851, to obtain his teaching credentials. Although Mendel failed his final exams, he returned to the monastery and began a 10-year program of experiments that were so well conceived and executed that his results form the cornerstone of modern genetics.

In Mendel's time, plant hybrids were highly desired for their unique ornamental varieties. But the inheritance of colorful hybrid flower patterns was perplexing and unpredictable. The inability to decipher general principles of inheritance was not for lack of trying. Many well-known scientists performed extensive plant-breeding experiments, but no fundamental principles of inheritance could be formulated from these endeavors.



Gregor Mendel,
1822-1884

Mendel's success where others failed can be attributed to his sound scientific approach to the problem. For his studies Mendel picked the garden pea, *Pisum sativum*, an excellent choice for several reasons. Because the pea was economically important, many varieties were available from seed merchants. The pea plant is also small, so that many plants can be grown in a confined space;

and it grows quickly, reaching maturity in one growing season. But perhaps more important than Mendel's choice of experimental organism was his approach to studying it. Others before him, in studying plant breeding, had looked at the plant as a whole, dooming a study of heredity from the start, because many genes control the overall appearance of an organism. Mendel focused instead on separate features of the plant, carefully observing isolated characteristics of the seeds, flowers, stem, and seed pods (Figure 2-1).

Mendel's First Law: Allele Pairs Segregate during Gamete Formation

Mendel spent the first two years growing different varieties of peas to ensure that they were true-breeding, or **purebred**. Then he carefully selected seven different pairs of traits and cross-pollinated plants with contrasting traits. For example, plants with round seeds were crossed with plants having wrinkled seeds. The parental plants are referred to as the **P generation** (*P* for parental). The **hybrid** offspring are called the **F₁ generation** (*F* for filial, from the Latin for “daughter”; *F₁* indicating first filial). This first generation produced only round seeds; the wrinkled-seed trait seemed to have disappeared. Mendel observed a similar result in crosses for all seven pairs of traits (Table 2-1). He referred to the trait that appears in the F₁ generation as the **dominant** trait, and the trait that disappears as the **recessive** trait (see Figure 2-1).

Mendel's finding that one trait is dominant and the other is recessive was novel and completely contrary to the prevailing view that parental traits blend together in the offspring. Other experimenters might have stopped at this new and dramatic discovery, but not Mendel. In his next experiment, he allowed F₁ plants to self-pollinate and produce the **F₂ generation** (second filial generation). Surprisingly, the F₂ generation was a mixture: most plants produced round seeds (the dominant trait), but some had wrinkled seeds. The recessive wrinkled-seed trait that disappeared in the F₁ generation reappeared in the F₂ generation!

Unlike other scientists before him, Mendel kept close track of the numbers of dominant and recessive offspring. He counted 5,474 dominant round seeds and 1,850 recessive wrinkled seeds in the F₂ generation, for a ratio of 2.96 dominant to 1 recessive trait. His experiments examining seed color produced a similar result. He observed 6,022 dominant yellow seeds and 2,001 recessive green seeds, for a ratio of 3.01:1. The other pairs of traits also displayed a 3:1 ratio of dominant to recessive offspring in the F₂ generation, as summarized in Table 2-1.

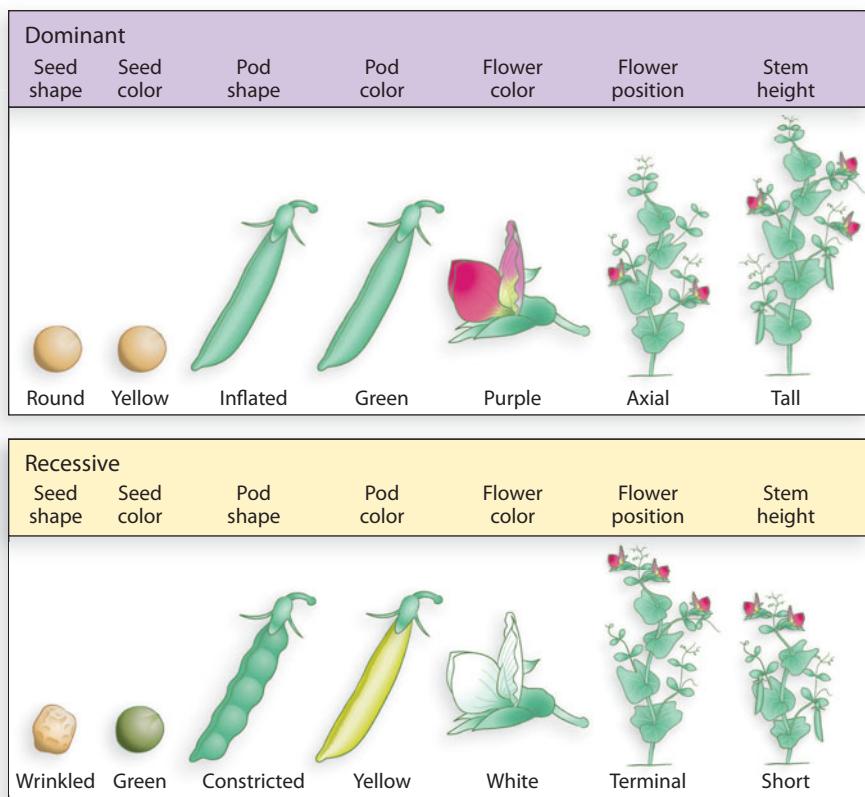


FIGURE 2-1 Traits of the garden pea that were examined by Mendel.

Mendel picked seven pairs of traits to study: seed shape, seed color, seed pod shape, seed pod color, flower color, flower position along the stem, and stem height. The dominant and recessive forms of the traits are shown.

Mendel's interpretation of these results was brilliant. Reappearance of the recessive wrinkled-seed trait in the F_2 generation suggested that traits do not really disappear. Mendel therefore proposed that traits are "hereditary particles"—now called **genes**—and that they come in pairs. Organisms that carry two copies of each gene are **diploid**, and the different variants of a given gene are called **alleles**. In other words, a diploid parental plant has two alleles of the seed-shape gene. Furthermore, Mendel proposed that one allele could mask the appearance

of the other. This explained why traits could disappear but then reappear in future generations. Even though the F_1 plant carried one dominant allele (round seed) and one recessive allele (wrinkled seed), only the dominant round-seed allele is evident in the outward appearance, or **phenotype**, of the F_1 plant. Mendel also reasoned that each parent contributes only one copy of each gene to the offspring; that is, the gamete cells are **haploid**, having only one allele of each gene. When two F_1 generation gametes combine and each carries the recessive allele for

Table 2-1 Results of Mendel's Single-Factor Crosses

Characteristic Isolated for Study	Parental Cross	F_1 Phenotype (dominant trait)	F_2 Phenotypes (dominant and recessive)	F_2 Ratio (dominant: recessive)
Seed shape	Round × wrinkled	All round	5,474 round and 1,850 wrinkled	2.96:1
Seed color	Yellow × green	All yellow	6,022 yellow and 2,001 green	3.01:1
Pod shape	Inflated × constricted	All inflated	882 inflated and 299 constricted	2.95:1
Pod color	Green × yellow	All green	428 green and 152 yellow	2.82:1
Flower color	Purple × white	All purple	705 purple and 224 white	3.15:1
Flower position	Axial × terminal	All axial	651 axial and 207 terminal	3.14:1
Stem height	Tall × short	All tall	787 tall and 277 short	2.84:1

Table 2-2 Commonly Used Terms in Genetics

Term	Definition
P generation	Parents used in a cross
F ₁ generation	Progeny resulting from cross in the P generation
F ₂ generation	Progeny resulting from cross in the F ₁ generation (succeeding generations are F ₃ , F ₄ , etc.)
Purebred	Individual homozygous for a given trait or set of traits
Hybrid	Progeny resulting from a cross of parents with different genotypes
Gene	Section of DNA encoding a protein or functional RNA
Allele	Variant of the gene encoding a trait (e.g., yellow or green seeds)
Phenotype	Outward appearance of an organism
Genotype	Alleles contained in an organism
Homozygous	Having two identical copies of an allele for one gene
Heterozygous	Having two different alleles for one gene
Dominant allele	Allele expressed in the phenotype of a heterozygous organism
Recessive allele	Allele masked in the phenotype of a heterozygous organism

seed shape, the resulting F₂ plant will produce wrinkled seeds.

We have introduced several genetic terms in the preceding paragraphs. These terms, and others that follow, part of the scientific language of genetics, are defined in Table 2-2.

Mendel was well-trained in mathematics, which helped him make sense of his results. To explain the ratios of phenotypes in the offspring in mathematical terms, he referred to the dominant allele with a capital letter (e.g., R for round) and to the recessive allele with a lowercase version of the same letter (r for wrinkled). A lowercase *w* might seem more fitting for the wrinkled-seed allele, but using different letters would make it harder to keep track of allele pairs of the same gene. A purebred round-seed plant has two R alleles, RR, and a purebred wrinkled-seed plant has two recessive alleles, rr. RR plants exhibit the dominant R trait (round seeds), and rr plants exhibit the recessive r trait (wrinkled seeds). This double-letter nomenclature, representing the allelic makeup of an organism, is a way of denoting the organism's **genotype**.

In a cross of purebred parents, round (RR) and wrinkled (rr), all F₁ progeny receive one allele from each parent and are thus Rr. The R allele is dominant to the r allele, so all F₁ Rr hybrid plants have round seeds. The F₁ plant produces R and r gametes in equal amounts, and therefore self-pollination of F₁ plants produces three different diploid genotypes: RR, Rr, and rr. A modern-day way of displaying the genes that come together during a cross such as this is a **Punnett square** analysis (Figure 2-2). In this analysis, the gamete genotypes are written along the top and side of the Punnett

square, and the various combinations in which the alleles can come together during pollination are entered in the grid. The results yield the three different F₂ genotypes in the following ratios: 1 RR, 2 Rr, and 1 rr. These genotypes, together with the concept of dominant and recessive alleles, explain the ratio of phenotypes that Mendel observed: 3 dominant (1 RR + 2 Rr) to 1 recessive (1 rr). We now refer to an organism with identical alleles for a given gene as **homozygous** for that gene (such as RR or rr). An organism with two different alleles, such as an Rr plant, is characterized as **heterozygous**.

To determine whether the F₂ plants really were of three genotypes in a 1:2:1 ratio, Mendel analyzed the offspring of self-fertilized F₂ plants (the F₃ generation). The F₂ recessive wrinkled-seed plants (25% of the total F₂ plants) all bred true, giving only wrinkled-seed F₃ progeny, and thus were homozygous rr. The dominant round-seed F₂ plants were of two types. One-third (25% of the total) bred true; their offspring always produced round seeds, and therefore these F₂ plants were homozygous RR. The remaining two-thirds of the round-seed F₂ plants (50% of the total) produced F₃ plants with round and wrinkled seeds in a 3:1 ratio, and therefore these F₂ plants were heterozygous Rr. The analysis fit the 1:2:1 ratio for the F₂ genotype exactly: 1RR:2Rr:1rr (see Figure 2-2).

In summary, Mendel hypothesized that traits are carried by particulate genes, that somatic cells contain two copies (two alleles) of each gene, and that gamete cells obtain only one allele for each gene during gamete formation. When two gametes fuse at fertilization, allele pairs are restored, producing the diploid genotype

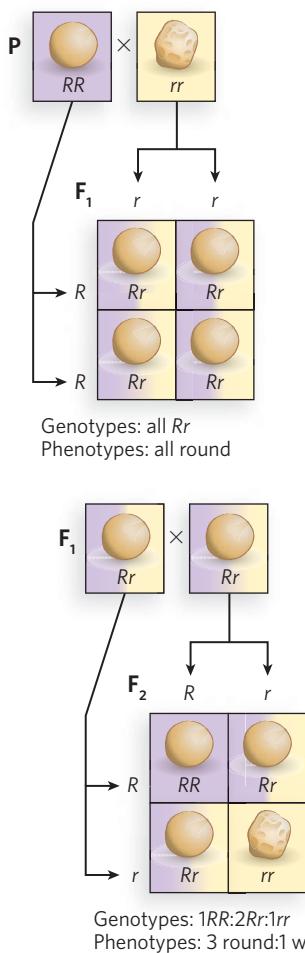


FIGURE 2-2 An example of Mendel's first law. Alleles of the same gene segregate independently into gametes. Plants that are homozygous for dominant round-seed shape (*RR*) were cross-pollinated with homozygous recessive wrinkled-seed plants (*rr*) to produce *F*₁ progeny. The Punnett square analysis shows the gametes of each parent along the top and left side of the grid, from which—if the gametes are formed in equal amounts—one can predict the possible progeny and their frequency. Punnett square analysis is a common way of illustrating genetic crosses today, but was developed after Mendel's work. In the *F*₁ generation, the only progeny that can be produced are *Rr* hybrids. *F*₁ plants were self-fertilized, and as the Punnett square analysis for the *F*₂ generation predicts (based on the assumption that the different alleles (*R* and *r*) segregate independently into *F*₁ generation gametes), round seeds and wrinkled seeds were produced in a 3:1 ratio.

of the offspring. The general principle summarizing this proposal is often referred to as Mendel's first law, or the **law of segregation**, which states that equal and independent segregation of alleles occurs during formation of gamete cells.

Mendel's Second Law: Different Genes Assort Independently during Gamete Formation

Mendel's results demonstrated that two alleles for one gene separate during gamete formation, but how do alleles for two *different* genes behave? There are two possibilities. Alleles for two different genes could separate during gamete formation, assorting randomly into the gametes. Alternatively, they could remain associated, traveling together into the same gamete cells. These two scenarios have distinct outcomes. For example, if particular alleles for seed shape and seed color stay together during the formation of gamete cells, future offspring will retain both the same seed shape and the same seed color as one parent or the other. But if alleles for the two genes separate during gamete formation, some of the *F*₂ offspring will exhibit new combinations of seed shape and color, distinct from those of either parent.

To test these hypotheses, Mendel's next experiments were two-factor crosses, analyzing the transmission of two different genes in each cross. He began by cross-pollinating purebred plants having round, yellow seeds with plants having wrinkled, green seeds. First let's consider the genotype of the two plants. We already know that the round-seed allele (*R*) is dominant to the wrinkled-seed allele (*r*). The genotype of the purebred plant with dominant yellow seeds is *YY*, and the purebred plant with green seeds is homozygous for the recessive green-seed allele, *yy*. Thus, the genotype of a purebred round, yellow-seed plant is *RRYY*, and the genotype of a plant with wrinkled, green seeds is *rryy*. A cross between these plants yields *F*₁ progeny of genotype *RrYy*, phenotypically round and yellow. If the four alleles for the two genes separate and assort randomly during gamete formation, *F*₁ plants will produce four different gametes (*RY*, *rY*, *RY*, and *ry*), and all combinations of seed shape and color will be observed in *F*₂ plants (Figure 2-3).

Mendel's observations of *F*₂ phenotypes are shown in Table 2-3. All possible combinations of traits occurred, and therefore the alleles of the two different genes are not physically connected; instead, they separate during the formation of gamete cells. The analysis of the four possible genotypes, as shown in Figure 2-3, predicts a 9:3:3:1 ratio, close to the observed result. Mendel performed many two-factor crosses analyzing different gene combinations, and the results were always consistent with the random assortment of genes during gamete formation. Mendel's second law, or the **law of independent assortment**, states that different genes assort into gametes independent of one another.

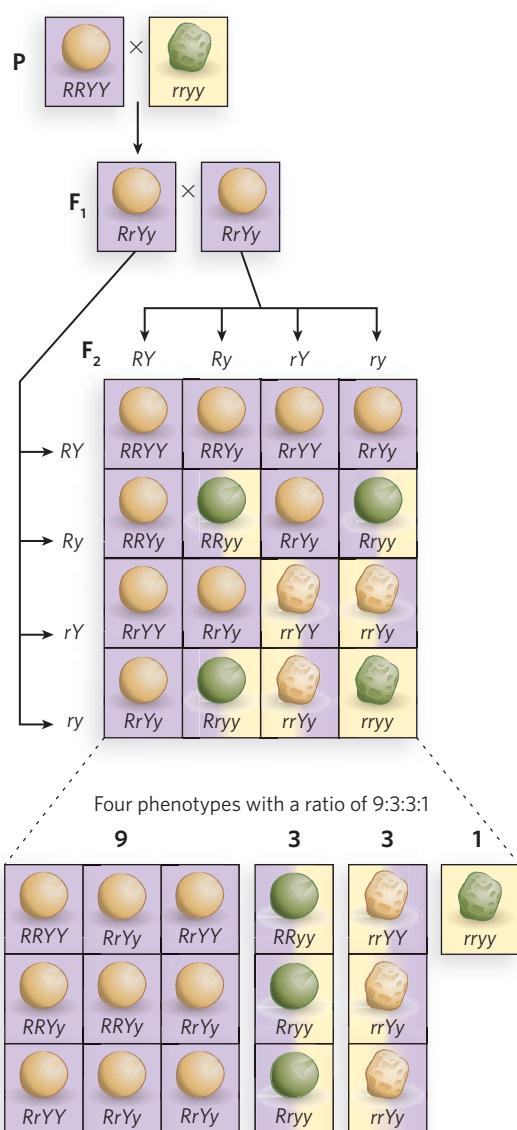


FIGURE 2-3 An example of Mendel's second law. Different genes assort independently into gamete cells. The parental cross (round, yellow seeds \times wrinkled, green seeds) yields uniform F₁ progeny with the dominant phenotype (round, yellow seeds) and genotype RrYy. The Punnett square analysis assumes random assortment of the different alleles into the gametes formed by the F₁ plant. All possible gamete genotypes are written across the top and left side of the grid. The predicted outcome for independent assortment is seeds of four different phenotypes in a 9:3:3:1 ratio, as illustrated below the Punnett square.

Mendel studied garden peas, but his basic principles hold true for sex-based inheritance in all animals and plants. Indeed, many human genetic diseases that can be traced through a family pedigree follow Mendel's simple rules of inheritance.

There Are Exceptions to Mendel's Laws

The transmission of dominant and recessive traits documented by Mendel is sometimes referred to as "Mendelian behavior." However, not all genes behave in such an ideal fashion. There are many exceptions to Mendel's principles of heredity, a few of which we review here.

Incomplete Dominance Some alleles of a gene are neither dominant nor recessive. Instead, hybrid progeny display a phenotype intermediate between those of the two parents. This type of non-Mendelian behavior is called **incomplete dominance**. An example of incomplete dominance can be seen in the gene for flower color in four o'clock plants (Figure 2-4). Homozygotes are either red (RR) or white (rr), but the F₁ heterozygote (Rr) is neither red nor white; it is pink. The molecular explanation for the pink heterozygote is the production of sufficient red color from the single R allele in the heterozygote to yield a pink coloration.

Table 2-3 Mendel's Results from a Two-Factor Cross

Characteristics Isolated for Study	Parental Cross	F ₁ Phenotype	F ₂ Phenotype	F ₂ Ratio
Seed shape and seed color	Round, yellow \times wrinkled, green	All round, yellow	315 round, yellow 101 wrinkled, yellow 108 round, green 38 wrinkled, green	8.3 2.7 2.8 1.0

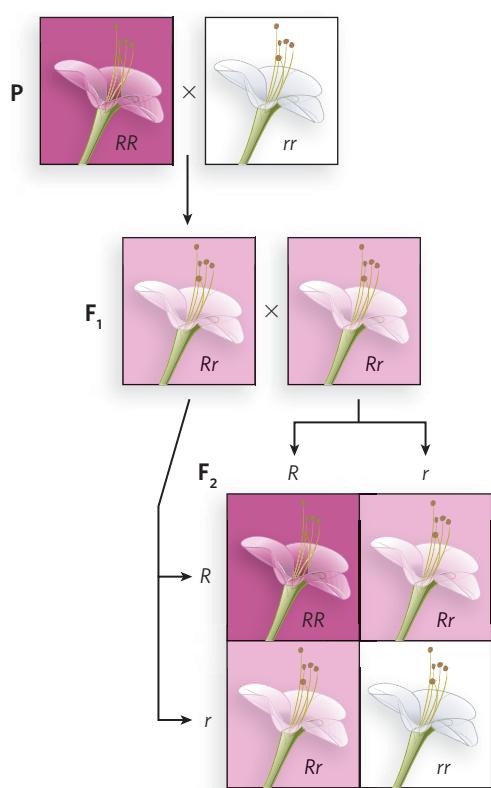


FIGURE 2-4 Non-Mendelian behavior: incomplete dominance in four o’clock plants. Cross-pollination of red- and white-flowered plants yields an F_1 plant with flowers of intermediate color (pink). Therefore, neither parental allele is completely dominant. The single R allele gives rise to sufficient red pigment to produce a pink coloration. Genotypes are given below each flower.

Interestingly, an example of incomplete dominance can also be found in Mendel’s published work. He studied different alleles for the gene controlling the pea plant’s flowering time. The F_1 progeny had a flowering time that was intermediate between the flowering times of the two parents.

Codominance Recessive alleles often produce non-functional proteins, or none at all. However, there are many examples of two alleles of a gene that produce two different functional proteins, neither of which is dominant to the other. This non-Mendelian behavior is known as **codominance**. An example of codominance is human blood type (Figure 2-5). The allele for A-type blood (I^A) results in a cell surface glycoprotein different from the glycoprotein encoded by the allele for B-type blood (I^B). People with A-type blood are homozygous $I^A I^A$, and those with B-type blood are homozygous $I^B I^B$. AB-type individuals are $I^A I^B$ heterozygotes. O-type individuals lack both varieties of surface glycoprotein; they are homozygous for the recessive i allele.

Linked Genes The most common non-Mendelian behavior is seen in **linked genes**, in which alleles for two different genes assort together in the gametes, rather than assorting independently. We now know that genes are located on chromosomes, and diploid organisms have two copies of each chromosome, known as **homologous chromosomes**, or homologs. During gamete formation, whole chromosomes, not individual genes, assort into gametes. Genes that are close together on one chromosome are inherited together, contrary to Mendel’s second law. Assortment of linked genes into gamete cells is shown in Figure 2-6 and is discussed in detail later in the chapter.

Mendel picked traits whose genes assorted independently. However, some of the genes that he studied are on the same chromosome. How could Mendel have observed independent assortment of genes on the same chromosome? As we describe later in the chapter, homologous chromosomes associate together during the cell divisions that lead to gametes. At that time, there is often an exchange of genetic material between the chromosome pair, resulting in some alleles previously

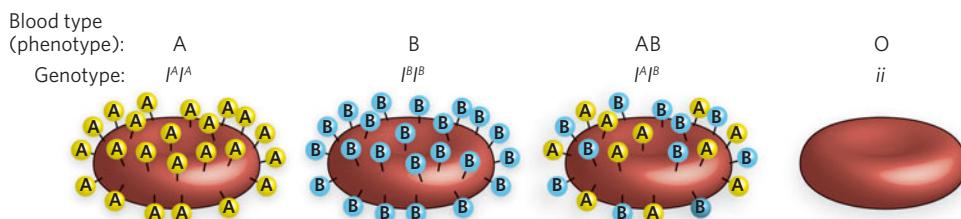


FIGURE 2-5 Non-Mendelian behavior: codominance in human blood types. The cell surface glycoprotein antigens on red blood cells (erythrocytes) determine human blood type. Two different alleles encode two variants of the enzyme glycosylase and produce different cell surface glycoproteins, A (allele I^A , yellow circles) and B (allele I^B ,

blue circles). Both alleles are expressed in the heterozygote (AB blood type, genotype $I^A I^B$). Because both alleles produce functional surface glycoproteins, neither allele is dominant to the other. Individuals with O-type blood have two null alleles (ii) and thus produce no surface antigens on their red blood cells.

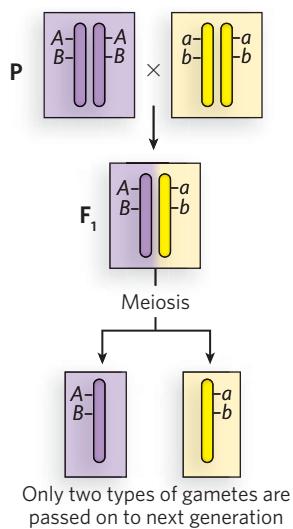


FIGURE 2-6 Non-Mendelian behavior: linked genes. Genes A and B are located on the same chromosome. Shown here is a cross between a homozygous dominant parent and a homozygous recessive parent that results in F_1 hybrid progeny, all $AaBb$. Because the alleles for the two genes are on the same chromosome (A, B and a, b), they cannot separate during the formation of gametes. The gametes contain one copy of this chromosome, and thus either the A and B alleles or the a and b alleles; no gametes containing A and b or a and B are formed. The two genes A and B are linked and do not assort independently during meiosis in the formation of gamete cells.

found on one chromosome now being found on the other. For the traits selected in Mendel's study, genes on the same chromosome were spaced far apart, and this swapping of genetic material occurred frequently between them. Therefore, the genes assorted as though they were on different chromosomes.

Many other types of non-Mendelian behavior exist besides those described above. These include traits determined by multiple genes, traits derived from interactions between different genes (epistasis), the inheritance of traits encoded by organelle genes (cytoplasmic inheritance), and traits that depend on whether the gene is inherited from the male or female parent (genomic imprinting).

SECTION 2.1 SUMMARY

- Mendel's studies on the garden pea revealed an underlying mathematical pattern in inheritance.
- Mendel postulated that genetic traits are particulate (now called genes). Diploid organisms contain two

copies, or alleles, of each gene and produce haploid gametes that contain one allele for each gene.

- Homozygous individuals have two identical alleles for a particular gene. In heterozygous individuals, the two alleles for a gene are different. The allelic makeup of an organism is its genotype.
- The different alleles for a gene may be dominant or recessive. In a heterozygote, the dominant allele masks the recessive allele in the outward appearance, or phenotype, of the organism.
- Mendel's first law, the law of segregation, states that the two alleles for each gene segregate independently into haploid gamete cells.
- Mendel's second law, the law of independent assortment, states that alleles for different genes assort into gametes randomly. However, we now know that genes reside on chromosomes and that chromosomes, not genes, assort randomly into gametes.
- There are exceptions to Mendel's laws. For example, a gene exhibits incomplete dominance when the phenotype of heterozygous progeny is intermediate between those of the two homozygous parents. Gene alleles exhibit codominance when both produce functional protein and neither is dominant to the other, as in human blood types. Alleles for two genes close together on the same chromosome are linked and can assort together into gametes.

2.2 Cytogenetics: Chromosome Movements during Mitosis and Meiosis

In 1865, Mendel presented his findings on inheritance in two lectures to the Brünn Society for the Study of the Natural Sciences, which were then published in an obscure journal. Only about 150 copies of this journal were printed, and Mendel's findings lay dormant for decades, to be resurrected only after his death. However, in his lifetime, Mendel was well appreciated at his monastery, was elected abbot, and managed one of the wealthiest cloisters in the land. His claim to fame was an incident in which he refused to pay a new tax imposed on the monastery by the Habsburg Empire. Mendel met the sheriff at the gate and dared him to take the keys from his pocket before he'd pay another pfennig! Of course, this is not what we know Mendel for today.

The years between 1880 and 1900 saw amazing discoveries in **cytology**, the study of cells, which intersected with the rediscovery of Mendel's work. Microscopes

had become more advanced, and chromosomes could be stained and visualized in the cell nucleus. Cytologists observed that, unlike other cellular components, chromosomes were meticulously divided between the two new cells during cell division. The diploid nature of somatic cells and the haploid nature of gametes were also discovered around this time. It was in this scientific environment of explosive growth in **cytogenetics** that, in 1900, Mendel's principles of heredity were rediscovered independently by three scientists: Hugo de Vries, Carl Correns, and Erich von Tschermark. The behavior of chromosomes was seen to remarkably mirror the behavior of Mendel's hereditary particles, and the idea that the nucleus, and perhaps the chromosomes themselves, formed the basis of heredity was bandied about.

In this section we describe the architecture of the cell and the chromosome movements that occur during somatic cell division and gamete formation, setting the stage for the chromosome theory of inheritance (the subject of Section 2.3).

Cells Contain Chromosomes and Other Internal Structures

Robert Hooke was the first to notice the cellular composition of a biological specimen, during his microscopic examination of cork in 1665 (Figure 2-7). In his famous book *Micrographia*, Hooke noted a multitude of tiny boxes in the cork sample and coined the word **cell** (Latin *cellula*, “small compartment”). By the early 1800s, it became clear that plants are made up of cells. In 1833, Robert Brown identified the nucleus, the first subcellular structure to be

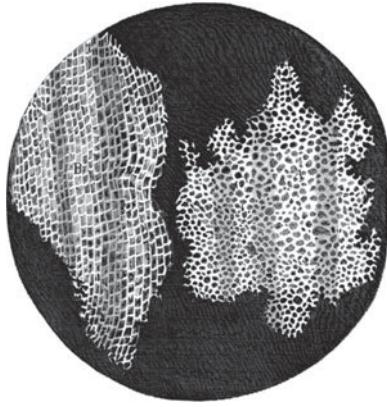


FIGURE 2-7 Hooke's microscopic examination of cork.

Robert Hooke used his compound microscope to visualize cells in cork and catalogued his work with meticulous drawings. His studies provided the first clue that organisms are cell-based.



FIGURE 2-8 The chromosomes of plant cells. A cross section of a rapidly dividing root tip of the onion (*Allium cepa*) shows the chromosomes as darkly stained bodies. Cells in different stages of division (mitosis) are apparent. [Source: Biodisc/Visuals Unlimited.]

discovered. In 1839, Theodor Schwann realized that animal tissue contains nuclei throughout the cells, and he proposed the **cell theory**, which states that all animals and plants consist of large assemblages of cells.

Microscopic studies of chromosomes within nuclei were first made in plant cells by Karl Wilhelm von Nägeli and Wilhelm Hofmeister between 1842 and 1849. **Chromosomes** were named (from the Greek for “colored body”) for their property of taking up large amounts of colored dye. The term was coined by Heinrich von Waldeyer in 1888. Figure 2-8 shows rapidly dividing cells of an onion, stained to show the chromosomes. Development of the electron microscope in 1931 eventually brought into view the detailed structure of the cell as we know it today. Each cell is bounded by the **cytoplasmic membrane**, encasing a variety of subcellular structures called **organelles**. Figure 2-9 is a schematic depiction of a typical animal cell, a eukaryotic cell; the caption describes each organelle and its function.

Mitosis: Cells Evenly Divide Chromosomes between New Cells

For biological information to be faithfully transmitted to daughter cells, it must be duplicated and then each complete information packet correctly partitioned into its own cell. As cells grow, they proceed through four phases of the

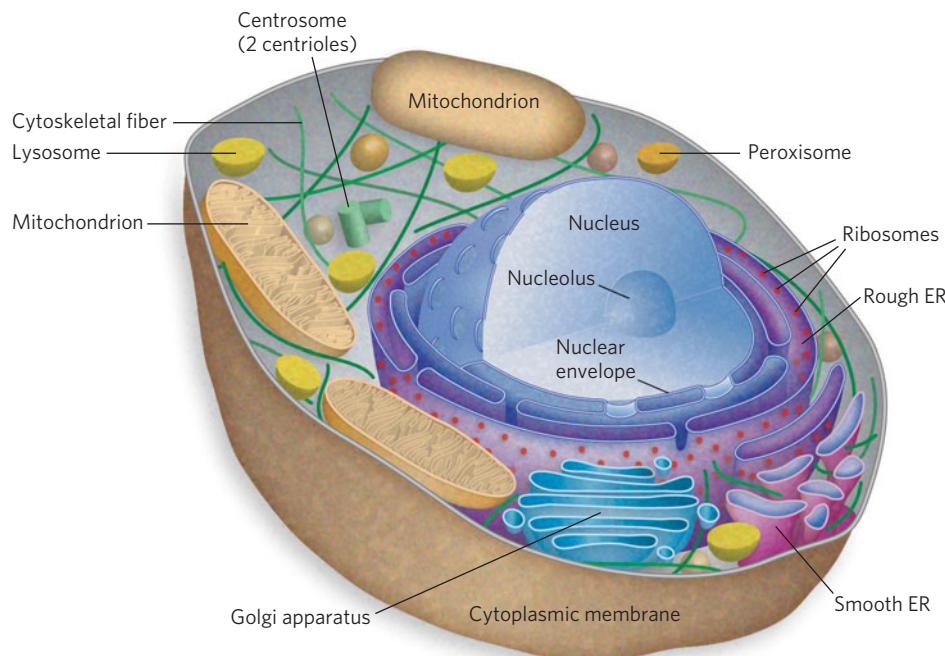


FIGURE 2-9 Animal cell structure. Eukaryotic cells are bounded by a cytoplasmic membrane. Chromosomes (not visible as individual structures in a nondividing cell) are located in the nucleus, and the nuclear envelope (nuclear membrane) is a double membrane with large pores through which RNA and proteins move. The nucleolus, a substructure within the nucleus, is the site of rRNA synthesis. All other organelles are in the cytoplasm. The centrosome consists of two perpendicular cylinders, the centrioles. During cell division, the centrosome duplicates and each new centrosome exudes a spindle apparatus, with spindle fibers connecting each centrosome to each chromosome. Mitochondria, the energy factories of animal cells, oxidize fuels to produce ATP. Lysosomes, containing

degradative enzymes, aid in digestion of intracellular debris and recycle certain components. Peroxisomes help detoxify chemicals and degrade fatty acids. The smooth endoplasmic reticulum (ER) is the site of lipid synthesis and drug detoxification. Ribosomes are ribonucleoprotein particles that act as protein-synthesizing factories; many attach to the ER, giving it a rough appearance. The rough ER sorts proteins destined for the cytoplasmic membrane or for other organelles; it is continuous with the outer membrane of the nuclear envelope. The Golgi apparatus, a membranous network, receives proteins from the ER and modifies and directs them to their proper compartments. Cytoskeletal fibers are a network of structural proteins that give shape to the cell and aid in cell movement.

cell cycle: G₁, S, G₂, and M (Figure 2-10). In **G₁ phase**, cells are diploid, containing two copies of each chromosome. The cellular chromosome content in G₁ cells is represented as 2n, where n is the number of unique chromosomes of that species. G₁ is also called the first gap, because it represents a gap in time before S phase.

During **S phase** (S for synthesis), each chromosome is duplicated, and the two identical chromosomes remain together as a **sister chromatid pair**. The point where the sister chromatids are joined is called the **centromere** (Figure 2-11). At the end of S phase, each homologous chromosome exists as a sister chromatid pair, thus the cell now contains four copies of each chromosome (i.e., the cell is 4n, or tetraploid). The cell next enters **G₂ phase**, or the second gap in time, after S phase.

The final phase of the cell cycle is **M phase**, or **mitosis**, in which the duplicated chromosomes separate and the cell divides into two daughter cells, each 2n. The two new cells reenter G₁ phase and then either continue through another division or cease to divide, entering a quiescent phase (G₀) that may last hours, days, or the lifetime of the cell (see Figure 2-10). Differentiated cells such as hepatocytes (liver cells) or adipocytes (fat cells) have acquired their specialized function and thereafter remain in G₀ phase.

Many scientists contributed to this description of the events during mitosis, and Walther Flemming figures most prominently among them. By the late 1870s, the quality of microscopes included such developments as the oil immersion lens and the substage condenser.

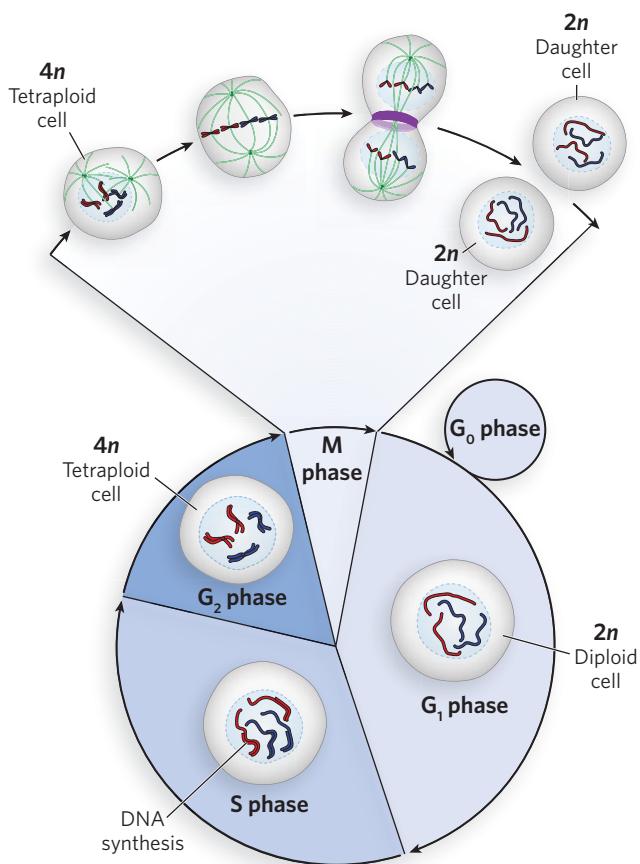


FIGURE 2-10 The eukaryotic cell cycle. Cells start in G₁ phase and progress to S phase, in which chromosomes are duplicated. In G₁, cells are diploid (2n, where n is the species' chromosome number). After S phase, cells are tetraploid (4n) and enter G₂ phase. In M phase, the duplicated chromosomes are equally divided, and the cell splits (by cytokinesis) into two daughter cells, each 2n. These cells can enter a quiescent phase, G₀, which removes them from the cell cycle, or can undergo further division. The duration of each phase varies with species and cell type. A typical human cell in tissue culture has a cell cycle of about 24 hours: G₁ phase, 6–12 hours; S phase, 6–8 hours; G₂ phase, 3–4 hours; and M phase, 1 hour.

These advances made possible Flemming's detailed observations of dividing cells, published in 1878 and 1882, revealing the stages of mitosis as we know them today. The steps of mitosis are illustrated on the left side of Figure 2-12 and summarized here.

Interphase. Cells not in mitosis are in interphase (which comprises G₁, S, and G₂). The cell is metabolically active and growing, and the chromosomes are duplicated (in S phase) in preparation for mitosis. The chromosomes are decondensed and not yet visible in the microscope.

Prophase. Prophase is the first stage of mitosis, following G₂ phase. As cells enter prophase, the sister chromatid pairs condense and become visible. Two

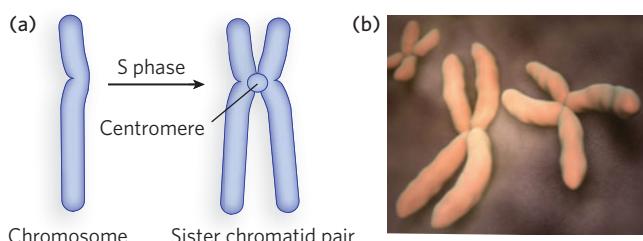


FIGURE 2-11 Formation of sister chromatid pairs by chromosome duplication. (a) Chromosomes are duplicated as cells proceed through S phase. Each resulting sister chromatid pair is held together at the centromere. (b) Three sister chromatid pairs, as seen by electron microscopy. [Source: (b) MedicalRF/The Medical File/Peter Arnold Inc.]

organelles, called **centrosomes**, move to opposite poles of the cell. There they give rise to the spindle apparatus, an organized structure of protein fibers, which also becomes visible during prophase.

Metaphase. In metaphase, the membrane surrounding the nucleus dissolves. The spindle apparatus becomes fully developed and attaches to the centromeres of the sister chromatid pairs, directing them to align in the equatorial plane of the cell, a site also known as the **metaphase plate**.

Anaphase. Each sister chromatid pair separates at the centromere, becoming two separate chromosomes. The spindle apparatus moves the separated chromosomes toward opposite poles of the cell.

Telophase. The two chromosome sets reach opposite cell poles, nuclear membranes (nuclear envelopes) re-form, and the chromosomes become less distinct as they decondense. Telophase ends with **cytokinesis**, the physical splitting of the cytoplasmic membrane and cell contents to form two daughter cells.

Meiosis: Chromosome Number Is Halved during Gamete Formation

Studies of fertilization in the late 1800s revealed that gamete cells contain only half the number of chromosomes found in somatic cells, and the union of two gametes reestablishes the diploid chromosome number; mitosis keeps this chromosome number constant during somatic cell division. These findings nicely explained how chromosome number is established and maintained in an organism, but they posed a new riddle: how are haploid gamete cells formed? The answer came in the 1880s from studies by Edouard van Beneden, Oskar Hertwig, and Theodor Boveri. Their studies of the ovary cells of a parasitic worm, *Ascaris*, revealed that the haploid female gamete, the egg

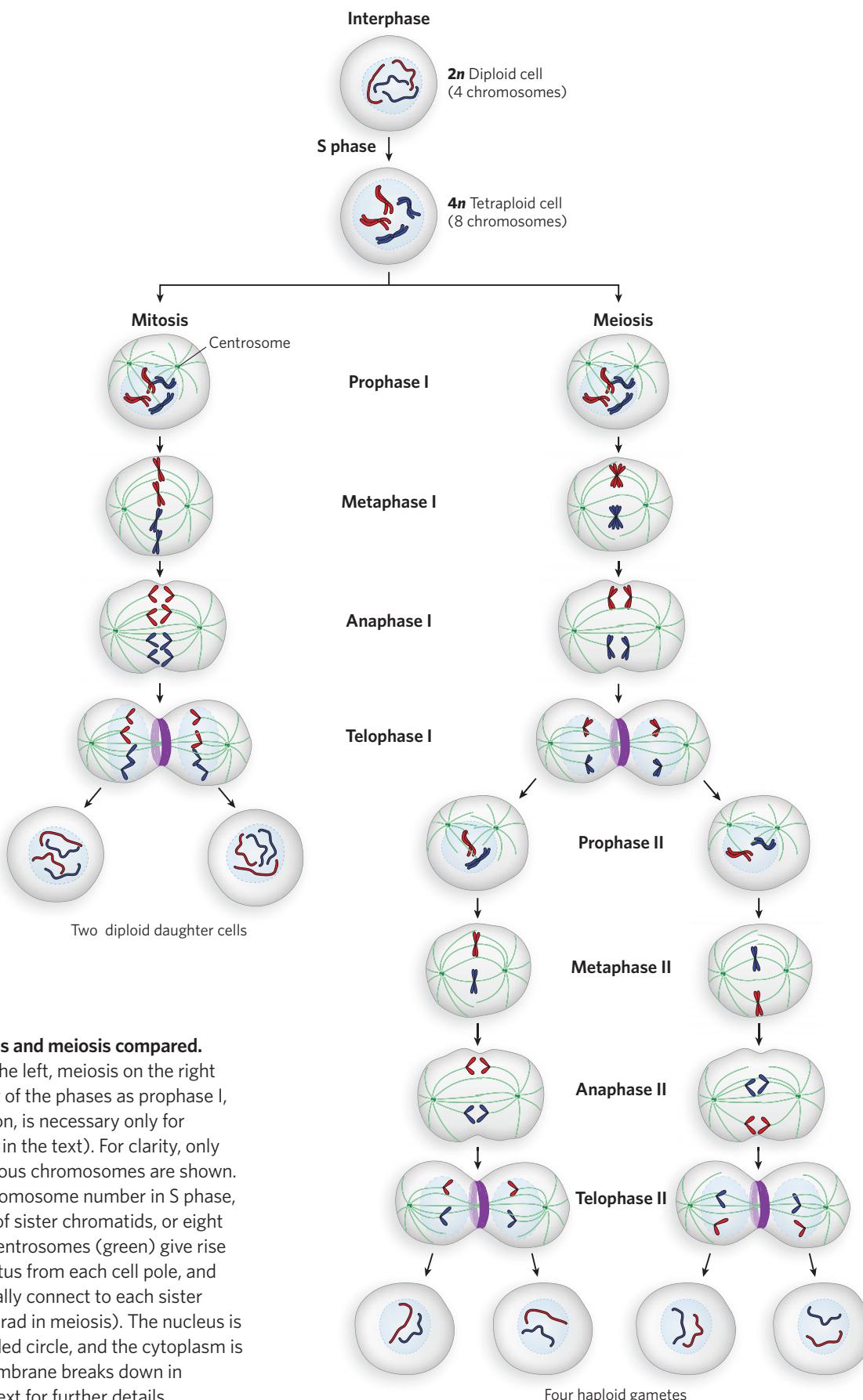


FIGURE 2-12 Mitosis and meiosis compared.

Mitosis is shown on the left, meiosis on the right (note that numbering of the phases as prophase I, metaphase I, and so on, is necessary only for meiosis, as explained in the text). For clarity, only two pairs of homologous chromosomes are shown. Cells double their chromosome number in S phase, producing four pairs of sister chromatids, or eight chromosomes. The centrosomes (green) give rise to the spindle apparatus from each cell pole, and spindle fibers eventually connect to each sister chromatid pair (or tetrad in meiosis). The nucleus is shown as a blue-shaded circle, and the cytoplasm is gray. The nuclear membrane breaks down in metaphase. See the text for further details.

(ovum), is formed by two consecutive cell divisions, in a process known as **meiosis** (see Figure 2-12, right). Later studies revealed that male gametes are also formed by meiotic cell divisions.

The most commonly studied organisms (such as *Ascaris*, sea urchin, or salamander larvae) had small, similar-looking chromosomes, and thus it was unclear whether chromosomes had unique identities. As far as anyone could tell, the cell simply divided an amorphous pool of chromosomes into equal parts to form daughter cells or gametes, rather than teasing apart two exact sets of different chromosomes. The unique nature of chromosomes, and the true precision with which the cell deals with them, was to come from work by Walter Sutton, in studies of the grasshopper. The grasshopper has chromosomes with unique morphologies that allow individual chromosomes to be observed during cell division. Sutton's observations demonstrated that during mitosis, one complete set of chromosomes is partitioned into each daughter cell. Chromosome segregation during meiosis proved to be just as precise, yielding one haploid set of chromosomes per gamete cell.

The Process of Meiosis Meiosis involves a halving of chromosome number to form haploid (n) gametes. One might expect that a diploid cell simply divides once to form two haploid gamete cells, but this is not so. Meiosis involves two successive cell divisions, and four haploid gametes are formed from one diploid cell. The meiotic cell first goes through S phase, just as in mitosis, thereby increasing the number of each chromosome to four copies per cell ($4n$). In sharp contrast to mitosis, however, the homologous chromosomes—each of which is a pair of sister chromatids—find each other in meiosis and physically associate to form a **tetrad**. In the first meiotic cell division, the tetrad splits and the two sister chromatid pairs segregate into two new cells, each $2n$. This differs from mitosis, in which each sister chromatid pair splits at the centromere, resulting in two chromosomes that segregate into the two daughter cells.

Whereas mitosis involves only one cell division, the daughter cells from this first meiotic division divide a second time, but without an intervening S phase (no additional chromosome duplication). This second cell division closely resembles mitosis, except that the cells are diploid ($2n$) going into the second meiotic division (rather than $4n$, as in mitosis), so the second division reduces the diploid chromosome number by half, to form haploid gametes (n). In other words, the second meiotic cell division resembles mitosis in that sister chromatids separate,

but in meiosis, for each chromosome, there is only one sister chromatid pair to split apart, whereas in mitosis the sister chromatid pairs of both homologous chromosomes are present at the metaphase plate, and each pair splits apart.

The phases of meiosis are summarized (and contrasted with mitosis) here and illustrated in Figure 2-12.

Interphase. Chromosomes are duplicated to form sister chromatid pairs; no obvious difference from mitosis.

Prophase I. Sister chromatid pairs become visible and the spindle apparatus forms. The difference from mitosis is that two homologous sister chromatid pairs find and associate with each other, forming a tetrad.

Metaphase I. The nuclear membrane breaks down, and the spindle apparatus moves the four homologous chromosomes to the metaphase plate as a tetrad, rather than moving two homologous but independent sister chromatid pairs as in mitosis.

Anaphase I. Centromeres stay intact, and sister chromatids do not separate. Instead, the tetrad splits and the two sets of sister chromatid pairs move to opposite poles. By contrast, in mitosis, sister chromatids split at the centromere and individual chromosomes move apart.

Telophase I. Telophase occurs as in mitosis. The nuclear membrane re-forms and the cell divides.

The second meiotic cell division is a lot like mitosis, but there is no S phase between divisions and the cell is diploid going into the second cell division.

Prophase II. As in mitosis, sister chromatid pairs are visible, but there are half as many as in mitosis because the homologous sister chromatid pair is no longer present (it is in the other daughter cell formed from the first division).

Metaphase II. As in mitosis, the nuclear membrane breaks down and sister chromatid pairs align in the equatorial plane.

Anaphase II. As in mitosis, the centromere splits and the two separated chromosomes move to opposite poles of the cell.

Telophase II. As in mitosis, cytokinesis results in two cells and the two nuclear membranes form. Unlike mitosis, the division produces daughter cells that are haploid (n).

Sex Determination Cytological studies in many types of cells documented the existence of one or two chromosomes that behaved strangely in meiosis during the formation of male gametes. Known as accessory or X



**Edmund B. Wilson,
1856–1939** [Source:
Columbia University.]



Nettie Stevens, 1861–1912
[Source: Columbia University
Archives.]

chromosomes, they either pair with a morphologically distinct partner chromosome or do not pair at all. Meiotic divisions therefore produce two types of sperm that differ in the accessory chromosome they contain. In 1905, Edmund B. Wilson and Nettie Stevens identified these accessory chromosomes in insects as the determinants of male and female sex, and referred to them as X and Y chromosomes, or **sex chromosomes**. All other chromosomes are called **autosomes**.

Sex can be determined in many different ways, depending on the type of organism. For example, in mammals, a common way is the XY system (Figure 2-13). In XY determination, the female is XX and the male is XY; the male gametes are of two varieties, carrying either the X or Y chromosome. In many insects, sex is determined by the XO system, in which females are XX and males have one X and no other sex chromosome. The male gametes contain either an X chromosome or no sex chromosome. In both XY and XO determination, the union of male and female gametes is equally likely to produce male or female offspring. In birds, some insects, and other organisms, the ZW system determines sex. It is like the XY system but in reverse: males have two of the same chromosome (ZZ), whereas females have one copy each of the Z and W chromosomes.

SECTION 2.2 SUMMARY

- Organisms are composed of cells, which have intricate intracellular structures, including chromosomes located in the nucleus.
- Cells that are not actively dividing contain two complete sets of unique chromosomes in the

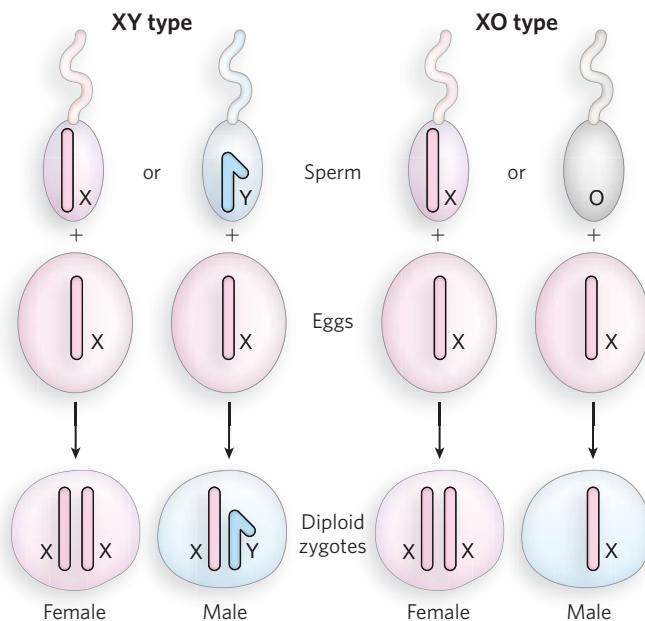


FIGURE 2-13 The chromosomal basis of sex determination. In the XY type of sex determination, meiosis produces sperm with either an X or a Y chromosome, in a 50:50 ratio (autosomes are not shown). All eggs have one X. Fertilization results in either an XX (female) or XY (male) zygote. The Y chromosome confers maleness. In XO determination, meiosis produces two types of sperm in a 50:50 ratio, either with or without an X chromosome (X or O). Fertilization results in either an XX (female) or an XO (male) zygote. The number of copies of X determines the sex.

nucleus; they are diploid, or $2n$ (except gametes, which are haploid, or n).

- The cell cycle consists of four stages; G₁ phase, S phase (synthesis), G₂ phase, and M phase (mitosis). The chromosomes of a diploid cell are duplicated in S phase (during interphase) and then carefully segregated into two daughter cells during mitosis, which proceeds through four stages: prophase, metaphase, anaphase, and telophase. The resulting daughter cells are also diploid.
- Meiosis is a specialized type of cell division that halves the diploid chromosome number ($2n$) to produce haploid gametes (n), each containing one complete set of chromosomes.
- Haploid gametes unite in fertilization to reestablish the diploid state of the organism.
- Sex is determined by an accessory chromosome that is paired either with a similar chromosome or with a distinct, differently shaped chromosome, or has no partner at all. These special chromosomes are called sex chromosomes; all other chromosomes are autosomes.

2.3 The Chromosome Theory of Inheritance

Walter Sutton's studies on chromosomes were performed just as Mendel's work was being rediscovered. Sutton found himself at a remarkable intersection of two fields: cytology and genetics. He made the connection between chromosomes and Mendel's particles of heredity in his classic 1903 paper, "The Chromosomes in Heredity" (see How We Know). He proposed that chromosomes contain Mendel's particles of heredity and that the particles come in pairs: chromosomes exist as homologous pairs in diploid cells. Mendel's particles—gene pairs—separate and assort independently into gamete cells; homologous chromosome pairs also separate and assort into haploid gamete cells.

Sutton's hypothesis that genes are located on chromosomes received much attention and became known as the **chromosome theory of inheritance**. But there was still no proof that genes were actually on chromosomes. This would be left for Thomas Hunt Morgan and his students to establish in their classic studies of fruit flies. Interestingly, Morgan did not initially believe in the chromosome theory of inheritance. But his experiments would inevitably lead him to this conclusion, and his name would become as linked to the chromosome theory as are genes themselves.

Sex-Linked Genes in the Fruit Fly Reveal That Genes Are on Chromosomes

In 1908, Morgan initiated his studies of the fruit fly, *Drosophila melanogaster* (see the Model Organisms Appendix). In those days, it was essential to keep costs to a minimum, as funding for science was scarce. The fruit fly is small; it could be grown in large numbers and was inexpensive to maintain. Flies also have an array of phenotypic features suitable for genetic studies, and just four homologous pairs of chromosomes that can be visualized under the microscope. Most important of all, the generation time of the fly is less than 2 weeks, and each female can lay hundreds of eggs. These features made fruit flies far superior to other model organisms of the day.



Thomas Hunt Morgan,
1866–1945 [Source: Caltech Archives.]

Morgan's famous Fly Room at Columbia University was small and cramped, but conducive to science; several of his students became famous for their discoveries in genetics. In 1910, Morgan noticed a male fly with white eyes that spontaneously appeared in a bottle of red-eyed flies. He immediately set up crosses to determine whether the white-eye trait was inheritable. It was, but in an unusual way. Figure 2-14 shows Morgan's first experiment. Normal flies have red eyes, straight wings, and a gray body, referred to as **wild-type** traits.

KEY CONVENTION

The allele that appears with the greatest frequency in a natural population of a species is called the wild-type allele. All other alleles are mutants. Wild-type alleles can be dominant or recessive to a mutant allele.

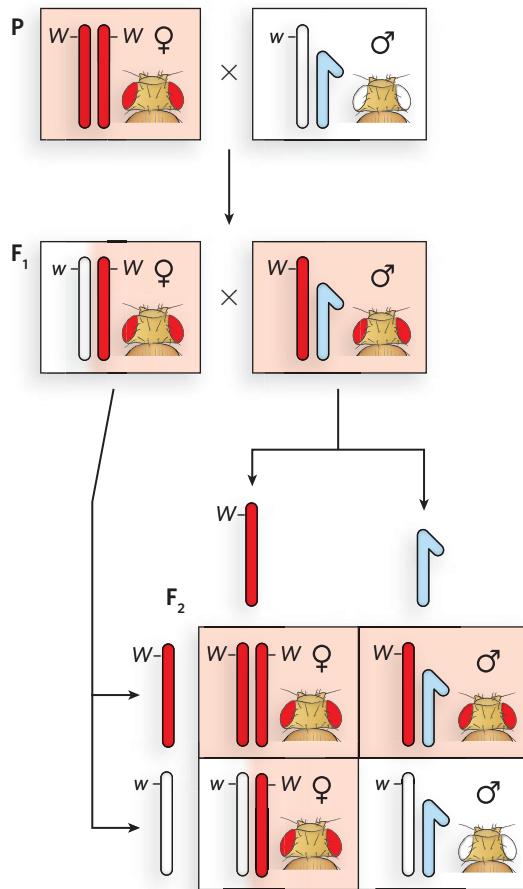


FIGURE 2-14 X-linkage of the white-eye allele. Results of a cross between a red-eyed female and a white-eyed male fly. F_1 flies are red-eyed. The white-eye trait reappears in the F_2 generation. Only males in the F_2 generation have white eyes. Morgan realized that these results make sense if the white-eye allele is located on the X chromosome. [Source: Adapted from T. H. Morgan et al., *Mechanism of Mendelian Heredity*, Henry Holt, 1915.]

Morgan crossed a red-eyed female with the mutant white-eyed male. All the F_1 hybrid progeny had red eyes, the wild-type phenotype. This told Morgan that the allele for red eyes is dominant to the allele for white eyes. His next cross, an F_1 female with an F_1 male, produced some F_2 progeny with white eyes, the expected result for a typical recessive allele. But surprisingly, all the white-eyed flies were male. All the F_2 females had red eyes, and about half the F_2 males had red eyes. It seemed that the trait for white eyes was somehow connected to sex. Morgan performed a variety of additional crosses and found, again to his surprise, that the white-eye trait mirrored the segregation behavior of the X chromosome (see Figure 2-14). Morgan's findings, linking a genetic trait to a particular chromosome, were convincing evidence that genes are located on chromosomes.

The alleles for this eye-color gene can be represented as X^W for red-eyed and X^w for white-eyed. In this genetic nomenclature, the X represents the X chromosome, a superscript W is used for the dominant red-eye allele and a superscript w for the recessive white-eye allele. The letters R and r (for red and white, respectively), which might be expected from the convention introduced earlier in the chapter, are not used in this case, because there are many different mutant alleles that affect eye color. There is only one wild-type color, so the wild-type and different mutant alleles are named according to the different mutant colors.

Further evidence that genes are located on chromosomes came from Calvin Bridges, an associate in Morgan's laboratory. Bridges hypothesized that if genes are located on chromosomes, then some genetic anomalies should also produce visible abnormalities in the chromosomes themselves. Bridges crossed white-eyed female flies (X^wX^w) with red-eyed males (X^WY) that he had in his fly collection. Most progeny were the expected white-eyed males and red-eyed females. However, Bridges noticed a few rare (<0.1%) white-eyed females and red-eyed males, which he called "primary exceptions." Bridges made the unusual prediction that if genes are truly on chromosomes, then primary exceptional flies will have an abnormal chromosome number. He reasoned that the primary exceptional phenotype might be explained by defective meiosis in the female parent, in which X chromosomes did not separate, producing an egg with two X chromosomes and an egg with no X chromosome (Figure 2-15a). Thus, exceptional white-eyed females, which must have two X^w chromosomes, received them from the abnormal X^wX^w egg, plus a Y chromosome from the sperm, for a genotype of X^wX^wY (note that an X^W sperm would bring in a dominant red-eye gene) (Figure 2-15b). By similar

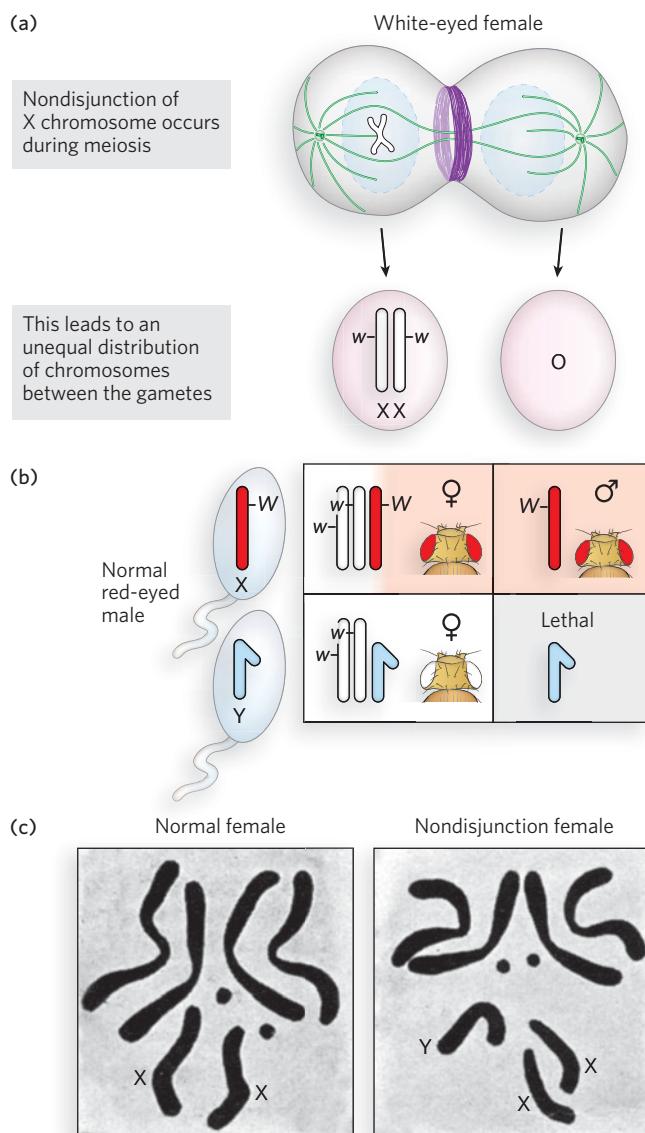


FIGURE 2-15 Nondisjunction. A rare occurrence during meiosis, nondisjunction produces gametes that have either one extra chromosome or one fewer chromosome. (a) Nondisjunction of the X chromosome is shown here; the white-eye allele is on the X chromosome. (b) Fertilization with normal sperm produces adult flies with odd numbers of chromosomes, as illustrated in the Punnett square. (c) Bridges predicted the occurrence of nondisjunction to explain rare progeny phenotypes, and cytologic examination of chromosomes in the rare progeny confirmed the predicted extra Y chromosome in a white-eyed female. [Source: (c) C. Bridges, *Genetics* 1:107–163, 1916.]

reasoning, the exceptional red-eyed male originated from fertilization of the abnormal egg with no X chromosome by a sperm containing a single X^W chromosome, for a genotype of X^WO . Note that although flies (like mammals) have X and Y chromosomes, sex in *D. melanogaster*

is determined by the number of copies of the X chromosome, not by the presence or absence of the Y chromosome. Thus, an XXY fly is female and an XO fly is male.

When Bridges examined the chromosomes of primary exceptional flies, the results followed his predictions precisely (Figure 2-15c). His study was an impressive demonstration that genes are located on chromosomes, because to explain the genetic results, he had hypothesized highly unusual outcomes that could be verified by examining the chromosomes directly. This abnormal assortment of chromosomes during meiosis is called **nondisjunction**.

Linked Genes Do Not Segregate Independently

Chromosomes, not individual genes, segregate into gamete cells, so one might expect two different genes on the same chromosome to stay together during meiosis and thus to be inherited together (i.e., they would not obey Mendel's second law). Take, for example, two genes, *A* and *B*, on the same chromosome. A cross of *AABB* and *aabb* parents will produce the *AaBb* *F*₁ hybrid, but particular combinations of alleles (*AB* and *ab*) are linked on the same chromosome. Therefore, the *F*₁ hybrid can produce only two types of gametes, *AB* and *ab*, rather than all four possible gametes produced if the genes separated and assorted randomly—*AB*, *Ab*, *aB*, *ab*.

To determine the genotype of an *F*₁ hybrid experimentally, it is crossed with a strain that is homozygous recessive (*aabb*), and the progeny reveal both the recessive and dominant alleles of the *F*₁ gametes. Such a cross is known as a **testcross**. If the two genes separate in the gametes of the *F*₁ hybrid, the *F*₂ generation will exhibit all four possible phenotypes. If the two genes are linked, the *F*₁ hybrid will produce only two types of gametes (*AB* and *ab*) and the *F*₂ generation will display only the two parental phenotypes (Figure 2-16).

An example of linked genes in *Drosophila* is illustrated in Figure 2-17, for a body-color gene with alleles *b* (black body) and *B* (gray body), and a wing-shape gene with alleles *v* (vestigial wings) and *V* (long wings). Consider the parental cross *BBvv* (gray body, vestigial wings) × *bbVV* (black body, long wings). All *F*₁ progeny (*BbVv*) have a gray body and long wings. To determine whether the two genes are linked, a testcross is performed between the *F*₁ fly and a double-recessive *bbvv* fly. The *F*₂ progeny are mainly of two types and exhibit the same characteristics as the *P* generation (gray body, vestigial wings; black body, long wings). Thus, the two genes are linked. Had the genes assorted completely independently, mixed phenotypes would have been observed in the *F*₂ generation (black body, vestigial wings; gray body, long wings) in amounts equal to the parental phenotypes.

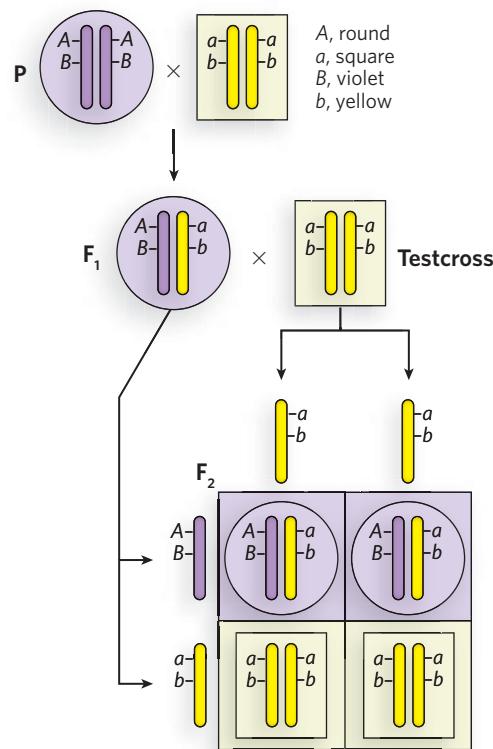


FIGURE 2-16 The inheritance of linked genes. Linked genes segregate together because they are on the same chromosome—that is, part of the same DNA molecule. The dominant and recessive genes, and their phenotypes, are: *A*, round; *a*, square; *B*, violet; *b*, yellow. The cross between homozygous dominant and homozygous recessive parents produces *F*₁ *AaBb* progeny with linked alleles *A*, *B* and *a*, *b*. A testcross with a double-recessive homozygous individual (*aabb*) reveals the genotypes of the gametes produced by the *F*₁ progeny. The Punnett square shows the expected results for completely linked genes. The *F*₁ generation can produce only *AB* and *ab* gametes, and thus only two types of *F*₂ progeny are observed; they have the same phenotype as the original *P* generation.

The results of the experiment, however, do not show complete gene linkage. There are some *F*₂ generation flies with mixed phenotypes, indicating that linked genes sometimes unlink. How can this happen—how do linked genes become unlinked?

Recombination Unlinks Alleles

Morgan noticed that linked genes do not always stay linked, but instead show a low, but reproducible, frequency of separating. Take, for example, the cross of flies with linked genes discussed above (see Figure 2-17). Linked genes should give only parental phenotypes in the *F*₂ progeny, yet a low frequency of mixed-phenotype *F*₂ progeny was observed. These **recombinant** flies could only be produced if the linked genes were unlinked and separated

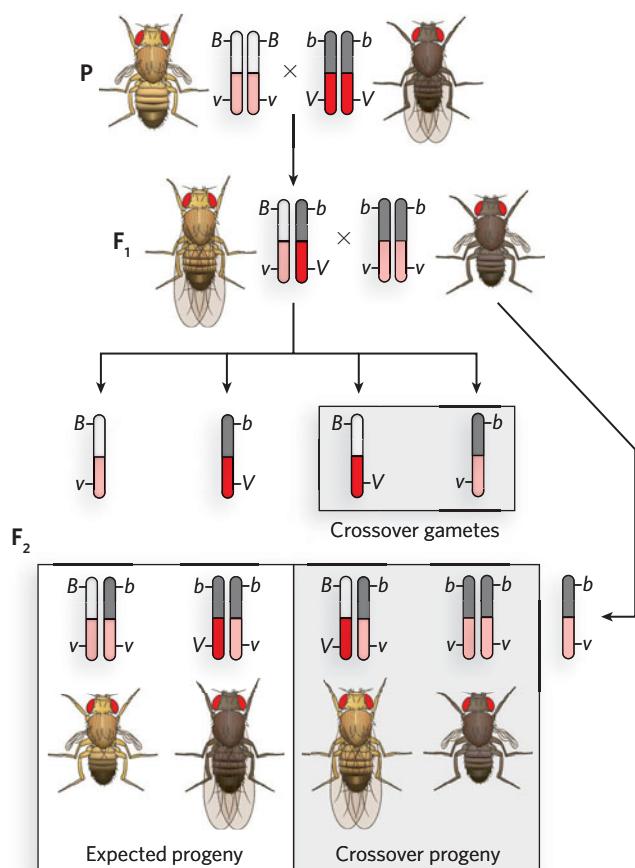


FIGURE 2-17 Unlinking genes by crossing over. Linked genes can become unlinked by chromosome recombination, or crossing over, during meiosis. The chromosomes containing the linked genes are illustrated in diploid cells and in gametes. To analyze gametes produced by F_1 flies, the F_1 hybrid is crossed with a double-recessive fly of genotype $bbvv$. In the F_1 gametes, B and v are linked, and b and V are linked, so all F_2 progeny are expected to contain these same two combinations. The double-recessive fly always contributes a bv gamete. But, in fact, four types of F_2 progeny are observed: two are the expected phenotypes, and the other two contain b , v and B , V , resulting from gametes in which the linked alleles were unlinked by recombination during meiosis. The two crossover phenotypes are produced at equal frequency (17% each). [Source: Adapted from T. H. Morgan et al., *Mechanism of Mendelian Heredity*, Henry Holt, 1915.]

during gamete formation. Both possible types of mixed-phenotype recombinant flies were produced (black body, vestigial wings; gray body, long wings) and appeared with equal frequency: 17% of the total F_2 population.

To explain how linked genes become unlinked, and why they produce equal amounts of the two mixed phenotypes, Morgan hypothesized that one of the linked alleles on one chromosome (e.g., the long-wing allele) trades places with the homologous allele (vestigial-wing allele) on the homologous chromosome (Figure 2-18). In other words, genes hop from one homologous

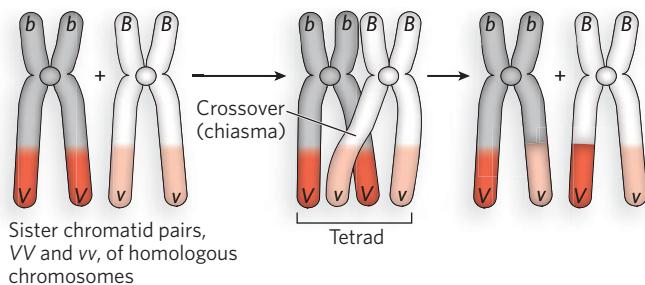


FIGURE 2-18 Crossing over in the tetrad. Two homologous chromosomes are duplicated in S phase to produce two sister chromatid pairs (left), one homozygous VV and the other homozygous vv . The sister chromatid pairs are homologous to each other, and they pair to form a tetrad before the first cell division of meiosis (middle). Recombination—as evidenced by the exchange of V and v alleles—occurs at crossovers, or chiasmata, the sites where chromosomes intertwine, resulting in genetic exchange between the two chromosomes of the sister chromatid pairs (right).

chromosome to the other and do so in a reciprocal fashion. This reciprocal exchange of alleles between chromosomes is called **recombination**, or **crossing over**.

The idea that chromosomes exchange genetic material had been suggested earlier, in cytological studies by F. A. Janssens in 1909. Janssens noticed that during meiosis, the four chromosomes of the tetrad coil around one another and form cross-shaped junctions, which he called **chiasmata** (see Figure 2-18). He proposed that, as the mechanical forces pull the sister chromatid pairs apart during the first division of meiosis, the intertwined chromosomes break at the same place and then rejoin, but with the opposite chromosome. The first experimental proof that genetic crossing over is mediated by physical recombination between two chromosomes came from a study of corn by Barbara McClintock and Harriet Creighton (see How We Know).

We now know that recombination events are mediated by specialized proteins that catalyze DNA breakage and rejoining within homologous chromosomes of the tetrad. Crossing over is a frequent event during meiosis (Figure 2-19), occurring at least once in each tetrad. It is thought that meiotic recombination was selected for during evolution because it helps generate diversity within a species. Homologous recombination is discussed in detail in Chapter 13.

Recombination Frequency Can Be Used to Map Genes along Chromosomes

Different pairs of linked genes exhibit different frequencies of crossing over, but the frequency is constant for a given pair of genes. Alfred Sturtevant, a student of Morgan's, rationalized this observation by assuming that

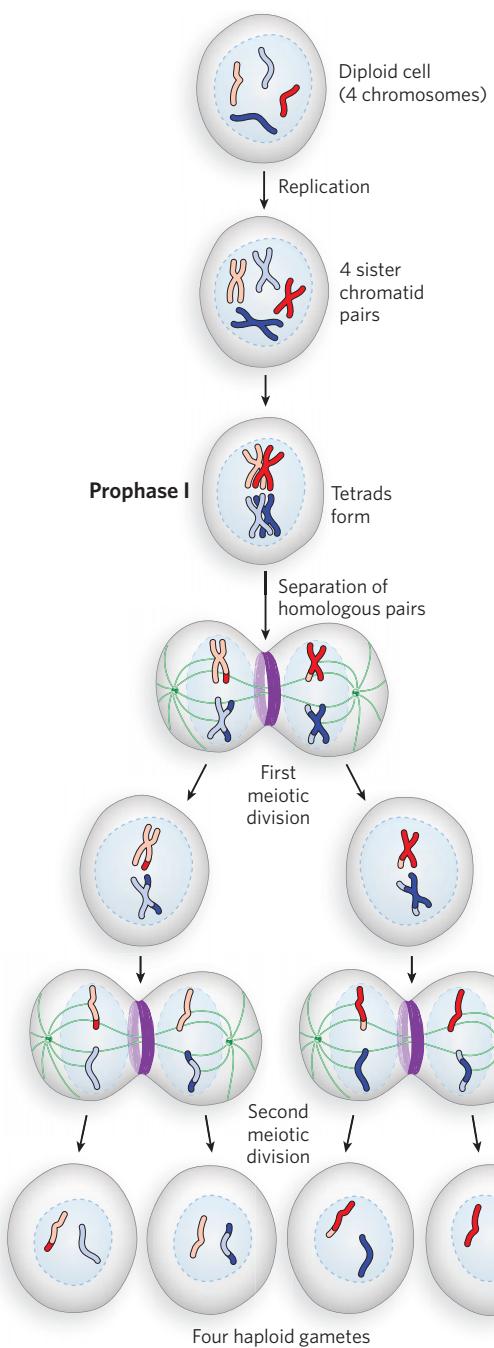


FIGURE 2-19 Recombination during meiosis. The chromosomes of a diploid cell (four chromosomes, two homologous pairs, are shown here) replicate, and each pair is held together at the centromere, forming four sister chromatid pairs. In prophase I, at the start of the first meiotic division, the two homologous sets of sister chromatid pairs align to form tetrads. Crossovers occur within the tetrads. In the first meiotic division, homologous pairs of chromosomes segregate into daughter cells. Each sister chromatid pair then lines up in preparation for the second meiotic division, which produces four haploid gamete cells. Each gamete has two chromosomes, half the number of the diploid cell. [Source: Adapted from D. L. Nelson and M. M. Cox, *Lehninger Principles of Biochemistry*, 5th ed., W. H. Freeman, 2008, Fig. 25-31.]

the frequency of crossing over corresponds to the distance between the two linked genes. He reasoned that the greater the distance, the more room there is for recombination to occur, thereby allowing linked genes to separate with greater frequency. With this logic, he used the frequency of crossing over to map the relative positions of pairs of linked genes along chromosomes (Figure 2-20). Genetic map units, calculated from the frequency of crossing over, are called centimorgans (cM) in honor of Thomas Hunt Morgan; however, they do not

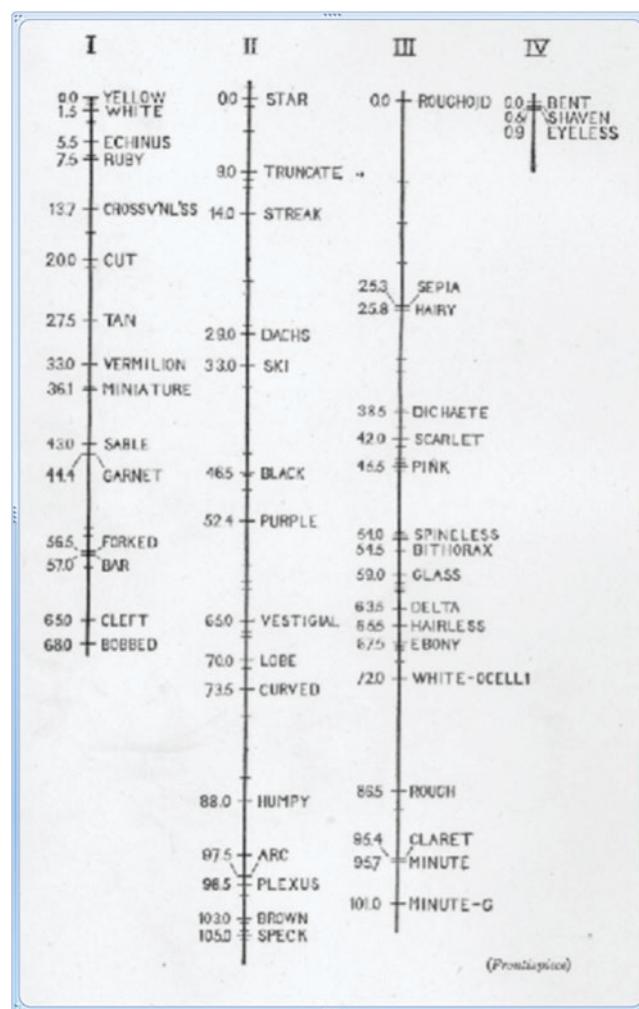


FIGURE 2-20 Using recombination frequency to create genetic maps. Sturtevant created genetic maps showing the positions of genes along the four *Drosophila* chromosomes, based on the frequency of crossing over between many pairs of linked genes. Linked genes fall into four groups, corresponding to the four different chromosomes in *Drosophila*, represented here by the vertical lines. Numbers on the left side of each chromosome are genetic map units, in centimorgans. Along the right side are the names of mutant alleles used in the crosses. [Source: T. H. Morgan et al., *Mechanism of Mendelian Heredity*, Henry Holt, 1915.]

necessarily reflect accurate physical distances between genes. Some regions of chromosomes tend to promote recombination, giving the impression that genes are farther apart than they really are; conversely, other regions repress crossing over, and genes seem to be closer than they are. The accuracy of genetic map distances is also limited by one crossing-over event interfering with another. However, recombination frequencies do provide useful genetic maps, because the data reveal the linear order of genes along a chromosome and provide a first approximation of the distance between them.

An example of **recombination mapping** is illustrated in [Figure 2-21](#) for linked genes A, B, and C. Consider the frequency of crossing over of the linked gene pair A and B in fruit flies (Figure 2-21a). A parent homozygous for dominant alleles is crossed with a

double-recessive fly ($AABB \times aabb$), and the frequency of crossing over (the frequency of production of Ab and aB gametes by the F_1 flies) is determined from the percentage of recombinant F_2 progeny ($Aabb$ and $aaBb$). This is repeated for the A, C pair and B, C pair. The results are shown in [Figure 2-21b](#).

The greater frequency of recombinants for the A, B pair than for the A, C pair indicates that genes A and B are farther apart than genes A and C. However, gene C could be between A and B or on the opposite side of A from B. The frequency of crossing over of the B, C pair resolves the ambiguity: C is between A and B.

The frequency of recombinants for the A, B pair (26%) is somewhat less than the added frequencies of recombinants for the B, C pair and C, A pair (28%). This is because the probability of multiple crossing-over

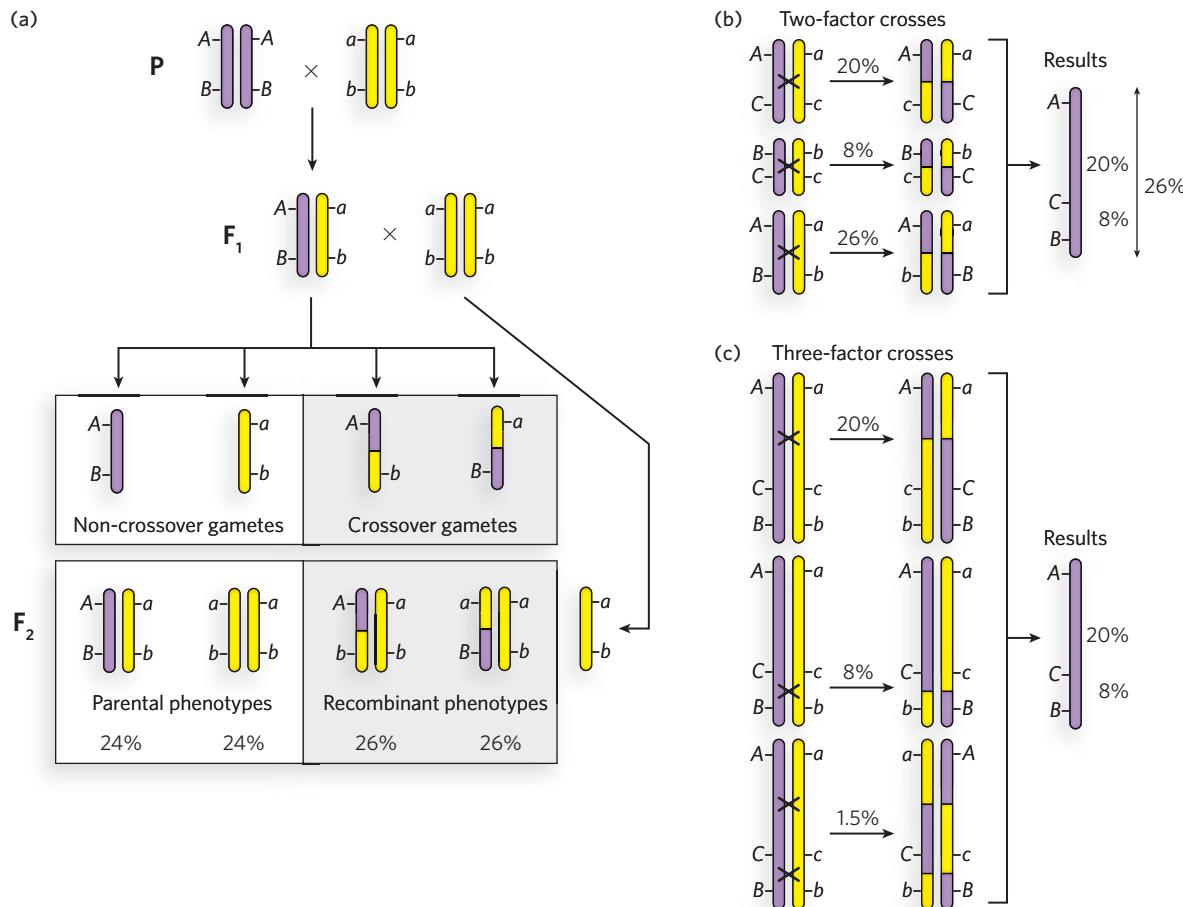


FIGURE 2-21 Recombination mapping. (a) The procedure for analyzing a two-factor cross. Diploid cells and gametes are illustrated to show the origin of recombinant F_2 progeny. Chromosomes with dominant linked genes are purple; chromosomes with recessive linked genes are yellow. Crossing over results in purple-and-yellow hybrid chromosomes. F_1 progeny are crossed with a homozygous

double-recessive fly to analyze the genotypes of gametes produced by the F_1 hybrid. (b) Analysis of linked genes using three two-factor crosses. Crossing over is indicated by an \times between two chromosomes. (c) Analysis of linked genes using three three-factor crosses. Two- and three-factor crosses lead to the same conclusion about gene order (ACB).

events between linked genes is higher the farther they are apart. For example, a single crossover unlinks the genes, but a second crossover links them again. Therefore, an odd number of crossovers will unlink genes, and an even number of crossovers relinks them, resulting in a maximum frequency of recombination of 50%. The frequency of independent assortment of genes on different chromosomes is also 50%, because there is a 50:50 chance that two chromosomes will segregate together into the same gamete. Because crossing over is a frequent occurrence in meiosis, genes on the same chromosome often assort independently. Therefore, recombination mapping is accurate only for pairs of linked genes that are close together.

Analysis of three genes in one experiment, known as a three-factor cross, provides a convenient method to identify or confirm their order along the chromosome. To illustrate this, consider a three-factor cross between genes *A*, *B*, and *C* (Figure 2-21c). A fly that is homozygous dominant for three linked genes is crossed with a fly that is double recessive for all three genes. The F_1 progeny ($AaBbCc$) are then crossed with a fly that is double recessive for all three genes. Most F_2 progeny exhibit the parental phenotypes, but crossing over will produce six possible recombinants: three recombinants containing two dominant traits and three reciprocal recombinants containing one dominant trait. If the gene order is *ACB*, generation of the *AcB* and *aCb* recombinants requires two crossover events—one between *A* and *C*, and another between *C* and *B*. The *aCB* (and *AcB*) or *ACb* (and *acB*) recombinants each require only one crossover. Because a double crossover is much less frequent than a single crossover, the far lower frequency of the double crossover (1.5% in this example, yielding *AcB* and *aCb*) reveals which gene (*C* in this case) is between the other two.

SECTION 2.3 SUMMARY

- Direct evidence that genes are located on chromosomes came from intensive studies of the fruit fly, *Drosophila melanogaster*, by Thomas Hunt Morgan. Segregation of the white-eye mutant allele with the X chromosome suggested that genes are associated with chromosomes.
- Calvin Bridges's correlation of mutant genes with chromosome abnormalities showed definitively that genes are located on chromosomes.
- Linked genes, genes on the same chromosome, violate Mendel's second law and assort together into

gametes. However, linked genes must be close together on the chromosome to stay linked. The farther apart they are, the more likely they are to be separated by recombination during meiosis.

- Recombination frequency can be used to map the relative positions of genes along a chromosome.

2.4 Molecular Genetics

The union of genetics and cytology in the early 1900s was an exceedingly productive time. Heredity was based in genes, and genes were located on chromosomes. But what are genes made of? To some scientists, genes were almost unreal, a mental construct to explain real phenomena. We now know that genes are made of DNA. In fact, DNA was discovered decades before its significance was understood. The recognition of DNA as the genetic material, and the solution of its chemical and three-dimensional structure, brought genetics out of the realm of imagination and into the realm of chemistry. These discoveries sparked the fusion of chemistry and genetics to give us an entirely new scientific discipline: **molecular genetics** or, more generally, **molecular biology**. In this section we outline some discoveries that led to our current understanding of DNA as the repository of biological information. We also describe how the information in DNA is translated into functional RNAs and proteins, and how this knowledge furthers our understanding of human health and disease.

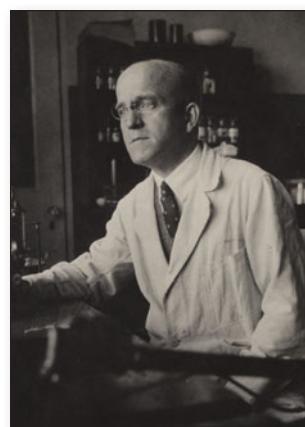
DNA Is the Chemical of Heredity

Deoxyribonucleic acid (DNA), as we've noted, was identified long before its importance was recognized. The history begins with Friedrich Miescher, who carried out the first systematic chemical studies of cell nuclei in 1868. Miescher obtained white blood cells from pus that he collected from discarded surgical bandages. He carefully isolated the nuclei and then ruptured the nuclear membranes, releasing an acidic phosphorus-containing substance that he called nuclein. Nuclein, a **nucleic acid**, was a new type of chemical polymer, different from all others previously identified. Around the turn of the century, Albrecht Kossel investigated the chemical structure of nucleic acids—both DNA and a similar molecule called **ribonucleic acid (RNA)**—and found that they contain nitrogenous bases, or nitrogen-containing basic compounds. Kossel identified five types of nitrogenous bases: adenine (A), guanine (G), cytosine (C), thymine (T), and uracil (U) (described in Chapters 3 and 6). By 1910, other investigators had determined that

nitrogenous bases were but one component of a larger unit called a **nucleotide**, which consists of a phosphate group, a pentose sugar, and a nitrogenous base. DNA is composed of A, G, C, and T nucleotides, and RNA is composed of A, G, C, and U nucleotides.

By the 1920s, the chemical basis of heredity was thought to lie in chromosomes, but chromosomes are composed of both DNA and protein. Which one is the hereditary chemical? DNA was initially ruled out because it was believed to be too simple—just a repeating polymer of four different nucleotides. Surely such a monotonous molecule lacked the complexity to code for the working apparatus of a living cell? Attention turned to the other biopolymer, protein, for a chemical explanation of heredity. Only after biochemical studies conducted in the 1940s pointed to DNA as the genetic molecule was attention refocused on the structure and function of this molecule.

DNA was shown to be the chemical of heredity in the 1940s, by Oswald T. Avery and his colleagues at the Rockefeller Institute in New York City. Their starting point was an observation made in 1928 by English microbiologist Frederick Griffith, who studied the pneumonia-causing bacterium *Streptococcus pneumoniae*. This pneumococcus exists as two types, virulent (disease-causing) and nonvirulent. Griffith noticed



Oswald Avery, 1877–1955

[Source: National Library of Medicine.]



Frederick Griffith,

1879–1941 [Source: Courtesy of Joshua Lederberg/Wiki.]

that virulent bacteria produced smooth colonies when grown on Petri plates, but colonies of a nonvirulent strain appeared rough. The difference in appearance lies in a capsule coat present only on the virulent strains. Griffith found that heat-killed virulent bacteria transformed live nonvirulent bacteria into live virulent bacteria (Figure 2-22a-d). The nonvirulent bacteria acquired the smooth-colony trait from the heat-killed

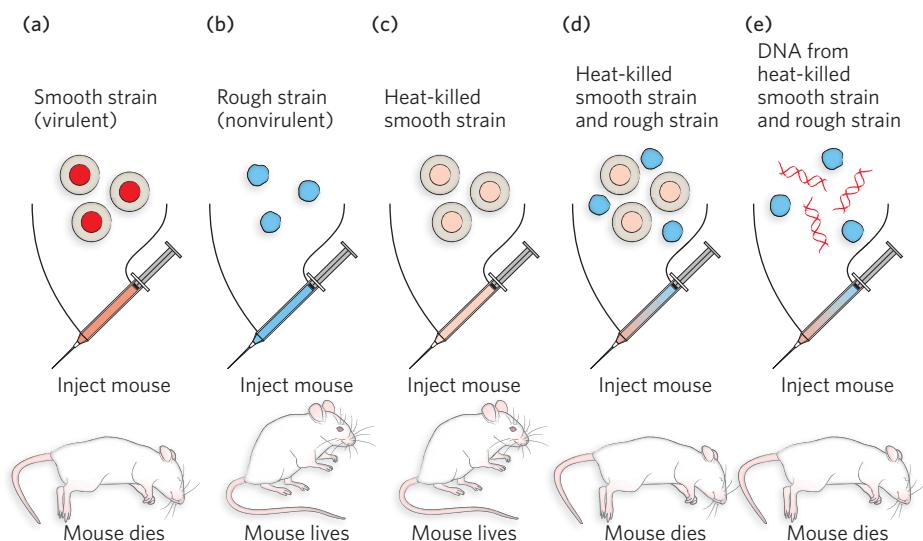


FIGURE 2-22 Transformation of nonvirulent bacteria to virulent bacteria by DNA. When injected into mice, the encapsulated strain of pneumococcus (*Streptococcus pneumoniae*) is lethal (a), whereas the nonencapsulated strain (b), and the heat-killed encapsulated strain (c), are harmless. (d) Griffith's research had shown that adding heat-killed virulent bacteria to a live nonvirulent strain (each

harmless to mice on their own) permanently transformed the live strain into lethal, virulent, encapsulated bacteria. (e) Avery and his colleagues extracted the DNA from heat-killed virulent pneumococci, removing RNA and protein as completely as possible, and added this DNA to nonvirulent bacteria, which were permanently transformed into a virulent strain.

bacteria. The results suggested that the genetic material coding for capsules remained intact even after the virulent (smooth-colony) bacteria were killed, and that this material could enter another cell and recombine with its genetic material.

Avery and his colleagues reproduced Griffith's results, and they analyzed the heat-killed virulent bacterial extract for the chemical nature of the transforming factor. They selectively removed either DNA, RNA, or protein from the heat-killed virulent bacterial extract by treatment with DNase, RNase, or proteases (enzymes that specifically break down one of these components). The DNase-treated extract lost the capacity to transform nonvirulent rough-colony cells into a virulent smooth-colony strain. The researchers then extracted DNA from virulent bacteria, purified it of contaminating proteins and RNA, and showed that this pure DNA was still capable of transforming nonvirulent bacteria into the virulent strain (Figure 2.22e). In 1944, Avery and colleagues reported their surprising conclusion that DNA was the carrier of genetic information. Another classic experiment, by Alfred Hershey and Martha Chase, supported this conclusion that DNA is the chemical of heredity (see How We Know).

Genes Encode Polypeptides and Functional RNAs

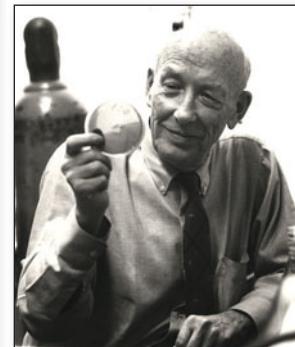
DNA and protein are chemically very different, and it was puzzling how a DNA sequence could code for a protein sequence. Regardless of the details, however, it now became easy to understand that mutations in a gene could lead to altered enzymes. In fact, even before the DNA structure was solved, the relationship between genes, mutations, and enzymes was well understood.

In 1902, the physician Archibald Garrod studied patients with alkaptonuria, a disease of little consequence for the patients, except that they excreted urine that turned black. Mendel's work had recently been rediscovered, and by noticing how alkaptonuria was inherited, Garrod realized that this disorder behaved as a recessive trait. It was already known that the synthesis and breakdown of biomolecules occurs in multistep pathways, each step requiring a different **enzyme**—a protein catalyst that facilitates the reaction. Garrod hypothesized that alkaptonuria was caused by a mutation that inactivated a gene required for the production of one enzyme in a metabolic pathway. Without this functional enzyme, the pathway was blocked, resulting in the buildup of an intermediate compound that was excreted and turned black. Garrod's reasoning drew the connection between a mutation in a gene and a mutation in an enzyme.



George Beadle, 1903–1989

[Source: Caltech Archives.]

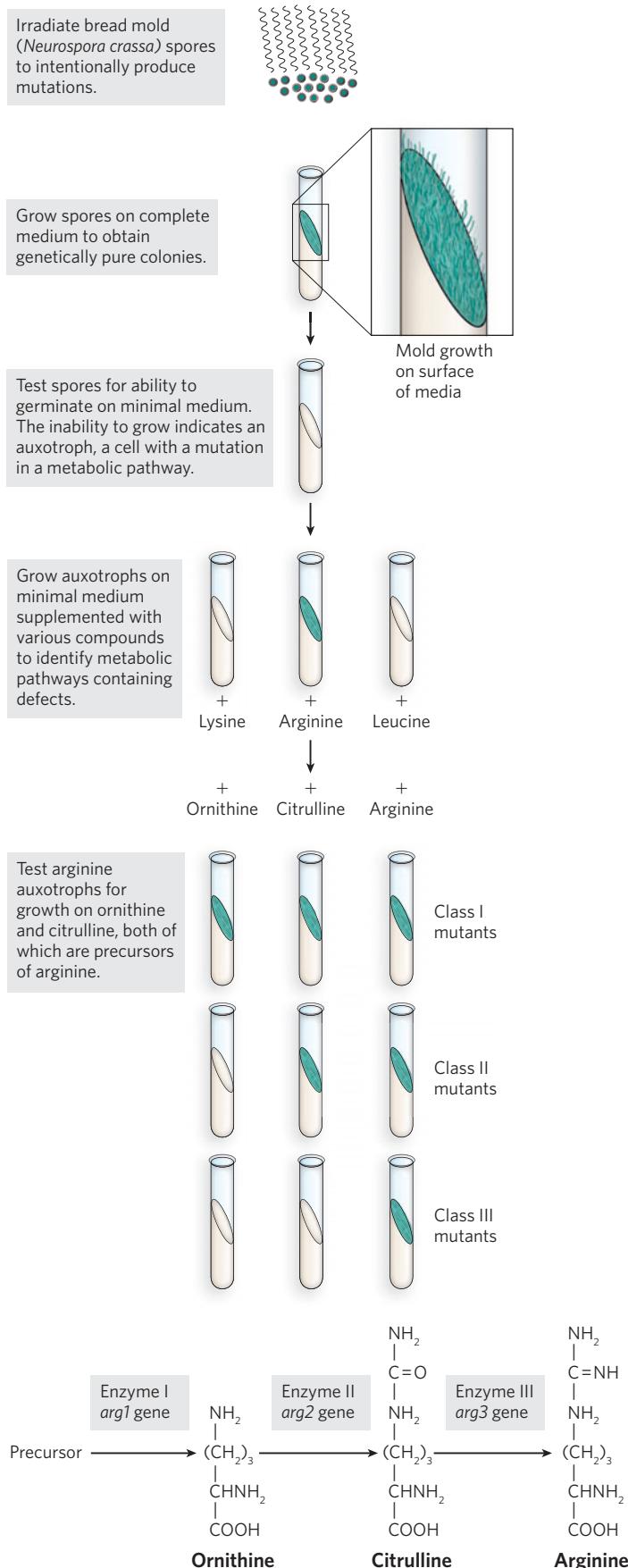


Edward Tatum, 1909–1975

[Source: Courtesy of Margaret & Barbara Tatum, Rockefeller Archive Center.]

Formal proof that genes encode enzymes came from a series of elegant experiments in the 1940s by George Beadle and Edward L. Tatum. They introduced a new microorganism into the study of genetics: the bread mold, *Neurospora crassa* (see the Model Organisms Appendix). This haploid organism can grow on a simple, defined medium, called minimal medium. Minimal medium contains sugar, nitrogen, inorganic salts, and biotin, and the cell must make all the rest of its own biochemicals that it needs to live from these simple starting compounds. Beadle and Tatum irradiated *Neurospora* spores to intentionally produce mutations, then germinated individual spores on a complete medium (i.e., one made with cell extracts that have all the amino acids, nucleotides, and vitamins) to obtain genetically pure colonies and their spores. These spores were then tested for their ability to germinate on the minimal medium. An inability to grow on minimal medium indicates a mutation in one of the metabolic pathways required for growth. These mutants are called **auxotrophs**. Spores of different auxotrophs were then analyzed for growth on a range of minimal media supplemented with selected compounds, to identify the defective metabolic pathways and the steps affected. An example of one type of study is illustrated in Figure 2.23, for auxotrophs of arginine metabolism.

Beadle and Tatum had a collection of *Neurospora* mutants that were auxotrophic for the amino acid arginine. The arginine synthetic pathway was known to include the intermediate compounds ornithine and citrulline, so they tested their arginine auxotrophs for growth on minimal media containing ornithine, citrulline, or arginine. The arginine auxotrophs fell into three classes, depending on which intermediate(s) they



required for growth (see Figure 2-23). Beadle and Tatum also mapped the mutant genes and found that mutants in a particular class of auxotrophs mapped to the same chromosomal location. They concluded that each class of mutant was caused by a single defective gene. Their findings also held true for genes in other metabolic pathways.

On the basis of these experiments, Beadle and Tatum proposed the *one gene, one enzyme* hypothesis, which stated that each gene codes for one enzyme. We now know that some enzymes are composed of multiple subunits encoded by different genes; furthermore, not all proteins are enzymes. So, the hypothesis was later revised to *one gene, one polypeptide*. A **polypeptide** is a chain of amino acids, and a functional protein can be composed of a single polypeptide or multiple polypeptide subunits. For a large number of genes, *one gene, one polypeptide* holds true. But as we will see throughout this textbook, even this hypothesis is not entirely accurate. Some genes code for functional RNAs rather than protein. And through a process called alternative splicing (see Chapter 16), some genes code for more than one polypeptide.

The Central Dogma: Information Flows from DNA to RNA to Protein

Watson and Crick's determination of DNA structure was a turning point in understanding how information flows in biological systems. Their model of DNA structure, which they reasoned from data collected by other scientists, consists of two strands of DNA wound about one another in a spiral, double helix. Each strand is composed of a long string of the four nucleotides

FIGURE 2-23 "One gene, one polypeptide" analysis of a *Neurospora crassa* auxotroph. Beadle and Tatum identified mutant *Neurospora* that were unable to synthesize the amino acid arginine (see text for details). To investigate the metabolic pathway of arginine synthesis, they analyzed arginine auxotrophs for growth on minimal medium plus ornithine or citrulline, both precursors of arginine (or on minimal medium plus arginine, to be sure that the mutant grows when it is supplied with arginine). They found that class I mutants grow when supplied with any of the three compounds, so these mutants lack an enzyme that is upstream of these three compounds (i.e., an enzyme catalyzing an earlier reaction) in the synthetic pathway. Class II mutants do not grow on ornithine, and thus lack an enzyme downstream of this intermediate but upstream of citrulline. Class III mutants grow only on arginine, and therefore lack an enzyme involved in the conversion of citrulline to arginine.

containing the bases adenine (A), guanine (G), cytosine (C), and thymine (T). The nucleotides in one strand pair with those in the other. Because A pairs only with T, and G pairs only with C, the sequence of each strand contains information about the sequence of the other, and the two strands are said to be **complementary**. The A-T and G-C pairs are referred to as **base pairs**. The detailed structure of DNA and the nucleotide bases, and how the nucleotides base-pair in a specific way, are described in Chapter 6.

The double-helical DNA structure immediately suggested a mechanism for the transmission of genetic information. The essential feature of the model is the complementarity of the two DNA strands. As Watson and Crick realized well before confirmatory data became available, the DNA could logically be replicated by separating the two strands and using each as a template to synthesize a new, complementary strand, thereby generating two new DNA duplexes that are identical to each other and to the original double-stranded DNA.

With discovery of the DNA structure, genetics could now be described in chemical terms. Both DNA and proteins are linear polymers, so the sequence of nucleotides in DNA must somehow be converted to a sequence of amino acids. But DNA is located in the nucleus, whereas proteins are synthesized in the cytoplasm. Therefore, there must be an intermediary molecule to shuttle information between the two locations. RNA was believed to play a role in this, and the similarities between DNA and RNA made it a simple matter to understand how an RNA molecule could be made from a DNA template. Crick proposed that biological information flows in the direction DNA→RNA→protein, and that DNA acts as a template for its own synthesis (DNA→DNA) (Figure 2-24). Crick's proposal is known as the **central dogma** of information flow. Although largely accurate, exceptions to the central dogma do exist. For example, certain enzymes can synthesize DNA from RNA (RNA→DNA), and some viruses use RNA as a template to make more RNA (RNA→RNA).

RNA was widely expected to be the molecule that mediates the transfer of information from DNA in the nucleus to the site of protein synthesis in the cytoplasm. However, no one imagined that three different types of RNA would be required for the process.

Ribosomal RNA In the early 1950s, Paul Zamecnik and his colleagues identified the site of protein synthesis as particles in the cytoplasm called **ribosomes**. Ribosomes are large structures composed of both protein and RNA.

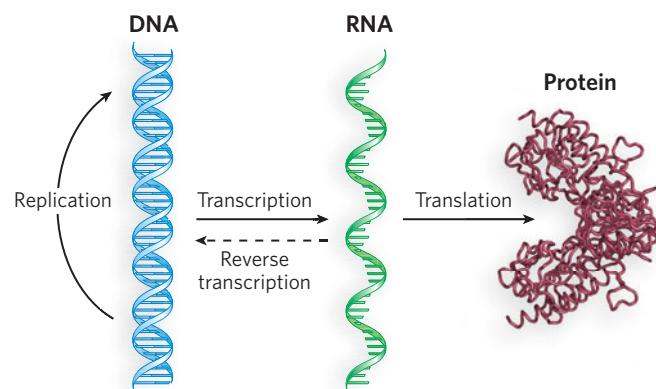


FIGURE 2-24 The central dogma of information flow: **DNA→RNA→protein**. The information to replicate DNA is inherent in its structure (curved arrow). Information flows from DNA to RNA by transcription. Information flows from RNA to protein by translation. In some instances, information can also flow backward, from RNA to DNA (reverse transcription). No evidence exists for information flow from protein to nucleic acid.

The RNA component is called **ribosomal RNA (rRNA)**. In bacteria and eukaryotes, ribosomes consist of a large subunit and a small subunit.

Messenger RNA The combined findings that ribosomes are the site of protein synthesis and that rRNA is the most abundant RNA in the cell (>80%) led most researchers to believe that rRNA was the carrier of information from DNA to protein. However, some features of rRNA are incompatible with its function as an information carrier. For example, rRNA is an integral part of the ribosome, so there would have to be specific ribosomes to make specific proteins. Further, the nucleotide composition of rRNAs from different organisms was relatively constant, whereas the nucleotide composition of chromosomal DNA varied considerably from one organism to the next.

Studies by Sydney Brenner, Jacques Monod, and Matthew Meselson in the early 1960s, using *Escherichia coli* (see the Model Organisms Appendix), suggested that another type of RNA carries the message from DNA to protein. They discovered a class of RNA that targets preexisting ribosomes, and the nucleotide composition of this RNA was more similar to chromosomal DNA than was rRNA. These properties are exactly those expected for a true messenger between DNA and protein. The investigators called this RNA **messenger RNA (mRNA)** and concluded that ribosomes are protein-synthesizing factories that use mRNA as a template to direct construction of the protein sequence.

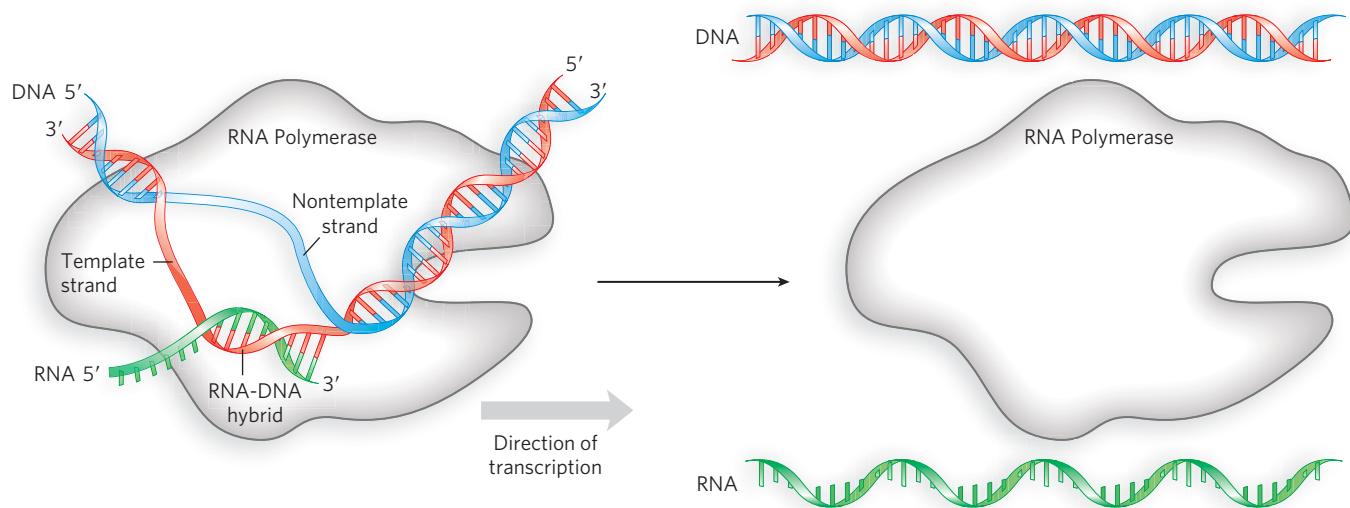


FIGURE 2-25 The process of transcription. RNA polymerase opens the DNA duplex and uses one strand as a template for RNA synthesis. The polymerase matches incoming nucleotides to the DNA template strand by base pairing and joins them together to form an RNA chain. As

RNA polymerase advances along the template strand, the two DNA strands reassociate behind it to re-form the double helix. When the gene has been completely transcribed, the polymerase dissociates from DNA, releasing the completed RNA transcript.

RNA synthesis is carried out by the enzyme **RNA polymerase**, which synthesizes RNA by reading one strand of the duplex DNA, pairing RNA bases to the bases in the DNA strand, to synthesize a single-stranded RNA molecule that has a sequence directed by the DNA sequence (Figure 2-25). This process of making single-stranded RNA copies of a DNA strand is known as **transcription**.

Transfer RNA The discovery of mRNA was a crucial piece of the information puzzle. But a problem remained: how is a sequence of nucleotides in mRNA converted to a sequence of amino acids in protein? Furthermore, DNA and RNA each consist of only four different nucleotides, whereas proteins have 20 different amino acids. Hence, one must assume the existence of a code that uses combinations of nucleotides to specify amino acids. Combinations of two nucleotides yield only 16 permutations (4^2). Combinations of three nucleotides yield 64 permutations (4^3), more than enough to specify a code for 20 amino acids.

In 1955, Crick hypothesized the existence of an adaptor molecule, perhaps a small RNA, that could read three nucleotides and also carry amino acids. It was not long after Crick's adaptor hypothesis (see Chapter 17) that Paul Zamecnik and Mahlon Hoagland discovered a small RNA to which amino acids could attach. This small RNA, later called **transfer RNA**

(**tRNA**), was the adaptor between nucleic acid and protein.

The discovery of tRNA, combined with the idea of a three-letter code, suggested how the DNA sequence could be converted to an amino acid sequence. Three bases in the tRNA form base pairs with a triplet sequence in the mRNA. When two amino acid-linked tRNAs align side-by-side on the mRNA by base-pairing to adjacent triplets, the amino acids attached to the tRNAs can be joined together. By continuing this process over the length of an mRNA strand, amino acids carried to the mRNA by tRNAs become connected together in a linear order specified by the mRNA sequence. These connections occur as the mRNA-tRNA complexes thread through the ribosome. The overall process of protein synthesis, involving three different types of RNA molecules, is known as **translation** (Figure 2-26). (Translation is covered in detail in Chapter 18.)

All RNAs, whether they code for protein or not, are transcribed from DNA genes. Messenger RNA is needed only transiently, to instruct the synthesis of proteins. But the end products of tRNA and rRNA genes are the RNA molecules themselves. These **functional RNAs** fold into specific three-dimensional shapes and constitute about 95% of the RNA in the cell. There are other types of functional RNA besides rRNA and tRNA (see Chapter 15), and new ones are

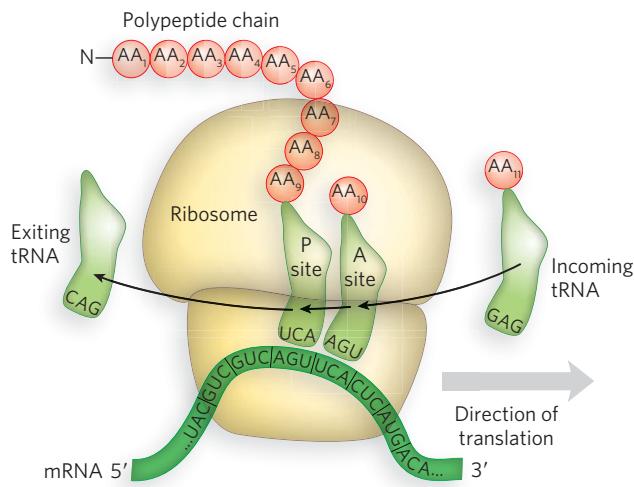


FIGURE 2-26 The process of translation. The ribosome, composed of a large and a small “subunit” (each consisting of many proteins and several rRNAs), mediates protein synthesis in cells. It associates with both mRNA and tRNAs as it synthesizes polypeptide chains. The ribosome has two major sites for binding tRNA molecules, the P site and the A site. Two tRNAs form base pairs with their respective, adjacent matching triplets on the mRNA: the tRNA in the P site carries the growing polypeptide chain, and the tRNA in the A site carries an amino acid (AA). The ribosome catalyzes the transfer of the polypeptide attached to the P-site tRNA to the amino acid on the A-site tRNA. The ribosome then shifts relative to mRNA so that the A-site tRNA, now holding the polypeptide, moves into the P site (and the tRNA previously at the P site departs). The next tRNA carrying an amino acid then binds to the vacated A site to continue extending the polypeptide chain.

almost certain to be discovered in the future. The increasing number of functional RNAs being discovered supports the proposal that many processes in early life forms were performed by RNA rather than proteins, and RNA continues to be of central importance to cellular function today.

Mutations in DNA Give Rise to Phenotypic Change

Most cellular functions are carried out by proteins. The precise sequence of amino acids in each protein molecule and the specific rules governing the timing and quantity of its production are programmed into an organism’s DNA. When changes in the DNA sequence occur, cellular function can be altered. Mutations in DNA can be beneficial or harmful to an organism, or can have no effect at all. For example, if the mutation does not change the sequence of a pro-

tein or how the protein is regulated, the mutation has no effect and is said to be silent. Evolution depends on mutations that are beneficial, and these usually alter the sequence or regulation of a protein in a way that enhances its function or confers a new, beneficial function that increases the viability of the organism. However, most mutations that change a protein sequence are harmful, because they lead to altered proteins with decreased function or new, detrimental function, and give rise to various diseases. When these DNA mutations occur in germ-line cells (cells that give rise to gametes), the disease can be inherited. There are many examples of inherited diseases, some of which have altered the course of history. One such disease is hemophilia.

Hemophilia occurred in the interrelated royal families of England, Russia, Spain, and Prussia in the 1800s. At the root of this malady is an inability of the blood to clot, resulting in excessive bleeding from even the slightest injury. The disease typically results in death at a relatively early age. Tracing hemophilia through the royal families of Europe indicates that it originated with Queen Victoria (Figure 2-27a). It is interesting that none of the current family members are carriers, presumably the result of natural selection against this trait.

Hemophilia is about 10,000 times more common in males than in females. This is because the blood-clotting factor involved in 90% of cases of this disease is factor VIII, encoded by a gene on the X chromosome. A mutant recessive allele of the factor VIII gene is responsible for hemophilia A, the most common form of the disease. Males have only one copy of the X chromosome, and the recessive allele, when inherited, is always expressed. Females have two X chromosomes, and if one X contains a wild-type allele, it masks the expression of the mutant recessive allele. A female with only one copy of the recessive allele is called a carrier, because she is phenotypically normal but may pass on this allele to her offspring (Figure 2-27b).

Many other inherited diseases have been mapped to their particular genes. One of the first to be identified was the gene involved in Huntington disease (Figure 2-28). The gene, *HTT*, is located on chromosome 4. The disease is associated with a region of the *HTT* gene where there can be a variable number of repeats of the triplet nucleotide sequence CAG (encoding the amino acid glutamine). The *HTT* gene in healthy individuals has about 27 or fewer of these repeats, but when the number exceeds 36, it is often associated with disease. The likelihood of having Huntington disease increases with the number of tri-nucleotide repeats in the *HTT* gene. The function of

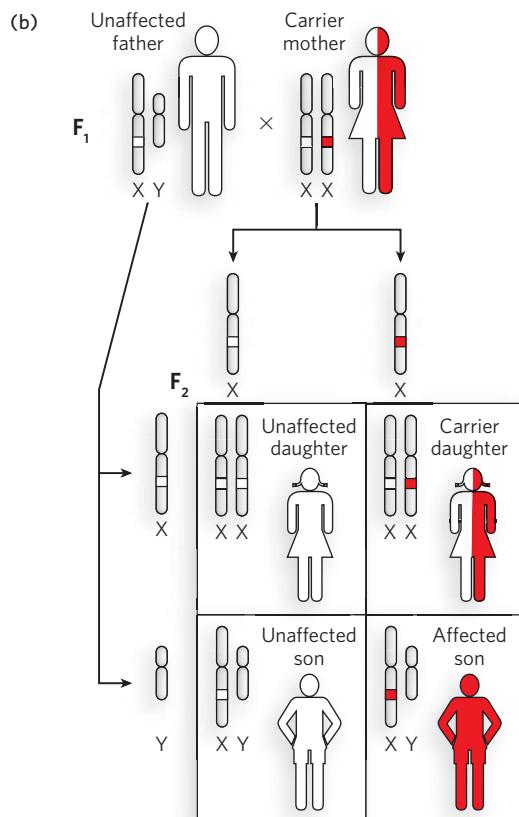
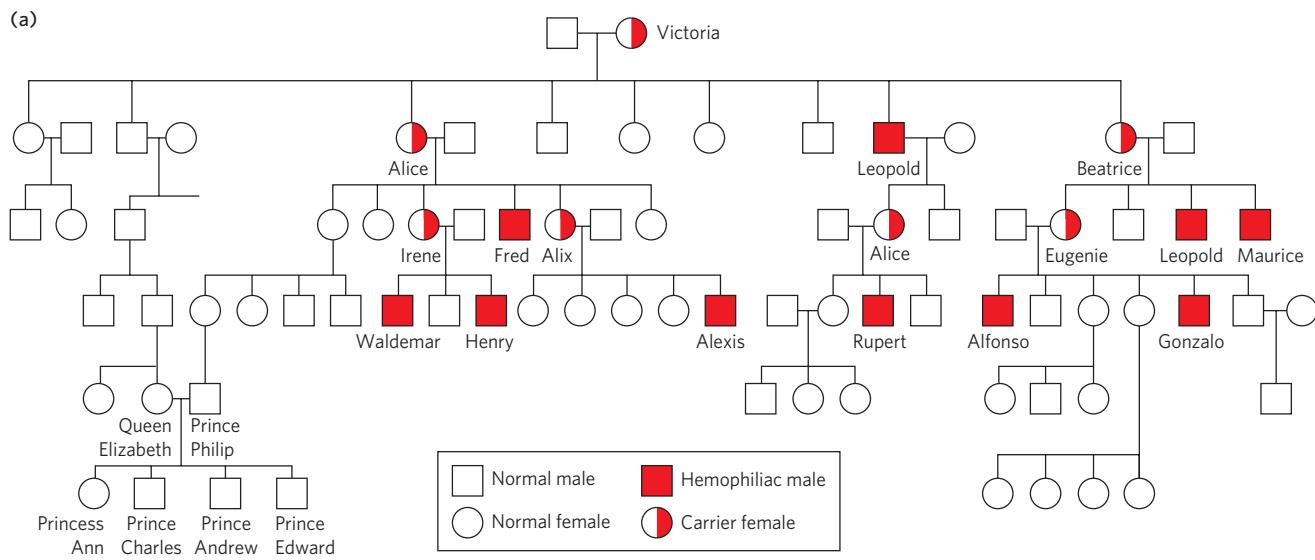


FIGURE 2-27 The inheritance of hemophilia. (a) The hereditary pattern of hemophilia in the royal families of Europe reveals that it is a recessive X-linked disease. (b) Because females have two copies of the X chromosome, they can carry one copy of the mutant gene for hemophilia without exhibiting disease; they have hemophilia only if both X chromosomes carry the mutant gene. Male offspring, having only one X chromosome, are more likely to have the disease; hemophilia occurs about 10,000 times more frequently in males than in females.

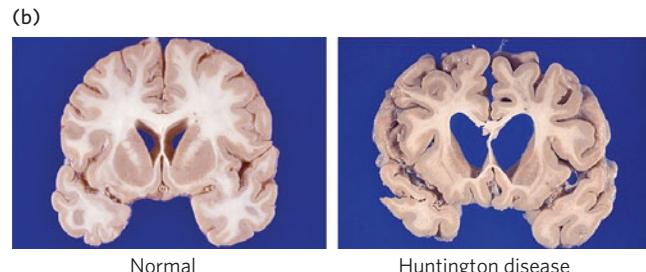
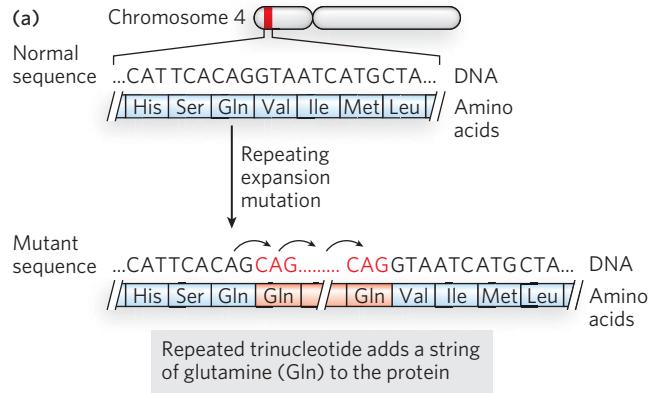


FIGURE 2-28 Huntington disease. Huntington disease is an inherited autosomal recessive neurological disease. (a) The gene for Huntington disease (*HTT*) is located on the short arm of chromosome 4, and the disease is associated with CAG repeats (CAG encodes glutamine) in this gene. When the number of CAG repeats increases above 36 copies, disease may occur in midlife. (b) Huntington disease affects the brain by causing degeneration of the basal ganglia and cerebral cortex. [Source: (b) U.S. National Library of Medicine.]

HIGHLIGHT 2-1 MEDICINE

The Molecular Biology of Sickle-Cell Anemia, a Recessive Genetic Disease of Hemoglobin

Genetics, molecular biology, and evolution by natural selection all converge in a striking fashion in sickle-cell anemia, a human hereditary disease. Sickle-cell anemia is a disease of the blood caused by a mutation in the hemoglobin protein. Hemoglobin, the oxygen-carrying protein of red blood cells (erythrocytes), is composed of four subunits, two α chains and two β chains. The sickle-cell mutation occurs in the β chain, and the mutant hemoglobin is called hemoglobin S. Normal hemoglobin is called hemoglobin A. Humans are diploid and thus contain two alleles of the β -chain gene. The two alleles are sometimes slightly different. About 50 genetic variants of hemoglobin are known, usually due to a single amino acid change, and most of these are quite rare. Although the effects on hemoglobin structure and function are often negligible, they can sometimes be extraordinary.

Sickle-cell anemia is a recessive genetic disease in which an individual inherits two copies of the β -chain allele for sickle-cell hemoglobin S (i.e., the sickle-cell allele). A heterozygous individual, with one sickle-cell allele and one normal allele, has nearly normal blood. Two heterozygous parents can potentially have a child who is homozygous for the recessive sickle-cell allele (Figure 1).

The nucleotide sequence of the sickle-cell allele usually contains a thymine (T) in place of an adenine (A), thereby changing one nucleotide triplet from GAG to GTG (Figure 2). This single base change results in hemoglobin S, which contains a

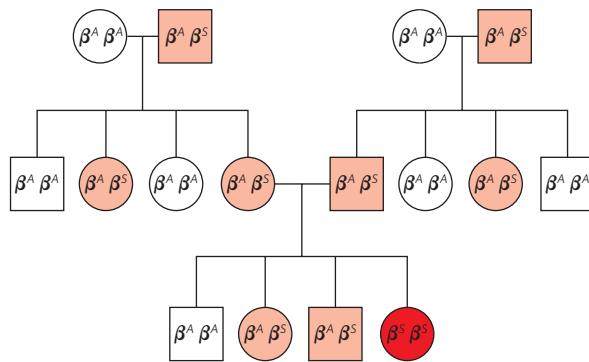


FIGURE 1 A family pedigree for sickle-cell anemia shows the genotypes for the hemoglobin β chain (circles, females; squares, males). The alleles are β^A for wild-type hemoglobin and β^S for sickle-cell hemoglobin S. Heterozygous individuals are shaded in pink, individuals homozygous for β^S in red.

Hemoglobin A

Triplet number	3	4	5	6	7	8	9
Gene sequence	...CTGACTCCTG	A GGAGAAGTCT...					
Amino acids	/ Leu Thr Pro	Glu Glu Lys Ser 					

Hemoglobin S

Triplet number	3	4	5	6	7	8	9
Gene sequence	...CTGACTCCTG	T GGAGAAGTCT...					
Amino acids	/ Leu Thr Pro	Val Glu Lys Ser 					

FIGURE 2 A single nucleotide change in the sickle-cell allele alters the hemoglobin β chain. In hemoglobin A, triplet 6 is GAG, which codes for glutamic acid (Glu). In hemoglobin S, the most common substitution is a T for the A in triplet 6 to form a GTG triplet, which codes for valine (Val).

the *HTT* protein is unknown, but the disease results in the degeneration of neurons in areas of the brain that affect motor coordination, memory, and cognitive function.

The number of triplet repeats in *HTT* can increase during gamete production, resulting in earlier onset and increased severity of the disease over successive generations. This is thought to occur by

template slippage (the same segment of DNA replicated more than once) during DNA synthesis due to the repetitive nature of the sequence. Other diseases caused by triplet expansion of this type have now been identified. These include Kennedy disease, spinocerebellar ataxia, and Machado-Joseph disease, all caused by an increase in CAG repeats. The CGG repeat is associated with fragile X syndrome, a

hydrophobic (water-fearing) valine residue at one position in the β chain, instead of a hydrophilic (water-loving) glutamic acid residue. This amino acid change causes deoxygenated hemoglobin S molecules to stick together, forming insoluble fibers inside erythrocytes and changing the shape of the cells. The blood of individuals with sickle-cell anemia contains many long, thin, crescent-shaped erythrocytes that look like the blade of a sickle (Figure 3). The sickle shape occurs only in veins, after the blood has become deoxygenated. Sickled cells are fragile and rupture easily, resulting in anemia (from the Greek for “lack of blood”).

When capillaries become blocked, the condition is much more serious. Capillary blockage causes pain and interferes with organ function—often the cause of early death. Without medical treatment, people with sickle-cell anemia usually die in childhood. Nevertheless, the sickle-cell allele is surprisingly common in certain parts of Africa. Investigation into the persistence of an allele that is so obviously deleterious in homozygous individuals led to the finding that in heterozygous individuals, the allele confers a small but significant resistance to lethal forms of malaria. Heterozygous individuals experience a milder condition called sickle-cell trait; only about 1% of their erythrocytes become sickled on deoxygenation. These individuals can live normal lives by avoiding vigorous exercise and other stresses on the circulatory system. Natural selection has thus resulted in an allele that balances the deleterious effects of the homozygous sickle-cell condition against the resistance to malaria conferred by the heterozygous condition.

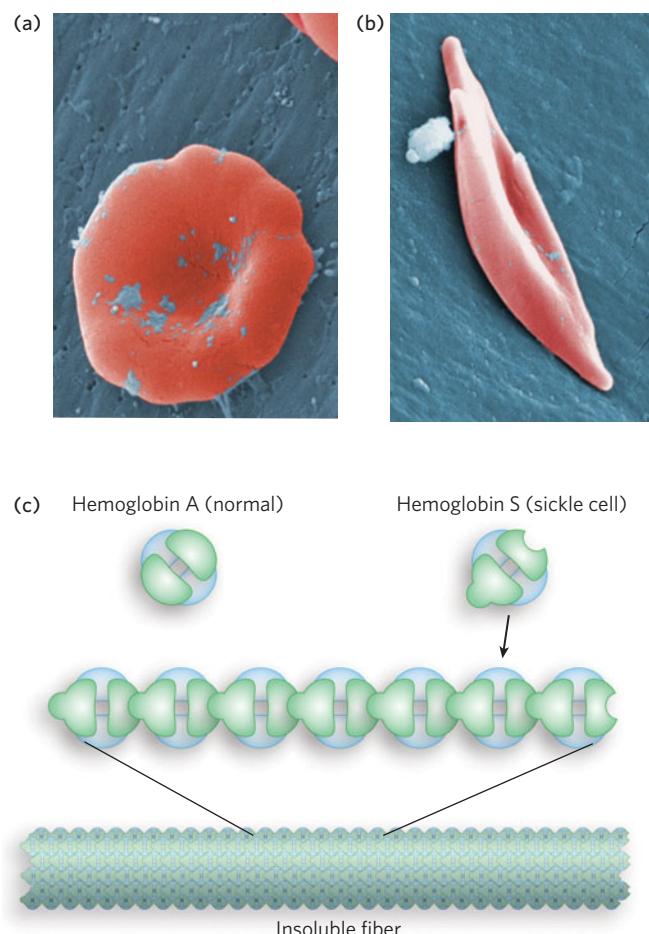


FIGURE 3 A comparison of (a) a normal uniform, cup-shaped erythrocyte with (b) a sickle-shaped erythrocyte seen in sickle-cell anemia. (c) The shape change in the hemoglobin molecule, due to the substitution of a Val for a Glu residue in the β chain, allows the molecules to aggregate into insoluble fibers within the erythrocytes. [Sources: (a) and (b) CDC/Sickle Cell Foundation of Georgia. Photos by Janice Haney Carr.]

neurological disorder; expansion of the CTG repeat is associated with myotonic dystrophy, a muscular wasting disease.

Cystic fibrosis is another genetic disease that has been identified at the molecular level. The gene (*CFTR*) is on chromosome 7 and encodes a chloride channel protein, the cystic fibrosis transmembrane conductance regulator (M_r 168,173). The protein contains five

domains: two domains that span the cytoplasmic membrane for chloride transport; two domains that bind and use ATP, the energy that fuels transport of the chloride ions; and a regulatory domain (Figure 2-29). The most common mutation (occurring in about 60% of cases) is *CFTRΔF508*, in which three nucleotides are deleted (denoted by the Δ), resulting in the deletion of phenylalanine (F) at position 508 in the amino acid

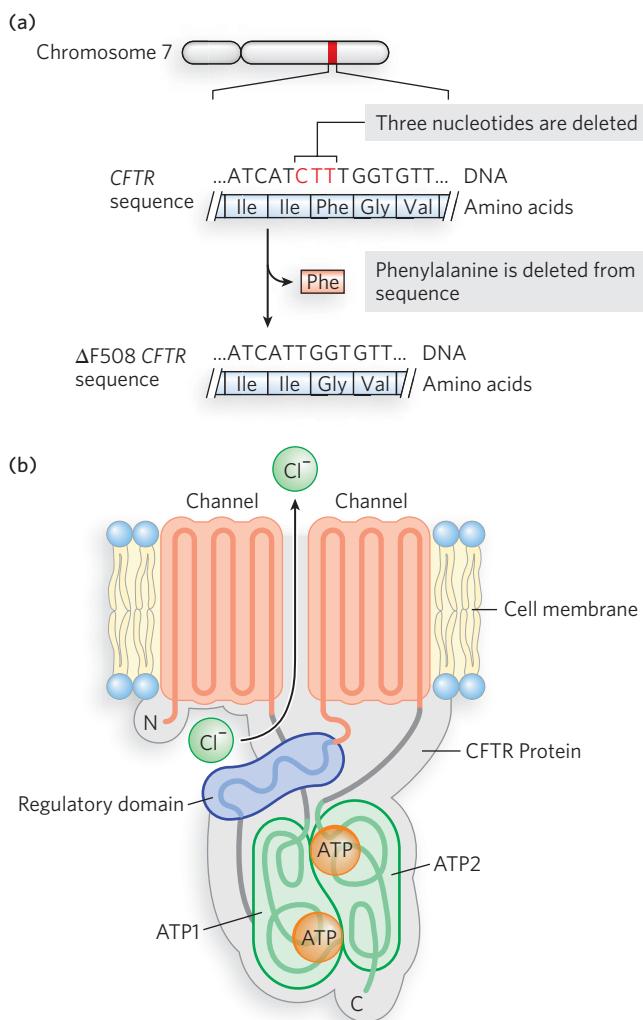


FIGURE 2-29 Cystic fibrosis. Cystic fibrosis is caused by a mutation that affects the function of a chloride ion channel. (a) The *CFTR* gene is on chromosome 7. It encodes a channel protein that transports chloride ions. The most common *CFTR* mutation leading to cystic fibrosis is a deletion of three nucleotides that results in the omission of phenylalanine (Phe) at position 508. The isoleucine (Ile) at position 507 remains the same, because both ATC and ATT code for an Ile residue. The omission of Phe⁵⁰⁸ prevents proper protein folding. (b) The chloride ion channel consists of five domains: two domains that form the channel across the cytoplasmic membrane, two domains that bind and use ATP as an energy source (ATP1 and ATP2), and a regulatory domain. Phe⁵⁰⁸ is in the ATP1 domain.

sequence. This residue is located in the first of the ATP-binding domains, and its deletion prevents proper folding of the protein. Many other mutations in *CFTR* have also been discovered. *CFTR* mutations are most prevalent in Caucasians from Northern Europe.

KEY CONVENTION

The molecular mass of a molecule is commonly expressed in one of two ways, and these are used interchangeably in this book. The first is molecular weight, also called relative molecular mass (M_r). The relative molecular mass of a molecule is the ratio of the mass of that molecule to one-twelfth the mass of carbon-12. Because it is a ratio, M_r is dimensionless and has no units. The second common method is molecular mass (m), which is the molar mass of the substance divided by Avogadro's number and is expressed in daltons (Da) or kilodaltons (kDa). One dalton is equal to one-twelfth the mass of carbon-12. For example, the molecular mass of a protein that is 1,000 times the mass of carbon-12 can be expressed by $M_r = 12,000$, $m = 12,000$ Da, or $m = 12$ kDa.

The $\Delta F508$ mutation in *CFTR* is autosomal recessive, and therefore an individual must inherit two copies of the mutant allele to develop cystic fibrosis, one from each parent. Without functional *CFTR* chloride channels, individuals with cystic fibrosis develop abnormally high sweat and mucus production, and a major complication is the buildup of mucus in the lungs. Patients experience breathing difficulties and often have pneumonia. Individuals with cystic fibrosis have typically had an average life span of about 30 years; however, as new treatments are developed, survival is increasing greatly.

Although many mutations are detrimental, other mutations can be beneficial. For example, the protein CCR5 is a coreceptor for HIV, the AIDS virus. There is much speculation about a 32 amino acid deletion mutation of CCR5 (due to the *CCR5Δ32* mutation), which is widely dispersed among people of European descent (an occurrence of 5% to 14% in these groups), although much rarer among Asians and Africans. Researchers speculate that this mutation may have conferred resistance to the bubonic plague or smallpox, thereby becoming enriched in the population, by natural selection, in endemic areas. Although the allele has a negative effect on T-cell (a type of immune cell) function, it seems to provide protection against HIV infection, as well as smallpox.

Another example of a mutation that confers some benefit is the one that causes sickle-cell anemia (Highlight 2-1), a mutation of hemoglobin. When the mutation is inherited from both parents, the result is misshapen red blood cells that can get stuck in capillaries and impede blood flow, with possibly fatal results.

However, people who are heterozygous for this mutation have enhanced resistance to malaria. Geographic areas where this mutation is prevalent in the population correlate with locations that are plagued by malaria.

Discovery of DNA as the hereditary material, and the understanding of how it is transcribed and translated into RNA and protein, is a most fascinating story in science. Darwin's theory of the origin of species through evolution by natural selection was compelling, but the mechanism that drove the variation on which natural selection could act remained a mystery in his lifetime. Yet, the key to understanding this mystery had already been discovered by Mendel. Mutations create the natural variation needed for the forces of natural selection to mold new species. It seems almost ludicrous that Mendel and Darwin were alive at the same time and separately uncovered secrets that together explained the diversity of planetary life. Lack of a robust means of communication kept these two vital pieces of information segregated for decades—an improbable situation today, given the rapid pace of global communication. Although most mutations are deleterious, the rare mutation that carries a beneficial change eventually enters the population through natural selection over the expanse of evolutionary time. Natural selection still drives change and the evolution of new species today.

SECTION 2.4 SUMMARY

- Nucleic acids (DNA and RNA) are composed of repeating units called nucleotides. Each nucleotide contains a phosphate group, a ribose sugar, and a nitrogenous base. Four different bases are found in DNA (adenine, guanine, cytosine, thymine). RNA also contains adenine (A), guanine (G), and cytosine (C), but uracil (U) instead of thymine (T). Information is encoded by the fact that G pairs specifically with C, and A pairs specifically with T (or U).
- Identification of DNA as the chemical of heredity was determined in experiments using virulent and nonvirulent bacteria. The DNA of virulent bacteria transforms nonvirulent bacteria into a virulent form.
- Even before the DNA structure was solved, studies of mutants drew the connection between genes and enzymes, as in the investigations of defective enzymes in the biosynthetic pathways of auxotrophic mutants of *Neurospora crassa*.
- Information flow in the direction DNA→RNA→protein is known as the central dogma. RNA is synthesized from a DNA template in the process of transcription. In translation, the RNA sequence is converted to protein. The duplication of DNA is replication. Exceptions to the central dogma exist (RNA→DNA, and RNA→RNA).
- Three types of RNA are required for DNA→RNA→protein. Ribosomal RNA combines with proteins to form ribosomes, which are factories for protein synthesis. Transfer RNAs are small adaptor RNAs to which amino acids become attached. Messenger RNAs encode proteins and are read by tRNAs in groups of three nucleotides, each of which specifies an amino acid.
- Functional RNAs are RNA sequences that are not translated into protein. Rather, the RNA sequences themselves perform functions in the cell. Both rRNA and tRNA are functional RNAs.
- Mutations are changes in DNA sequence. When a mutation affects the function of a protein or functional RNA, it results in a phenotypic change. Changes in the DNA sequence of germ-line cells underlie inherited human diseases, including hemophilia, Huntington disease, cystic fibrosis, and sickle-cell anemia. Mutations are not always deleterious—sometimes they can be beneficial and, indeed, are vital in creating the diversity needed for the evolution of new species.

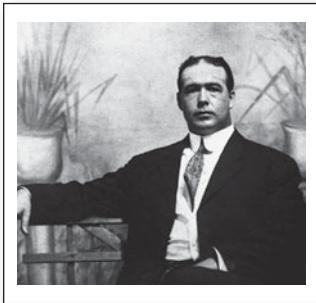
How We Know

Chromosome Pairs Segregate during Gamete Formation in a Way That Mirrors the Mendelian Behavior of Genes

Boveri, T. 1902. Ueber mehropologe Mitosen als Mittel zur Analyse des Zellkerns. Verh. Phys. Med. Ges. Wurzburg 35:67–90.

Sutton, W.S. 1902. On the morphology of the chromosome group in *Brachystola magna*. Biol. Bull. 4:24–39.

Sutton, W. S. 1903. The chromosomes in heredity. Biol. Bull. 4:231–251.



Walter Sutton, 1877-1916

[Source: University of Kansas Medical Center Archives.]

earlier. But now there were new ways of looking at organisms—namely, observing individual cells under the microscope. Sutton was particularly interested in the process of gamete production, in which one cell undergoes two divisions; in the second division, the chromosome number is halved relative to that of the parent. This process fascinated him. Others who studied these cell divisions used organisms with chromosomes that were too small to allow the observer to discern their individual identity. But Sutton studied the great lubber grasshopper, *Brachystola magna*, which had large chromosomes with distinctive shapes (Figure 1). This allowed him to see that, in meiosis, each chromosome paired with a look-alike partner (a homologous chromosome) and, during the second cell division, the two members of each pair assorted into different cells. On the union of sperm and egg, the homologous pairs were reestablished.

The behavior of chromosomes mimicked the Mendelian behavior of segregation of traits, but on a subcellular level. Sutton hypothesized that paternal and maternal chromosomes exist in pairs and separate into gametes during meiosis, explaining the diploid particles of heredity in Mendel's laws.

Today, a scientist making a groundbreaking discovery of this caliber would have established a solid reputation in science. But in Sutton's day, there were no graduate student stipends or regular sources of scientific funding. So Sutton became a physician and went back to his hometown in Kansas to practice medicine.

Theodor Boveri, a talented German scientist, conducted similar studies in the same year as Sutton. Boveri, too, was interested in meiosis and studied this process in sea urchin eggs. Although sea urchin chromosomes are small and cannot be observed individually, Boveri reached the same conclusions as Sutton, linking chromosomes with the particles of Mendelian inheritance. He also observed that eggs from which the nucleus was removed could be fertilized and then develop into normal—albeit haploid—larvae, and that normal larvae could develop from eggs with only the female set of chromosomes in the nucleus (also haploid). He concluded that each chromosome set, contributed by either parent, had a complete set of instructions for development of the organism.

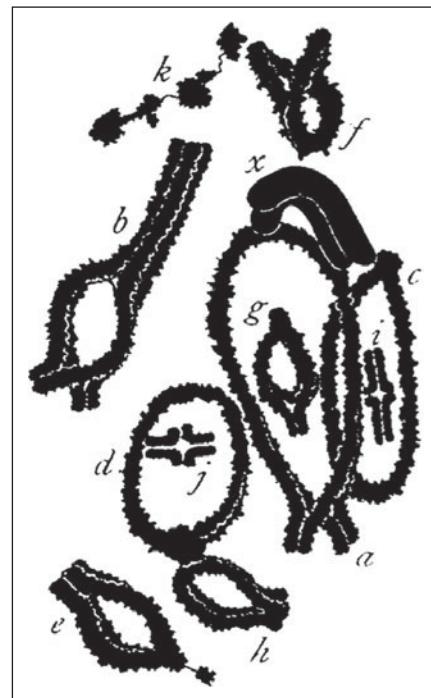


FIGURE 1 Chromosomes of the grasshopper *Brachystola magna* have unique shapes and sizes. The pairs of chromosomes are labeled *a* through *k* and *x*. [Source: W.S. Sutton, Biol. Bull. 4:24–39, 1902.]

Corn Crosses Uncover the Molecular Mechanism of Crossing Over

Creighton, H., and B. McClintock. 1931. A correlation of cytological and genetical crossing-over in *Zea mays*. *Proc. Natl. Acad. Sci. USA* 17:492-497.



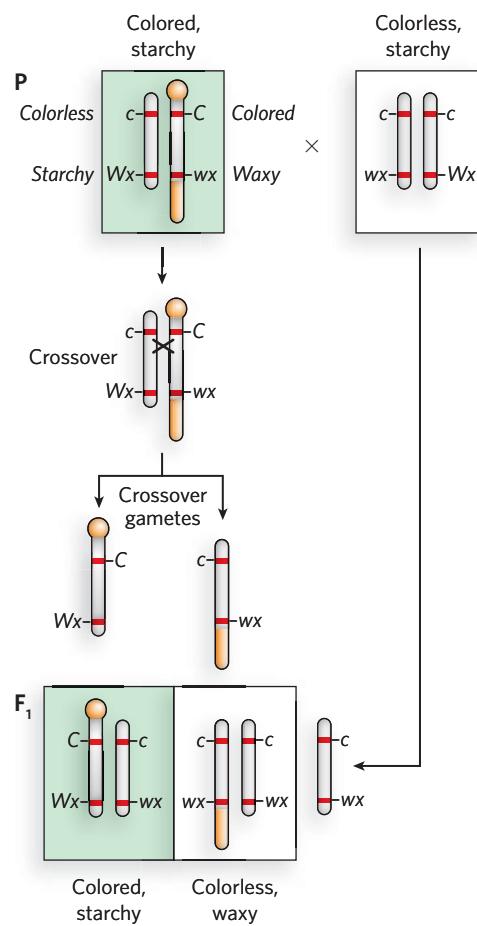
Barbara McClintock, 1902–1992 (left); Harriet Creighton, 1909–2004 (right) [Source: Cold Spring Harbor Laboratory.]

Fruit flies have taught us how our body plan is determined. Who would have guessed that fruit flies would teach us so much? Among the many fruitful (no pun intended) discoveries by Thomas Hunt Morgan, who developed the fly as a model for genetic study, was the finding that genes cross over between chromosomes. Although researchers presumed that genetic recombination occurred through material exchange between homologous chromosomes, there was no proof that this was indeed the case. Direct proof came in 1931 from a now classic study in corn (maize) by Harriet Creighton and Barbara McClintock.

The insightful experiments of Creighton and McClintock combined genetics and cytologic methods. To visualize crossing over between two homologous chromosomes, one first needs to find two homologous chromosomes that look different—no easy task. Creighton and McClintock searched until they found a plant with an odd-shaped chromosome; chromosome 9 had a knob on one end and an extension on the other. Next, they showed that this plant could be crossed with a plant having a normally shaped chromosome 9 to produce offspring having a homologous pair of chromosome 9's that did not look alike. They then mapped two alleles on chromosome 9 to follow recombination genetically. These alleles were seed color—C (colored) and c (colorless); and seed texture—Wx (starchy) and wx (waxy).

FIGURE 2 The gametes at the top left represent a corn plant with colored, starchy seeds that is heterozygous for these seed-color and seed-texture genes ($CcWwx$). One chromosome (chromosome 9) has abnormal extremities. This plant was crossed with a corn plant (top right) having colorless, starchy seeds ($ccWwx$). Genetic crossing over in the colored, starchy plant produced colorless, waxy progeny of genotype $ccwxw$. Microscopic examination confirmed that genetic crossing over involves physical recombination of chromosomes: one end of the abnormal chromosome 9 was replaced with a normal end, containing the colorless-seed gene.

Creighton and McClintock crossed the two plants represented at the top of **Figure 2** and looked for colorless, waxy progeny (i.e., progeny that produce colorless, waxy seeds). Genetic crossing over between the misshapen chromosome 9 and its homolog is required to produce a colorless, waxy plant of genotype $ccwxw$. If genetic crossing over results from physical recombination between the two chromosomes, then the chromosomes of the colorless, waxy progeny should contain chromosome pairs with the misshapen chromosome 9 having only one abnormality—either a knob or an extension at one end (see Figure 2). Indeed, chromosomes of the rare colorless, waxy offspring looked exactly as predicted, thereby confirming that genetic recombination occurs through the physical exchange of material between homologous chromosomes.



Hershey and Chase Settle the Matter: DNA Is the Genetic Material

Hershey, A.D. and M. Chase, 1952. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J. Gen. Physiol.* 36:39–56.



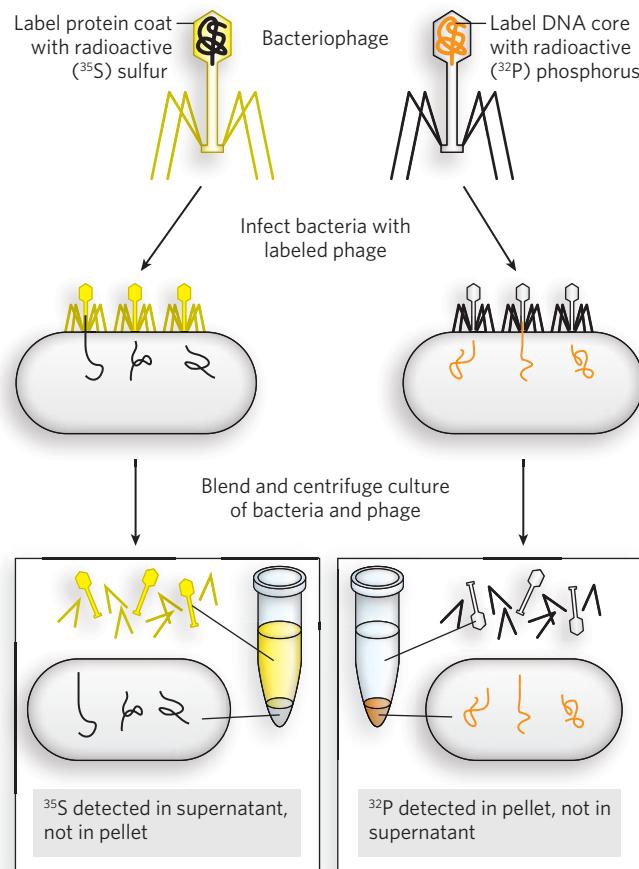
Martha Chase, 1927–2003
(left); Alfred Hershey,
1908–1997 (right) [Source:
Cold Spring Harbor Laboratory.]

In 1952, Martha Chase and Alfred Hershey performed a now classic experiment, the results of which would convince the world that DNA is the genetic material. They used a bacterial virus, mainly composed of protein and DNA, and set out to determine which of these components carries the hereditary material. Bacteriophage T2, or T2 phage, like other bacterial viruses, consists of a protein coat and a DNA core. Hershey and Chase took advantage of a key chemical difference between these two macromolecules. Using the fact that sulfur is found in proteins but not in DNA, and that phosphorus is found in DNA but not in proteins, they prepared radiolabeled T2 phage using either ^{35}S (only protein is radioactively labeled) or ^{32}P (only DNA is radioactively labeled). The two T2 phage samples were allowed to attach to their bacterial host, *Escherichia coli*, in two separate flasks (Figure 3). After infection, the bacteria were transferred to a blender and agitated to strip away any T2 phage material from the outside of the bacterial cell walls. Cells were collected by centrifugation, leaving unattached phage in the supernatant. The results were clear: ^{32}P -labeled DNA had transferred into the cells, while ^{35}S -labeled protein remained in the supernatant.

FIGURE 3 Bacterial cells infected with ^{32}P -labeled phage contained ^{32}P after blender treatment, indicating that viral ^{32}P -DNA had entered the cells. Cells infected with ^{35}S -labeled phage had no radioactivity after blender treatment. Progeny virus particles contained ^{32}P -DNA acquired from the cells infected with ^{32}P -labeled phage.

Therefore, it is the DNA that carries out the genetic program of the phage. In fact, the progeny phage produced in the infected cells contained ^{32}P and no ^{35}S , further proof that the DNA is the genetic material.

Although earlier experiments by Avery had suggested that DNA was the genetic material, the Hershey and Chase experiment finalized this important conclusion and inspired Watson and Crick in their quest to determine the structure of DNA. Hershey shared the 1969 Nobel Prize in Physiology or Medicine with Max Delbrück and Salvador E. Luria for their discoveries on the replication mechanism of viruses.



Key Terms

gamete cell, p. 24	law of segregation, p. 28	cytokinesis, p. 34
somatic cell, p. 24	law of independent assortment, p. 28	meiosis, p. 36
gene, p. 26	chromosome, p. 32	tetrad, p. 36
diploid, p. 26	G ₁ phase, p. 33	crossing over, p. 41
allele, p. 26	S phase, p. 33	deoxyribonucleic acid (DNA), p. 44
phenotype, p. 26	centromere, p. 33	ribonucleic acid (RNA), p. 44
haploid, p. 26	G ₂ phase, p. 33	ribosome, p. 48
genotype, p. 27	M phase, p. 33	RNA polymerase, p. 49
homozygous, p. 27	mitosis, p. 33	transcription, p. 49
heterozygous, p. 27		translation, p. 49

Problems

1. Two purebred pea plants are crossed. One strain has dominant round seeds; the other has recessive wrinkled seeds. (a) What phenotypes would be seen in the F₁ generation plants, and in what proportions? (b) What phenotypes would be seen in the F₂ generation plants, and in what proportions? (c) If an F₁ generation plant is crossed with a plant producing wrinkled seeds, what phenotypes are seen in the progeny, and in what proportions?
2. Two pea plants with round seeds are crossed. In the F₁ generation, all the plants have round seeds. What can you say about the genotype of the parental plants?
3. The F₁ plants from the cross in Problem 2 are next crossed at random. There are 129 plants in the F₂ generation. The majority, 121 plants, produce round seeds. However, there are 8 plants that produce wrinkled seeds. From this information, what were the genotypes of the original parental plants?
4. Purebred white-eyed male fruit flies are crossed with wild-type red-eyed females. If the progeny are crossed with each other repeatedly, which generation will be the first to contain white-eyed female flies?
5. Purebred wild-type male flies are crossed with purebred white-eyed female flies. If the progeny are crossed with each other repeatedly, which generation will be the first to contain white-eyed male flies?
6. A new species of fruit fly is found on an uncharted island. The flies are brightly colored, with blue and green bodies. After studying these insects for a year or two, researchers find one male with an all-black body. When this male is crossed with wild-type females, all of the male progeny in the F₁ generation are black, and all of the female progeny have the blue and green coloring. This same pattern (all black males and colored females) is repeated in the F₂, F₃, and F₄ generations. Explain these observations.
7. Two purebred flowering plants are crossed. One has red flowers and small leaves (RRLl) and the other has white flowers and large leaves (rrLL). Using a Punnett square analysis, and assuming that the genes are unlinked, predict the type and frequency of phenotypes in the F₂ generation.
8. If the F₁ plants in Problem 7 had genes for red flower color that exhibited incomplete dominance, the heterozygous Rr flower color would be pink. In that instance, what percentage of the F₁ plants in Problem 7 would have pink flowers? What percentage of the F₂ generation would have pink flowers?
9. Two purebred fruit flies are crossed. The male has white eyes and vestigial wings. The female has red eyes and normal wings. All F₁ flies have red eyes and vestigial wings. Using a Punnett square analysis, predict the percentage of F₂ generation males that will have red eyes and normal wings. Assume that the wing trait is not sex-linked.
10. A new and exotic species of fly is found, with green eyes (G) and striped wings (S). A mutant fly of the same species is found that has orange eyes (g) and clear (unstriped) wings (s). The mutant is cultured for many generations to obtain a purebred strain with the double-mutant phenotype. A ggss female fly is mated with a wild-type GGSS male. The F₁ progeny all have green eyes and striped wings, as expected. An F₁ male is mated with an F₁ female. Among the F₂ progeny of this cross, only two kinds of flies are observed: 75% with green eyes and striped wings, and 25% with orange eyes and clear wings. Some expected F₂ progeny (such as flies with green eyes and clear wings) are absent. Explain this result.
11. Both meiosis and mitosis are initiated with a complete replication of the cell's chromosomes in S phase. The replication of each chromosome produces a pair of sister chromatids. During the cell division immediately following replication, how are the chromosomes in the sister chromatid pairs distributed to daughter cells in mitosis and meiosis?
12. Two purebred plants, with genotypes AABBCCDDEEFF and aabbccddeeff, are crossed. In the F₁ generation, all individuals are heterozygous for all traits. Geneticists probe the linkage of these various genes by doing a series of crosses,

examining two traits in each cross. When all the crosses are done and the data are tabulated, the researchers find that in the F_1 plants, meiosis produces gametes that contain the following combinations of alleles at the indicated frequencies (which correspond to crossover frequencies):

$A + b$	9%
$A + e$	13%
$A + d$	50%
$B + c$	6%
$C + f$	50%
$C + e$	10%
$D + f$	16%

With these data, determine how the genes are distributed along the chromosomes. Draw a map, using the crossover frequencies as distances.

13. On one chromosome there are three linked genes designated M , N , and O . If crossing over occurs between M and O 5% of the time, and between N and O 8% of the time, what are the possible arrangements of the genes on the chromosome?
14. In the central dogma developed by Francis Crick and others, three kinds of RNA play important roles: rRNAs, tRNAs, and mRNAs. Explain two features that are characteristic for each type of RNA.
15. In the classic Hershey-Chase experiment (see How We Know), the T2 phage was labeled with either ^{35}S or ^{32}P prior to using it to infect a bacterial host. In this experiment, would it have been possible for these researchers to label one batch of T2 phage with both ^{35}S and ^{32}P and still get a compelling result? Why or why not? What would the results of the experiment be?

Additional Reading

Many of the books and papers listed here are available online, free of charge, from Electronic Scholarly Publishing, a collection of source material on the foundations of classical genetics (www.esp.org-foundations/genetics/classical).

General

Watson, J.D. 1968. *The Double Helix: A Personal Account of the Discovery of the Structure of DNA*. New York: Atheneum Publishers. Watson's original account of his adventures; a quick and very interesting read, warts and all.

Mendelian Genetics

Bateson, W. 1909. *Mendel's Principles of Heredity*. Cambridge: Cambridge University Press.

Mendel, G. 1866. Versuche über Pflanzen-Hybriden. In *Verhandlungen des naturforschenden Vereines (Proceedings of the Natural History Society)*. Brünn. Fewer than 150 copies were produced; Darwin owned one of them, but the evidence from examining his copy indicates that he did not open it to Mendel's work.

Cytogenetics: Chromosome Movements during Mitosis and Meiosis

Hooke, R. 1664. *Micrographia: Some Physiological Descriptions of Minute Bodies Made by Magnifying Glasses with Observations and Inquiries Thereupon*. London (available at www.gutenberg.org/etext/15491).

Sutton, W.S. 1903. The chromosomes in heredity. *Biol. Bull.* 4:231–251. An outline of the rationale behind Sutton's

proposal that chromosomes carry the material of heredity.

The Chromosome Theory of Inheritance

Creighton, H.B., and B. McClintock. 1931. A correlation of cytological and genetical crossing-over in *Zea mays*. *Proc. Natl. Acad. Sci. USA* 17:492–497.

Morgan, T.H., A.H. Sturtevant, H.J. Muller, and C.B. Bridges. 1915. *Mechanism of Mendelian Heredity*. New York: Henry Holt and Company. This book contains the classic work from Morgan's lab, and illustrates the way in which scientific discoveries were published before scientific journals became the norm.

Molecular Genetics

Avery, O.T., C.M. MacLeod, and M. McCarty. 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *J. Exp. Med.* 79:137–158.

Beadle, G.W., and E.L. Tatum. 1941. Genetic control of biochemical reactions in *Neurospora*. *Proc. Natl. Acad. Sci. USA* 27:499–506.

Crick, F. 1970. Central dogma of molecular biology. *Nature* 227:561–563. The classic paper in which Crick proposes the central dogma of information flow in biology.

Hershey, A.D. and M. Chase. 1952. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J. Gen. Physiol.* 36:39–56.

Chemical Basis of Information Molecules



Judith Klinman [Source: Courtesy of Judith Klinman]

banned energy state, might in fact contribute to biological reaction mechanisms. However, tunneling was thought to occur primarily at temperatures near absolute zero (0 K), whereas biological reactions take place at ~ 300 K!

To investigate a possible role for tunneling in biology, we measured changes in reaction rates catalyzed by the enzyme alcohol dehydrogenase that occurred when hydrogen atoms (^1H) in the reaction substrate were replaced with deuterium (^2H) or tritium (^3H). The resulting kinetic isotope effects revealed significant hydrogen atom tunneling, a result we published in *Science* in 1989 [see How We Know]. But then things got even more interesting. Some enzyme reactions were exhibiting properties that required a new conceptual framework, in addition to tunneling, to describe their behavior. This incredibly exciting discovery focused attention away from static enzyme structures seen by x-ray crystallography, and provided a direct link between protein dynamics and the making and breaking of chemical bonds at enzyme active sites. Hydrogen tunneling depends on the exact geometry between a hydrogen bond donor and acceptor, often requiring structural adjustments in the enzyme subsequent to substrate binding. As a result, our discovery led beyond an initial correction to classical enzyme theory to new theories for the ways that hydrogen atoms move within an enzyme active site. I wish I had made this discovery about enzyme function earlier in my career because there are so many avenues to now explore; much of this exciting research will be carried out not by me but by current and future students.

—**Judith Klinman**, on her discovery of hydrogen tunneling in enzyme-catalyzed reactions

- 3.1 Chemical Building Blocks of Nucleic Acids and Proteins** 62
- 3.2 Chemical Bonds** 68
- 3.3 Weak Chemical Interactions** 73
- 3.4 Stereochemistry** 78
- 3.5 The Role of pH and Ionization** 81
- 3.6 Chemical Reactions in Biology** 84

Molecular biology involves the study of molecules that store and process genetic information. The chemical properties of these molecules and the principles that govern their behavior are central to understanding the maintenance and transfer of that information. Key to the storage and use of genetically encoded information are nucleic acids and proteins, macromolecules that are major constituents of all cells. These high-molecular-weight polymers are assembled from relatively simple precursors and can form three-dimensional structures that mediate a wide variety of biological activities.

The functions of nucleic acids and proteins stem from their chemical properties. Shape, electrical charge distribution, propensity to form weak or strong chemical bonds, and preference for hydrophobic (water-excluding) or hydrophilic (water-including) interactions—all of these contribute to the ability of DNA to function as the primary repository of genetic information and the ability of proteins to enhance biochemical reaction rates. Proteins also play important structural roles in cells, and they enable cells to communicate and respond to their environment. RNA, a chemical cousin of DNA, shares the information-bearing properties of DNA and also has some structural and functional similarities to proteins. As we shall see throughout this book, new biological activities and functions for RNA continue to be discovered. RNA is an important and, until recently, underappreciated controller of gene expression in all cells.

The flow of biological information in cells and organisms makes sense only in the context of the underlying chemical behavior of these biomolecules. Although molecular biologists often say that a protein “recognizes” a fragment of DNA, or that an RNA molecule “binds” to a protein, or that several proteins “assemble” into a multisubunit complex, what do these phrases really mean? A more quantitative framework for understanding cellular function requires a familiarity with the chemical principles by which molecules fold, react, and interact.

In this chapter we discuss these chemical principles, many of which developed from concepts originally drawn from the study of small molecules. We begin with the chemical building blocks of nucleic acids and proteins and a discussion of the kinds of chemical bonds that hold them together. We then discuss constraints on the behavior of biomolecules stemming from their stereochemistry, their ionization properties within the cellular environment, and their propensities to react and interact with other large and small molecules. Understanding the chemistry that governs protein and nucleic acid function provides a strong

foundation for exploring the many facets of biological behavior described throughout this book.

3.1 Chemical Building Blocks of Nucleic Acids and Proteins

We start by focusing on the underlying chemical properties that control the behavior of nucleic acids and proteins. Nucleic acid structures are discussed in more depth in Chapter 6 and amino acids in Chapter 4.

Nucleic Acids Are Long Chains of Nucleotides

Deoxyribonucleic acid (DNA) and **ribonucleic acid (RNA)** store and transmit genetic information, in part by coding for proteins. In addition, some RNA molecules function catalytically or structurally within larger, multimolecular complexes. Both DNA and RNA are composed of building blocks (monomers) called **nucleotides**, which are linked together in long, unbranched chains. Nucleic acids can reach chain lengths of up to many millions of nucleotides and molecular masses of up to several billion daltons. A nucleotide molecule has three components: a nitrogenous base, a five-carbon (pentose) sugar, and a phosphate group. A sugar and base without the phosphate group is referred to as a **nucleoside**.

The nucleotides that make up DNA polymers are **deoxyribonucleotides** (Figure 3-1a), named for the type of pentose sugar found in DNA: deoxyribose. Whereas the phosphate group and the type of pentose remain constant, each deoxyribonucleotide contains one of four different nitrogenous bases: **adenine (A)**, **cytosine (C)**, **guanine (G)**, or **thymine (T)** (Figure 3-1b). The type of base establishes the identity of individual deoxyribonucleotides. Thus, the information in DNA is written in a four-letter alphabet.

Chemically, RNA is very similar to DNA. Like DNA, it is a long, unbranched polymer of nucleotides. And like DNA nucleotides, all RNA nucleotides contain the same pentose and a phosphate group, and one of four different nitrogenous bases. Two small differences in their chemical components, however, give rise to important distinctions between the structures and functions of RNA and DNA. The first is the type of pentose present. RNA nucleotides contain ribose, and thus are named **ribonucleotides** (see Figure 3-1a). Ribose has one more hydroxyl ($-OH$) group on the sugar ring than does deoxyribose, which defines the RNA polynucleotide as *ribonucleic acid* rather than *deoxyribonucleic acid*. The second distinction is the

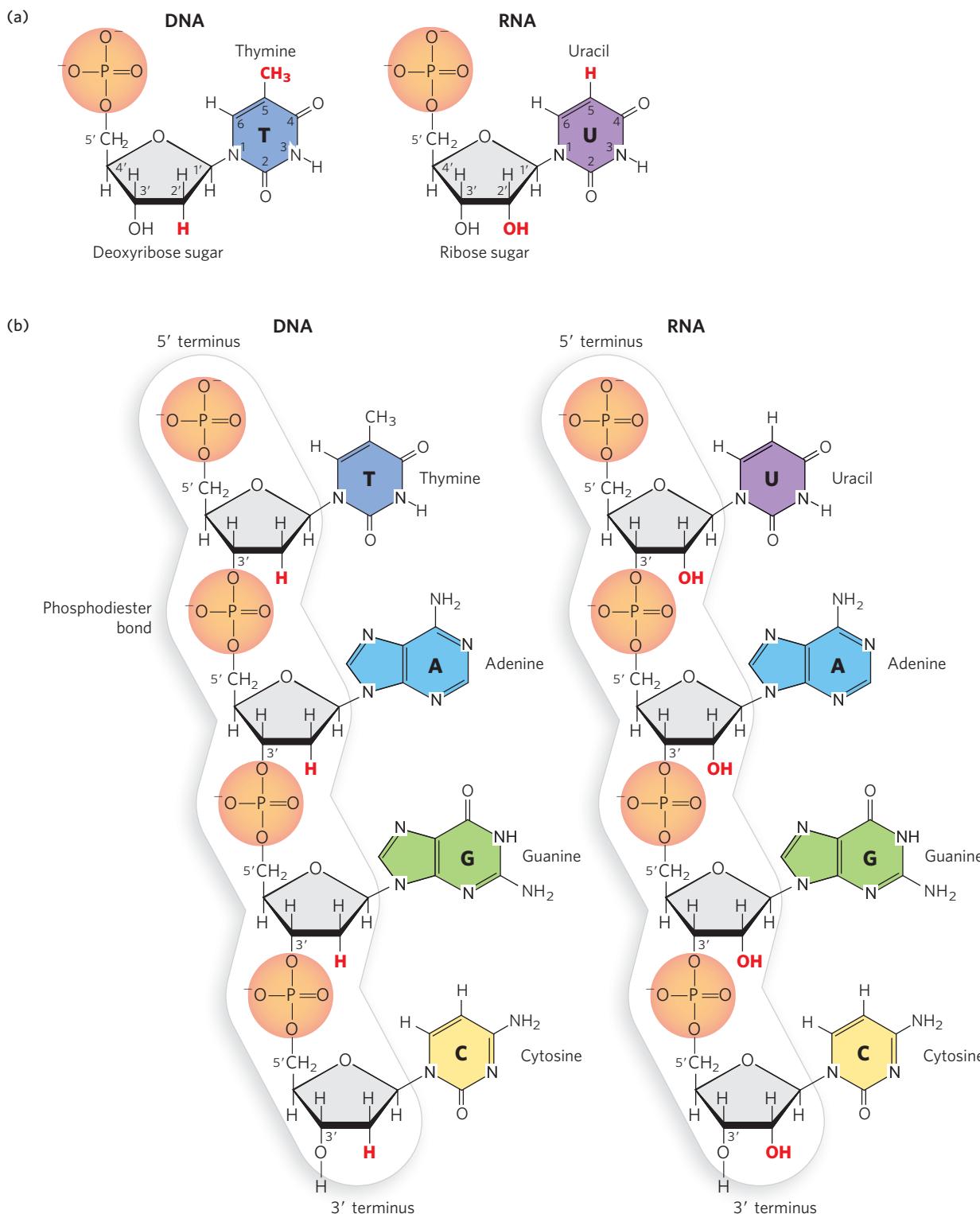


FIGURE 3-1 Chemical building blocks of DNA and RNA.

(a) Chemical differences between the nucleotides of DNA and RNA. Atoms in the base ring are numbered sequentially, starting with the nitrogen that is bound to the sugar; carbon numbering in the sugar ring is denoted by a prime (''). DNA nucleotides contain the sugar deoxyribose, whereas RNA contains ribose (gray). The difference between these two sugars is a single hydroxyl group on the 2' carbon. In addition,

DNA contains thymine, rather than the uracil found in RNA; these bases differ by a methyl group on carbon 5 of the base (red). (b) Segments of deoxyribonucleotide (DNA) and ribonucleotide (RNA) chains. Phosphodiester bonds connect the individual nucleotide units; the sugar-phosphate backbone (outlined) is polar and directional, with free 5' and 3' termini. The 2' hydrogens of DNA and the 2'-hydroxyl groups of RNA are highlighted in red.

Table 3-1 Nucleic Acid Nomenclature

Base	Nucleoside	Nucleotide	Abbreviation		Nucleic Acid
			One-letter	Three-letter	
Adenine	Adenosine	Adenylate	A	AMP	RNA
	Deoxyadenosine	Deoxyadenylate	dA	dAMP	DNA
Guanine	Guanosine	Guanylate	G	GMP	RNA
	Deoxyguanosine	Deoxyguanylate	dG	dGMP	DNA
Cytosine	Cytidine	Cytidylate	C	CMP	RNA
	Deoxycytidine	Deoxycytidylate	dC	dCMP	DNA
Thymine	Thymidine or deoxythymidine	Thymidylate or deoxythymidylate	T or dT	TMP or dTMP	DNA
Uracil	Uridine	Uridylate	U	UMP	RNA

Note: Nucleoside and nucleotide are generic terms that include both ribo- and deoxyribo-forms. Also, ribonucleosides and ribonucleotides are here designated simply as nucleosides and nucleotides (e.g., riboadenosine as adenosine), and deoxyribonucleosides and deoxyribonucleotides as deoxynucleosides and deoxynucleotides (e.g., deoxyriboadenosine as deoxyadenosine). Both forms of naming are acceptable, but the shortened names are more commonly used. Thymine is an exception; "ribothymidine" is used to describe its unusual occurrence in RNA.

assortment of nitrogenous bases found in RNA. Ribonucleotides contain three of the same bases found in DNA—adenine, cytosine, and guanine—but instead of thymine, the fourth base in RNA is **uracil (U)** (see Figure 3-1b). Uracil is structurally identical to thymine except for the absence of the methyl ($-\text{CH}_3$) group. The nucleotides of DNA and RNA are represented by both three-letter and one-letter abbreviations (Table 3-1).

KEY CONVENTION

DNA and RNA are defined by the type of sugar in the polynucleotide backbone (deoxyribose or ribose), not by the presence of thymine or uracil.

Even with just four types of nucleotides each, the number of possible DNA and RNA sequences (4^n , where n is the number of nucleotides in the sequence) is enormous for even the shortest molecules. Thus, an almost infinite number of distinct genetic messages can exist.

Proteins Are Long Polymers of Amino Acids

Proteins, like nucleic acids, are unbranched polymers. The building blocks of protein chains are **amino acids** (Figure 3-2a). When two or more amino acids are joined together, a peptide is formed. Longer chains of amino acids are called **polypeptides** (Figure 3-2b). Once incorporated into a polypeptide chain, the individual amino acids of the chain are referred to as amino acid residues (see Chapter 4). A functional

protein may be formed from one polypeptide chain or several interacting polypeptides. Proteins are abundant in all cells. They perform many functions, including catalyzing biochemical reactions, serving structural roles, receiving and transmitting chemical signals within and among cells, and transporting specific ions and molecules across cellular membranes. The bulk of all proteins found in cells and viruses are composed of just 20 different amino acids.

All 20 common amino acids have a similar structure: a central carbon atom, the **alpha carbon atom (α carbon, or C_α)**, bonded to four different chemical groups. For this reason, they are called α-amino acids. The α-amino acids have a carboxyl ($-\text{COOH}$) group, an amino ($-\text{NH}_2$) group, and a hydrogen atom, all bonded to the α carbon. Each amino acid also has a unique side chain, or **R group**, bonded to the α-carbon atom (see Figure 3-2). The R groups vary in structure, size, electrical charge, and hydrophobicity. The diverse chemical properties of R groups are what give proteins the ability to form many different three-dimensional structures and to perform many different kinds of activities in biological systems (see Figure 4-3). The 20 common amino acids are represented by both three-letter and one-letter abbreviations, which are used to indicate the composition and amino acid sequence of proteins (see Table 4-1). There are also many, less-common amino acids, found both in proteins and as cellular constituents not incorporated into proteins. Note that with 20 different amino acid building blocks, the number of possible protein

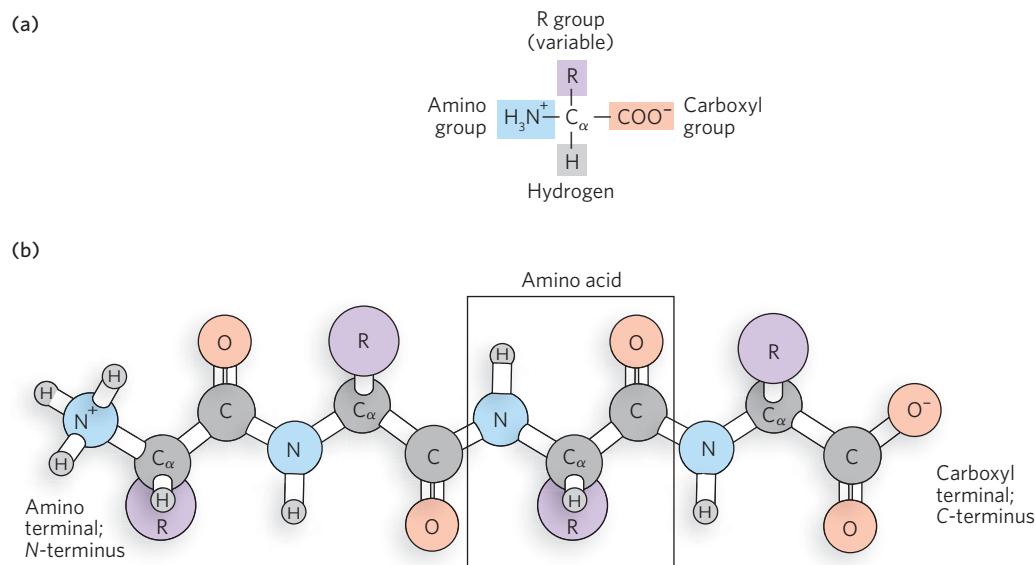


FIGURE 3-2 Chemical building blocks of proteins. (a) The structure of an amino acid. The central carbon atom (C_α) bonds to an amino group (blue), a carboxyl group (pink), a hydrogen, and a side chain (R, purple). The amino and carboxyl groups are in the ionized forms found in solution at physiological pH. (b) A segment of a polypeptide chain. Note

that polypeptide chains have directionality, with a free amino group at one end (the amino terminus, or N-terminus) and a free carboxyl group at the other (carboxyl terminus, or C-terminus). Peptide bonds connect the amino acid residues in the polypeptide chain.

sequences (20^n , where n is the number of amino acids in the sequence) is vast!

Chemical Composition Helps Determine Nucleic Acid and Protein Structure

The fact that some of the crucial requirements for life are met by polymeric molecules makes good sense, from a biosynthetic standpoint. As we have noted, a huge variety of nucleic acids and proteins can be produced by varying the sequence of nucleotide or amino acid monomers in the chains. DNA molecules are typically many millions of nucleotides long, but they form relatively uniform overall structures in which the nucleotide bases in two strands pair up along their length to produce a double helix (Figure 3-3a). RNA molecules, except for those that store the genetic information of viruses, are much shorter and more structurally diverse than DNA. A single strand of RNA can fold back on itself to form short helices that come together in a three-dimensional shape (Figure 3-3b). These differences between DNA and RNA structure stem from the role of RNA's 2'-hydroxyl groups in altering the shape and chemical properties of the sugar-phosphate backbone (see Chapter 6).

Of all biological polymers, proteins have the greatest variety of three-dimensional structures and

range of functional groups, resulting from the different types of amino acid side chains. This variety underlies the role of proteins as the primary catalysts of chemical reactions (Figure 3-4). Of course, proteins also perform many other, noncatalytic cellular functions, made possible by the chemical diversity of their amino acid building blocks.

Chemical Composition Can Be Altered by Postsynthetic Changes

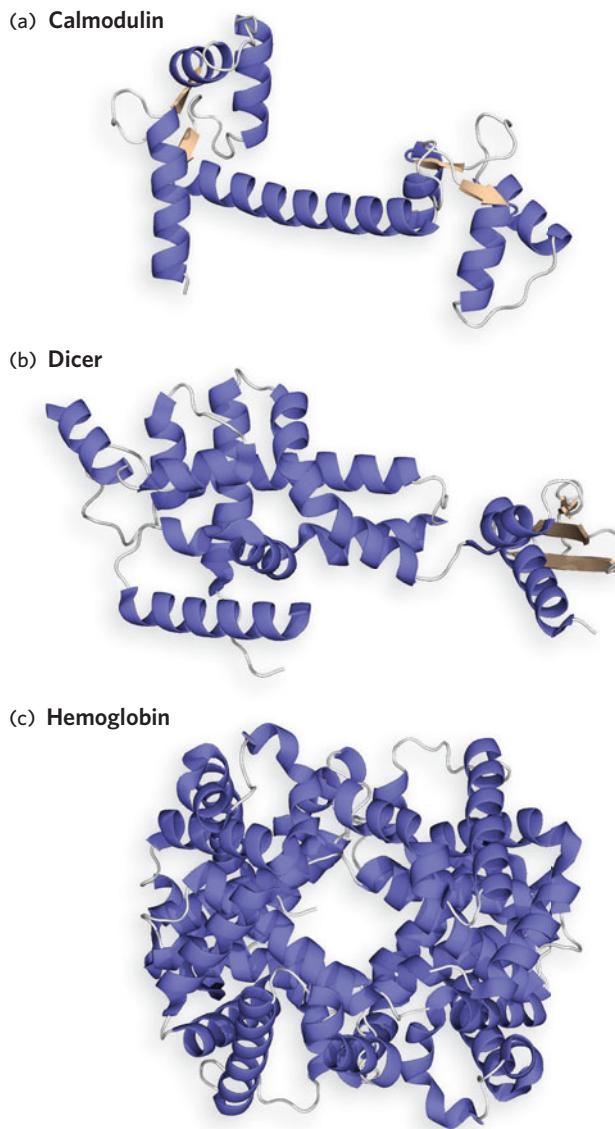
Chemical modifications of nucleotides and amino acids often occur after a DNA, RNA, or protein molecule has been synthesized. Sometimes these modifications are required for the molecule to attain its biologically active structure or to bind other molecules.

The primary modification of DNA nucleotides is the addition of methyl ($-\text{CH}_3$) groups to the C, A, and G bases (Figure 3-5a). DNA base methylation is critical for accurate DNA replication and, in bacteria, for the protection of DNA from degradative enzymes; and in human and other eukaryotic cells, it is essential for activating and silencing gene expression. RNA molecules can be modified in a greater variety of ways, including the addition of methyl groups to the nucleotide bases or the 2'-hydroxyl group of the ribose

**FIGURE 3-3** The helical structure of DNA and RNA.

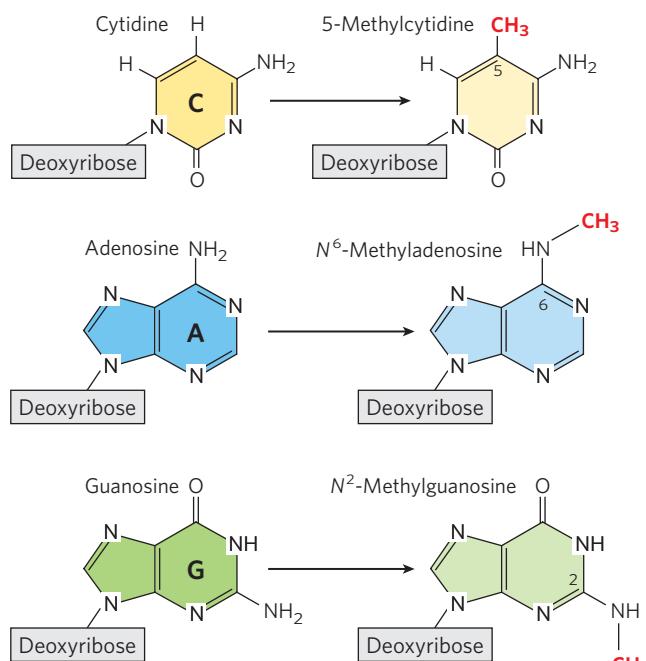
(a) Ribbon model of a DNA double helix, consisting of two strands of DNA. Base pairs in the helix twist around the central axis. (b) Ribbon models of three RNA molecules, each consisting of a single strand of RNA: phenylalanine-tRNA from yeast, a self-cleaving RNA from the hepatitis delta virus (HDV), and a self-splicing intron from *Tetrahymena*. Each RNA includes short stretches of helical structure that fold into a three-dimensional shape. The color coding of bases is as follows: cytosine, yellow; adenine, light blue; guanine, green; thymine, dark blue; uracil, purple. [Sources: (b) PDB ID 1TRA (top), PDB ID 1DRZ (middle), PDB ID 1U6B (bottom).]

and the substitution of less-common bases for the usual A, C, G, or U (Figure 3-5b). Such chemical changes affect the ability of RNA molecules to fold into their correct three-dimensional structure and to interact with proteins.

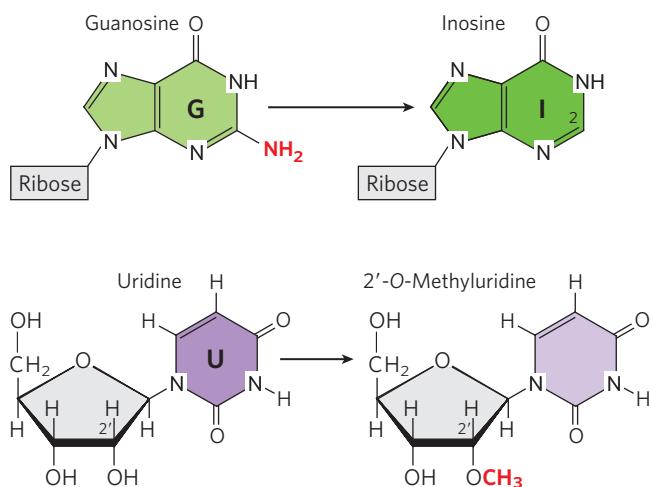
**FIGURE 3-4** Examples of protein structures. Proteins can form a wide range of three-dimensional structures. Shown here are (a) calmodulin, a Ca^{2+} -binding protein; (b) Dicer, an enzyme that cleaves double-stranded RNA; and (c) hemoglobin, the oxygen carrier in red blood cells. See Figure 4-10 for an explanation of how molecular structures of proteins are represented. [Sources: (a) PDB ID 1CLL. (b) PDB ID 3C4B. (c) PDB ID 1HGA.]

Proteins are often modified by the addition of phosphate groups to hydroxyl groups in the side chains of amino acids such as serine and tyrosine, and this can dramatically change a protein's shape and function. The phosphorylation and dephosphorylation of proteins is an important mechanism by which signals are transmitted within and among cells. Proteins are also sometimes modified by the addition of methyl groups and sugars (glycosylation), with functional consequences (Figure 3-6). For example, glycosylated proteins

(a) DNA modifications

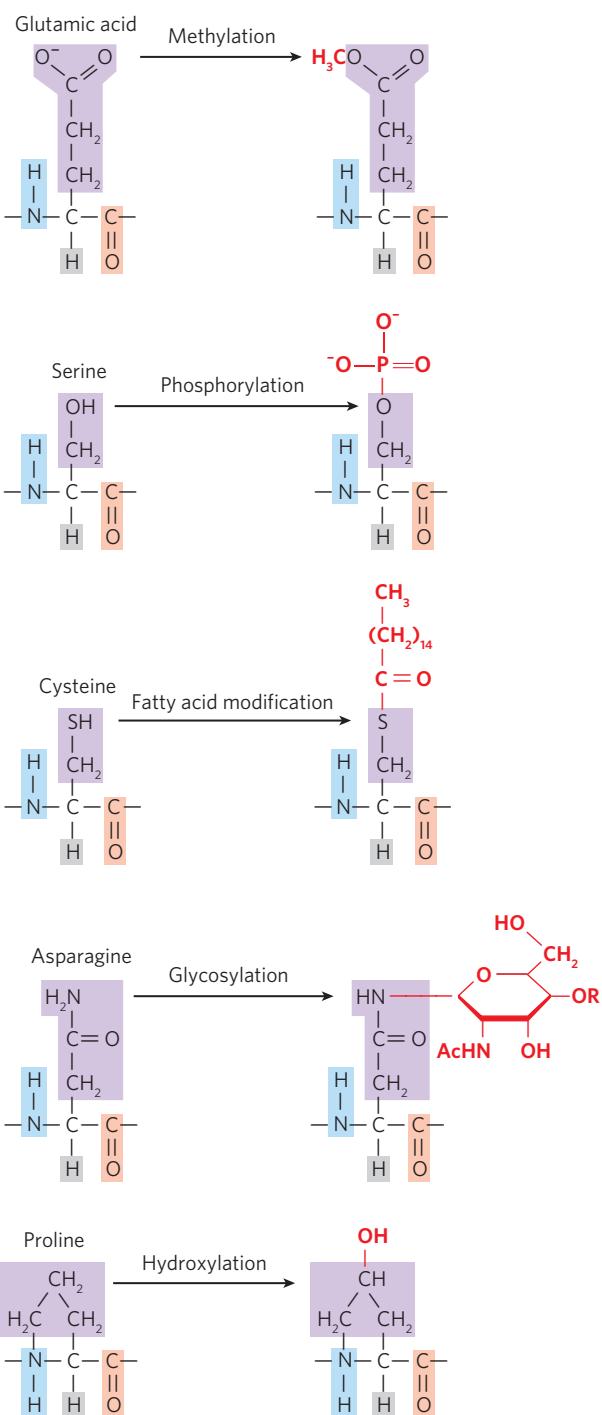


(b) RNA modifications

**FIGURE 3-5** Chemical modification of nucleotide bases.

(a) Examples of methylation modifications in DNA nucleotides. The extra methyl group on 5-methylcytidine, N⁶-methyladenosine, and N²-methylguanosine is highlighted in red. (b) Examples of modifications in RNA nucleotides. Loss of the exocyclic amino group from the guanine ring gives rise to inosine; uridine can be methylated at the 2' position to produce 2'-O-methyluridine. Modification sites are indicated in red.

provide chemical signatures on the surfaces of cells that help distinguish “self” from “nonself.” All these chemical modifications can substantially change the behavior of proteins, as we discuss in Chapters 5, 18, and 19.

**FIGURE 3-6** Chemical modification of some amino acids.

Glutamic acid can be methylated on its side-chain carboxyl group; the –OH group of serine is a frequent site of phosphorylation, as are those of tyrosine and threonine (not shown); the –SH group of cysteine can be modified with a lipid or fatty acid chain; asparagine can be modified with N-acetylglucosamine or glucosamine chains; and proline is sometimes hydroxylated. In each case, the modification alters the behavior of the protein containing the changed amino acid residue. Modification sites are indicated in red.

SECTION 3.1 SUMMARY

- Polymeric molecules play crucial roles in all organisms.
- The nucleic acids, DNA and RNA, are polymers of nucleotides. Each nucleotide has three components: a deoxyribose (in DNA) or ribose (in RNA) pentose sugar, a phosphate group, and a nitrogenous base. The four bases in DNA are adenine, guanine, cytosine, and thymine; the four bases in RNA are adenine, guanine, cytosine, and uracil.
- DNA and RNA are chemically similar, with two small differences. The ribose of RNA has a hydroxyl ($-OH$) group on the 2' carbon of the sugar ring, but the deoxyribose of DNA does not; and instead of thymine, RNA nucleotides contain uracil, an unmethylated form of the thymine base. The nucleotides of DNA or RNA are linked by phosphodiester bonds.
- Proteins are polymers of amino acids. Twenty amino acid building blocks are commonly found in proteins, each consisting of a central α carbon atom bonded to four different groups: a carboxyl group, an amino group, an R group, and a hydrogen atom. The R groups, or side chains, have chemical properties that contribute to the functional and structural diversity of proteins. The amino acid residues in proteins are linked by peptide bonds.
- DNA molecules form a two-stranded double helix, whereas RNA structures consist of a single polynucleotide strand that folds back on itself to create various three-dimensional shapes. Protein structures are even more diverse, due in part to the different chemical properties of the amino acid side chains.
- Postsynthetic chemical modifications of DNA, RNA, and proteins can dramatically affect the structure and biological activity of these macromolecules. In DNA, methylation of bases A, C, and G is common and leads to changes in gene expression. In RNA, modifications are more varied and include methylation of bases and/or ribose and other, more substantial alterations of bases. Protein modifications include the addition of phosphate, lipid, sugar, methyl, and hydroxyl groups to specific amino acid side chains.

3.2 Chemical Bonds

All molecules, whether table salt (sodium chloride) or a segment of DNA, are atomic aggregates in which the atoms are held together by attractive forces known as **chemical bonds**. Some chemical bonds are strong and

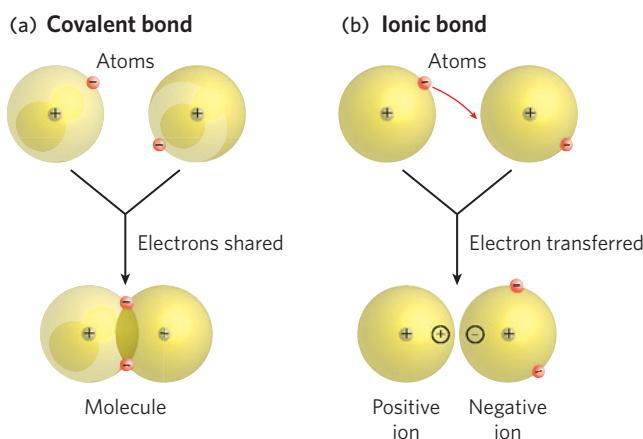


FIGURE 3-7 Covalent bonds and ionic bonds compared.

(a) A covalent bond forms when two atoms share electrons so that their outer electron shells are filled. (b) An ionic bond forms when one or more electrons are completely transferred from one atom to another, such that one atom bears a formal positive charge, and the other a negative charge. Note the space between the atoms paired in an ionic bond; they are not as close together as atoms joined by a covalent bond.

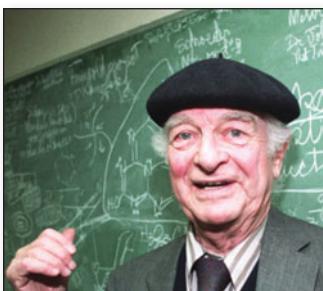
can hold two atoms together indefinitely, but others are relatively weak and transient. In this section we discuss strong chemical bonds—covalent bonds and ionic bonds. Section 3.3 focuses on weaker chemical bonds and interatomic interactions.

Electrons Are Shared in Covalent Bonds and Transferred in Ionic Bonds

As first described by Gilbert Lewis, a **covalent bond** is formed when two atoms share a pair of electrons between their positively charged nuclei (Figure 3-7a). Atoms joined by a covalent bond tend to share electrons such that their outer electron shells are filled. Another type of chemical bond was discovered through Richard Abegg's work on the chemical stability of noble gases. In contrast to covalent bonds, **ionic bonds** involve the complete transfer of one or more electrons from one atom to another. When electrons are transferred, the atoms are converted into ions—one having a positive charge and the



Gilbert Newton Lewis,
1875–1946 [Source: Lawrence Berkeley National Laboratory.]



Linus Pauling, 1901-1994

[Source: © Bettmann/Corbis.]

chemical bonds lie somewhere in between. To determine whether two atoms in a molecule are more likely to form a covalent or an ionic bond, Pauling introduced the concept of **electronegativity**, the propensity of an atom within a molecule to attract electrons to itself. Unequal electron sharing reflects different affinities of the bonded atoms for electrons. Atoms with a tendency to gain electrons are referred to as **electronegative atoms**, those with a propensity to lose electrons as **electropositive atoms**. In general, as atomic radius decreases, electronegativity increases and the atom has a greater likelihood of forming an ionic rather than a covalent bond. Ionic bonds often form between a metal and a nonmetal; the metal atom donates one or more electrons to the nonmetal atom to form a salt.

such as sodium chloride. The difference in electronegativities of two atoms can be used to predict the type of bonding between them ([Figure 3-8](#)). When this difference is zero or very small, the bond is purely covalent; when the difference is greater than zero but less than 1.67, the bond is considered to be **polar covalent**, meaning that the electrons are shared between the atoms but biased toward one “pole” of the two-atom bond. A difference in electronegativities of greater than 1.67 gives rise to an ionic bond.

Although typically weaker than covalent bonds, ionic bonds do not restrict the relative orientations of the bonded atoms; thus, they are very useful in macromolecules. For example, ionic bonds—also called salt bridges—can form between pairs of oppositely charged amino acid side chains, such as arginine and glutamic acid (glutamate), to stabilize protein structure (Figure 3-9). In highly charged molecules such as nucleic acids, metal ions form ionic bonds that help stabilize three-dimensional structure. RNA molecules require ionic bonding with magnesium ions to form complex three-dimensional structures, which involve close packing of the negatively charged sugar-phosphate backbone (Figure 3-10).

The strength of an ionic bond can vary with the salt concentration and hydrophobicity of the environment. Some ionic bonds are very strong indeed.

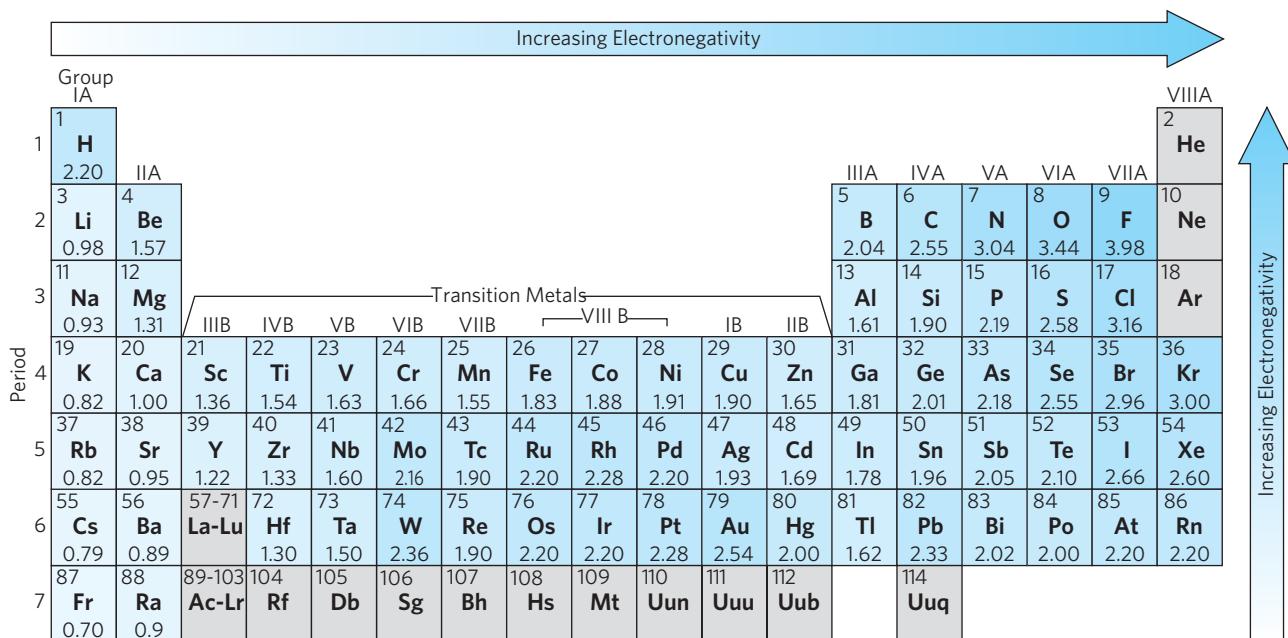


FIGURE 3-8 A periodic table of electronegativity, using the Pauling scale. The electronegativity of an atom is affected by both its atomic weight and the distance of its outer electrons.

from the positively charged nucleus. Electronegativity is not strictly an atomic property, but rather a property of an atom within a molecule.

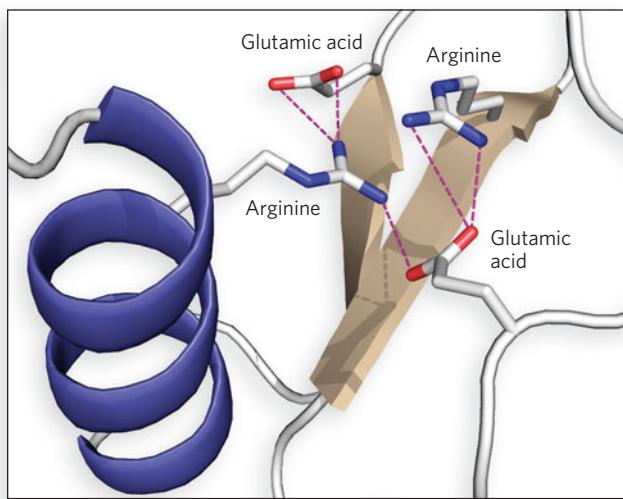


FIGURE 3-9 Salt bridges between oppositely charged amino acid side chains. In this representation of the three-dimensional structure of the backbone of a polypeptide, dotted lines indicate salt bridges between the positively charged amino group of an arginine side chain (blue) and the negatively charged carboxyl group of a glutamic acid side chain (red). Only the side chains of the four residues involved in the salt bridges are shown. See Figure 4-10 for an explanation of how molecular structures of proteins are represented. [Source: PDB ID 1HGD.]

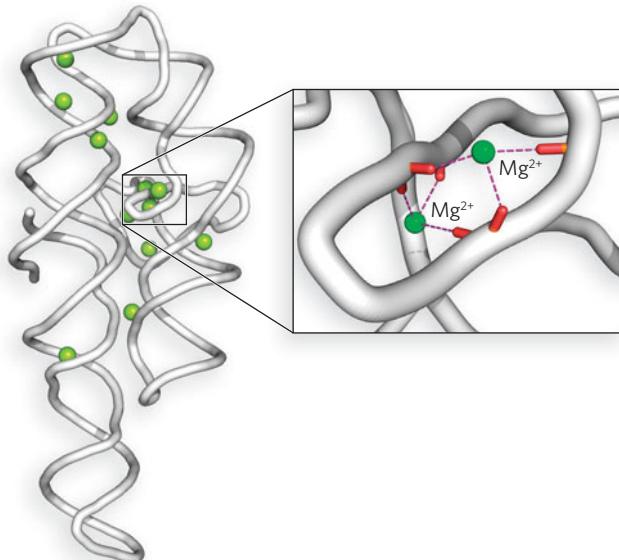


FIGURE 3-10 Stabilization of RNA structures by magnesium ions. In the P4-P6 domain of the *Tetrahymena* self-splicing group I intron, several magnesium ions (green) participate in salt bridges in the center of the folded RNA. The enlarged view shows magnesium ions stabilizing the RNA structure, with dotted lines indicating ionic bonds to the $-\text{O}^-$ (red) groups of adjacent phosphates in the RNA backbone. [Source: PDB ID 1GID.]

Sodium chloride is a good example of a molecule with a single strong ionic bond. Like other metals, sodium tends to form ionic bonds because it loses an electron, producing a positive ion that is strongly attracted to the negatively charged chloride ion. However, many ionic bonds found in biological molecules are weaker than the strong ionic or covalent bonds.

Chemical Bonds Are Explainable in Quantum Mechanical Terms

Although the idea of shared electron pairs provides a useful qualitative description of covalent bonding, the nature of the strong and weak forces that produce chemical bonds remained unknown to chemists until the development of quantum mechanics in the 1920s. The German scientists Walter Heitler and Fritz London offered the first successful quantum mechanical explanation of molecular hydrogen in 1927, laying the foundation for predicting the structures and properties of other simple molecules. Their work was based on the **valence bond model**, which posits that a chemical bond forms when there is suitable overlap between the electron clouds, or **atomic orbitals**, of participating atoms. These atomic orbitals are known

to have specific angular interrelationships, and thus the valence bond model can predict the bond angles observed in simple molecules. Today, the valence bond model has been supplemented with the **molecular orbital model**, in which the atomic orbitals of bonded atoms interact to form hybrid molecular orbitals. These molecular orbitals extend between the two bonding atoms.

Each element forms a characteristic number of bonds necessary to give it a complete outer shell of electrons. Because a complete outer shell, for most atoms, contains eight electrons, this is known as the octet rule. The maximum number of covalent bonds a particular atom can form is called its **valence**. The valence of the atoms commonly found in biological molecules dictates the shape, chemical properties, and ultimately the behavior of these molecules, even for large polymers such as nucleic acids and proteins. Hydrogen, oxygen, nitrogen, and carbon have valences of 1, 2, 3, and 4, respectively. Thus, hydrogen can form just one covalent bond, and O, N, and C can form any combination of single or multiple bonds to make up the total allowable number (Figure 3-11). A **single bond** between two atoms involves two electrons. Four shared electrons between two atoms produce a **double bond** (Figure 3-12).

Atom	Outer electrons	Usual number of covalent bonds	Bond geometry
Hydrogen	H	1	
Oxygen	··O··	2	
Nitrogen	··N··	3	
Carbon	··C··	4	

FIGURE 3-11 Valences of atoms that are common in biological molecules. Valence electrons, shown as dots in the Lewis structures in the second column, are the electrons in the outer shell of an atom that are available for chemical bonding. Examples of the resulting bond geometry are shown in the fourth column. [Source: Adapted from H. Lodish et al., *Molecular Cell Biology*, 5th ed., W. H. Freeman, Table 2-1.]

The angle between two bonds originating from a single atom is called the **bond angle**. The angle between two specific types of covalent bonds is always approximately the same. For example, the four single covalent bonds of a carbon atom are directed toward the corners of a tetrahedron (bond angle = 109.5°) (Figure 3-13). Covalent bonds differ in the degree of rotation they allow. Single bonds permit free rotation of the bound atoms around the bond, whereas double bonds are more rigid.

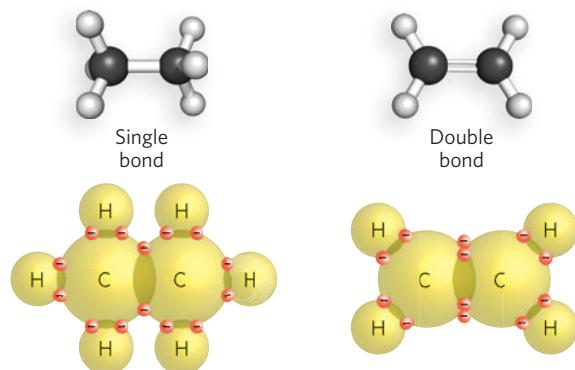


FIGURE 3-12 Shared electrons in single and double covalent bonds. Two electrons are shared between two atoms in a single covalent bond; four electrons are shared in a double covalent bond.

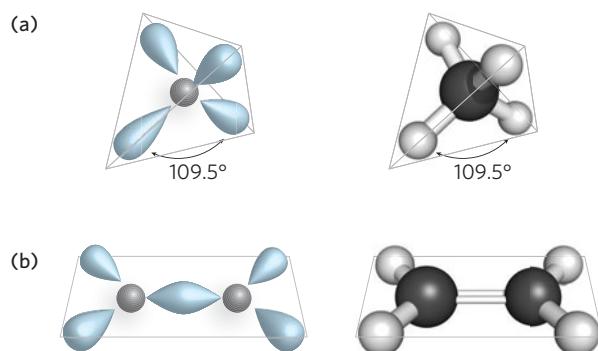
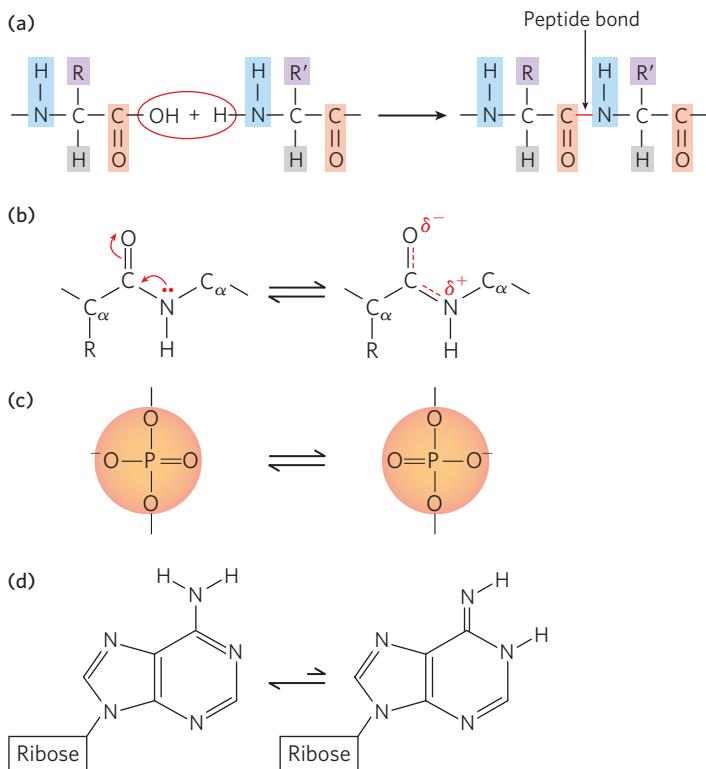


FIGURE 3-13 Geometry of single- and double-bonded carbon. (a) The four covalent bonds of single-bonded carbon point to the corners of a tetrahedron. (b) The covalent bonds of double-bonded carbons lie in a plane. Molecular orbitals of the carbon atoms are shown on the left.

Many molecules that contain single and double bonds adjacent to each other can exist as an average of multiple structures, a phenomenon called **resonance**. A **resonance hybrid** is a molecule that exists in an average of two possible forms. A classic example of this in biology is the molecular structure around the bond that links together two amino acids, known as a peptide bond (Figure 3-14a). The peptide bond links the carboxyl group of one amino acid to the amino group of another. The resulting carbonyl (C=O) and imino (C=N) bonds each have both double- and single-bond properties (Figure 3-14b). As a result, chemical groups bound together by the peptide bond in proteins must be located in the same plane, because the partial double-bond character of the carbonyl and imino bonds restricts rotation about these positions. As we shall see, this has profound consequences for the structure and function of proteins (see How We Know; also see Chapters 4 and 5).

Resonance also affects the behavior of nucleic acids, in multiple ways. The individual nucleotides of DNA and RNA are linked together covalently by phosphodiester bonds, which have a tetrahedral geometry and include two bonds to oxygen atoms that are related by resonance (Figure 3-14c). As a consequence, the negative charge in the phosphate group can shift between the two oxygen atoms that are not bonded to sugars in the backbone. Also, the bases are conjugated ring systems—that is, they have alternating double and single bonds—giving rise to shared electrons around the ring(s) (Figure 3-14d). As we discuss in Chapter 6, the accuracy of base pairing between two DNA strands, A with T and C with G, results from the dominance of particular resonance structures of the bases.

**FIGURE 3-14** Resonance in peptides and nucleic acids.

(a) Peptide bonds covalently link the carboxyl group and amino group of adjacent amino acid residues in a protein. (b) Resonance between the resulting carbonyl and imino bonds gives each the properties of both a single and a double bond. Rotation is restricted about these bonds, and thus the attached chemical groups must lie in the same plane. For electron movements, a full arrowhead (\rightarrow) indicates two electrons; a half arrowhead ($\overrightarrow{}$) denotes one electron. The partial positive and negative charges are represented by δ^+ and δ^- . (c) Resonance in the phosphate group (of the phosphodiester bond) of nucleic acids. (d) Resonance in the bases of nucleic acids; the adenine nucleoside is shown here. See Chapter 6 for resonance structures of other bases.

Both the Making and Breaking of Chemical Bonds Involve Energy Transfer

For a chemical bond to form, the total energy of the system—defined as the molecule and its environment—must be lower in the bonded state than in the nonbonded state. Therefore, bonding is an **exothermic** process, releasing energy on bond formation. The strength of a covalent bond increases with decreasing bond length; thus, two atoms connected by a single strong covalent bond, or by a double bond, are always

closer together than identical atoms held together by a single weak covalent bond. As mentioned above, the electronegativity of an atom can be used to predict the type and strength of bonding between that atom and any other atom. Stronger bonds release more energy on formation than weaker bonds. For two atoms A and B, the rate of bond formation is directly proportional to the frequency with which A and B bump into each other.

A calorie is the amount of energy needed to raise the temperature of 1 gram of water by 1 degree Celsius, from 14.5°C to 15.5°C. Energy changes in chemical reactions are typically expressed in kilocalories per mole (kcal/mol), because thousands of calories are involved in forming or breaking a **mole** of—that is, 6.02×10^{23} —chemical bonds. But if energy is given off when two atoms combine to form a covalent bond, then the separated atoms must have more total energy than the molecule. The amount of energy required to break a chemical bond exactly equals the amount that was released on its formation. This equivalence follows from the first law of thermodynamics, which states that energy, except where interconvertible with mass, can be neither created nor destroyed (see Section 3.6). This, then, is what holds atoms together in covalent bonds: they cannot separate unless they are given the required amount of energy.

Bond breakage frequently occurs on heating, because heat speeds up molecular motions, leading to intermolecular collisions in which some of the kinetic energy of a moving molecule is released as it pushes apart two bonded atoms. Higher temperatures produce faster-moving molecules and hence a greater chance that collisions will break bonds. Therefore, molecules are less stable at higher temperatures.

Electron Distribution between Bonded Atoms Determines Molecular Behavior

All chemical bonds, whether strong or weak, are the result of attractions between electrical charges. For example, the hydrogen molecule (H:H) has a symmetric distribution of electrons between its two hydrogen atoms, so both atoms are uncharged and the bond they share is purely covalent. In contrast, the polar covalent bonds of a water molecule (H:O:H) have a nonuniform distribution of charge. In water, the bonding electrons are unevenly shared due to the different electronegativities of H and O atoms (see Figure 3-8). In this case, the oxygen atom holds the bonding electrons more strongly and thus has a considerable negative charge, whereas the two

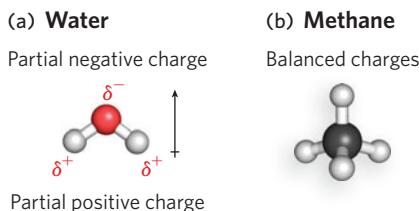


FIGURE 3-15 Polar and nonpolar molecules. (a) The polar water molecule has positive and negative poles and carries an electric dipole moment. Dipole moment is a vector quantity and is represented by a small arrow, pointing from the positive charge toward the negative charge. (b) A nonpolar methane molecule has no separation of charges.

hydrogen atoms together have an equal amount of positive charge (Figure 3-15a). Such a combination of separated positive and negative charges is called an **electric dipole moment**.

Molecules such as water that have a dipole moment are referred to as **polar molecules**. **Nonpolar molecules** are those with no effective dipole moment; an example is methane (Figure 3-15b). The large size of proteins and nucleic acids allows polar and nonpolar regions to exist within the same molecule. For example, the outer surfaces of proteins that function in the aqueous environment of the cell cytoplasm tend to be polar, thereby favoring interactions with polar water molecules. In contrast, proteins that function in the nonpolar environment of cellular membranes tend to have nonpolar surfaces, thus fostering contacts with the nonpolar fatty acid chains of the membrane.

SECTION 3.2 SUMMARY

- All molecules consist of atoms linked together by strong and/or weak chemical bonds.
- Covalent bonds share electrons equally between two atoms, whereas ionic bonds have electrons that are completely transferred from one atom to another, such that the atoms are drawn together by electrostatic forces. Electronegativity, a measure of how strongly an atom attracts electrons to itself, can be used to predict the type of bonding between two atoms.
- Valence is the maximum number of covalent bonds an atom can form, and the valence of the atoms in biological molecules dictates the shape of these molecules. Carbon, with a valence of 4, forms four

single bonds to neighboring atoms arranged in a tetrahedral geometry.

- Resonance is an aspect of valence bond theory used to graphically represent and mathematically model molecules for which no single, conventional Lewis structure can satisfactorily represent the observed molecular structure or explain its behavior. Such molecules are considered to be an intermediate or average, or resonance hybrid, of several Lewis structures that differ only in the placement of the valence electrons.
- Single bonds, in which a pair of atoms share two electrons, give rise to variable geometries, whereas double bonds, involving four shared atoms, give rise to planar molecular geometries.
- Exothermic bond formation is energetically favorable, and the total energy of the system is decreased in the process. The amount of energy released when a chemical bond breaks is the same as that required for its formation. Energy changes in chemical reactions are expressed in kilocalories per mole.
- Attractions between electrical charges in atoms lead to chemical bond formation. Two bonded atoms are uncharged when the bonding electrons are positioned equally between them. When one atom holds the bonding electrons more tightly than the other, due to differences in the atoms' electronegativities, an unequal charge distribution results. One end of the molecule carries a net negative charge, the other a net positive charge. The molecule is said to have an electric dipole moment.
- Polar molecules are those with a dipole moment, such as water. Some molecules, such as methane, lack a dipole moment and are nonpolar. The polarity of biomolecules governs their locations within cells and their interactions with other molecules.

3.3 Weak Chemical Interactions

The macromolecules of most interest to molecular biologists—nucleic acids and proteins—are formed by the covalent joining of their constituent atoms. Because covalent bonds are relatively strong, stable, and not subject to spontaneous breakage under physiological conditions, they were once thought to be solely responsible for holding together the atoms in molecules. In contrast, weak chemical interactions, sometimes called

weak chemical bonds, involve greater distances between atoms, are easily broken, and, individually, are transient. These properties, however, can be useful in biological systems, where transient chemical interactions are an essential part of cellular functions. For example, weak bonds mediate the interactions of proteins with small molecules, DNA, or other proteins.

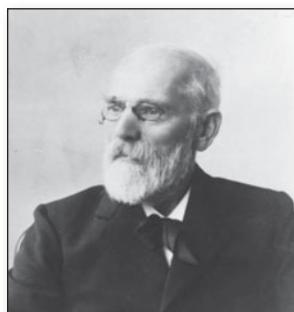
When arranged in ordered groups, weak bonds can persist for a long time and can thus play central roles in the formation and stability of the active three-dimensional shapes of macromolecules. DNA is a case in point: although DNA is a linear polymer of covalently linked nucleotides, its shape and its ability to encode genetic information are determined by the stable yet dynamic double-helical structure that it adopts. This structure is defined by a large number of individually weak contacts between nucleotides that are not covalently bonded. Likewise, protein structures are largely determined by weak interactions between amino acids that are not necessarily adjacent in the polypeptide sequence. Therefore, although they are not strong enough individually to effectively bind two atoms together, weak chemical interactions play central roles in the structure and behavior of biological macromolecules.

Three kinds of weak chemical interactions are important in biological systems: van der Waals forces, hydrophobic interactions, and hydrogen bonds. Most macromolecules also use weak ionic interactions, along with these weak chemical interactions, to bind other molecules and to form three-dimensional structures.

Van der Waals Forces Are Nonspecific Contacts between Atoms

The Dutch chemist Johannes van der Waals was the first to document the intermolecular forces that result from the polarization of atoms. When two atoms approach each other, as they

get closer, induced fluctuating charges between them cause a weak, non-specific attractive interaction. This **van der Waals interaction** depends heavily on the distance between the interacting atoms. As the distance decreases below a certain point, a more powerful van der Waals repulsive force is caused by overlap of the atoms' outer



Johannes van der Waals,
1837–1923 [Source: NIH/
NLM.]

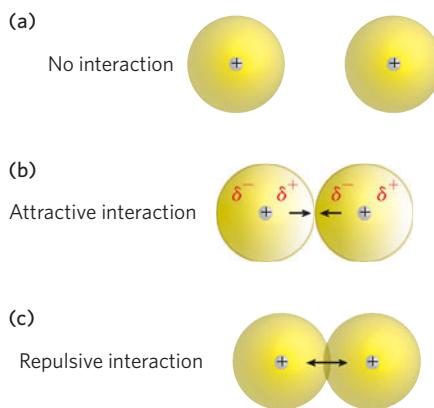


FIGURE 3-16 Attractive and repulsive van der Waals interactions. (a) Atoms that are farther apart than the sum of their van der Waals radii do not experience van der Waals forces. (b) Attractive interactions arise when atoms are separated by a distance equal to their combined van der Waals radii. (b) Repulsive interactions arise when atoms are closer than their combined van der Waals radii.

electron shells. The **van der Waals radius** of an atom is defined as the distance at which these attractive and repulsive forces are balanced and is characteristic for each atom (Figure 3-16).

The van der Waals bonding energy between two atoms that are separated by the sum of their van der Waals radii increases with the size of the atoms, but on average is only about 1 kcal/mol—just slightly above the average thermal energy of molecules at room temperature. This means that van der Waals forces are an effective binding force under physiological conditions only when they involve several atoms in each of the two interacting molecules. For several atoms to interact effectively in this way, the intermolecular fit must be exact, because the distance between any two interacting atoms must not be much different from the sum of their van der Waals radii.

The strongest kind of van der Waals contact arises when a macromolecule contains a cavity that precisely fits the shape of the molecule that it binds. This is the case for antibodies, proteins that recognize antigens—specific molecules of viruses, bacteria, or other foreign particles that enter the body. Antibodies contain clefts with the same shape as the antigen they bind, enabling van der Waals contacts along the length of the bound antigen (Figure 3-17a). The additive effects of many van der Waals forces can be exceptionally strong; for instance, they are responsible for a gecko's ability to climb vertically and hang on a glass surface using only one toe (Figure 3-17b).

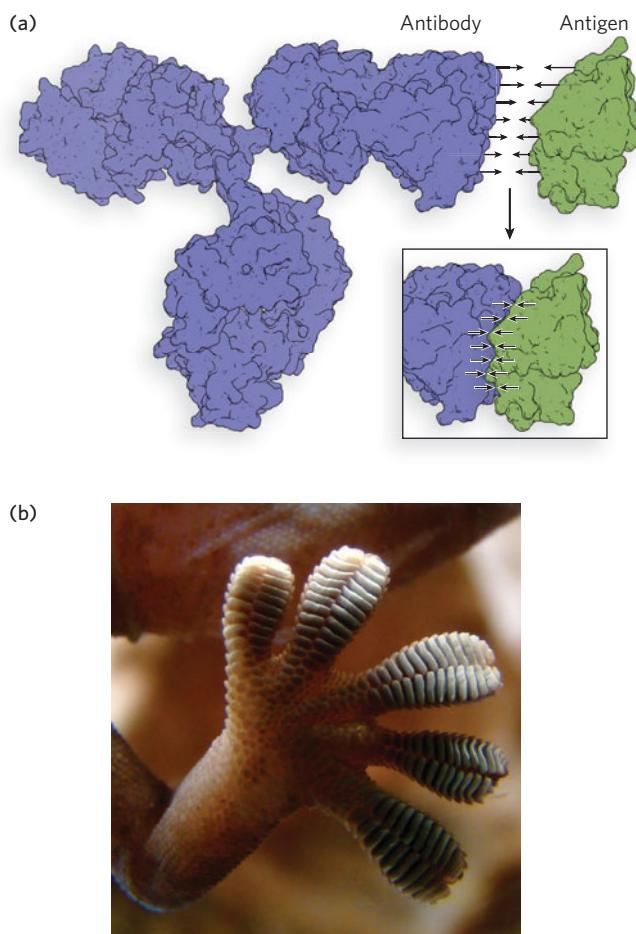


FIGURE 3-17 Examples of van der Waals interactions in nature. (a) The binding pockets of antibodies and the molecules they recognize, called antigens, typically fit together through van der Waals interactions. (b) It is also van der Waals interactions that enable a gecko to climb vertical surfaces, through the enormous number of interactions between its foot pads and the molecules of the surface material, such as glass. [Sources: (a) Adapted from H. Lodish et al., *Molecular Cell Biology*, 6th ed., W. H. Freeman, Fig. 24-13; PDB ID 1IGT and PDB ID 3HFM. (b) Bjorn Christian Torrisen.]

Hydrophobic Interactions Bring Together Nonpolar Molecules

Hydrophobic interactions arise from the strong tendency of water to exclude nonpolar groups, forcing these groups into contact with one another. Such hydrophobic contacts stabilize protein structures, accounting for most of the energy required for protein folding. The helical structures in DNA and RNA are also stabilized by hydrophobic interactions that arise from the stacked arrangement of base pairs. In accordance with the molecular orbital model introduced earlier, the electrons of the atoms in a base's aromatic

rings are found in a decentralized cloud above and below the rings, in an arrangement known as pi bonding. The overlap of pi bonds of adjacent stacked nucleotides is what contributes to the stability of the helical structure (Figure 3-18). Hydrophobic interactions are critical to many other cellular functions as well, including protein insertion into membranes and secretion of hormones and other signaling molecules.

The word *hydrophobic* means “water-fearing,” which describes the apparent behavior of nonpolar molecules in water. For example, when some drops of oil are added to water, they combine to form a larger drop. This is because water molecules are attracted to one another, due to their polarity, whereas oil molecules are nonpolar and therefore have no charged regions to repel or attract other molecules. The attractive forces between water molecules, and the resulting unfavorable organization of water molecules in the vicinity of hydrophobic molecules, have the effect of squeezing the oil drops together to form a larger aggregate, thereby minimizing the surface area in contact with water. Such hydrophobic effects are common in nature and are closely tied to such processes as the folding of protein molecules into their functional shapes and the binding of nucleic acids and proteins to other molecules.

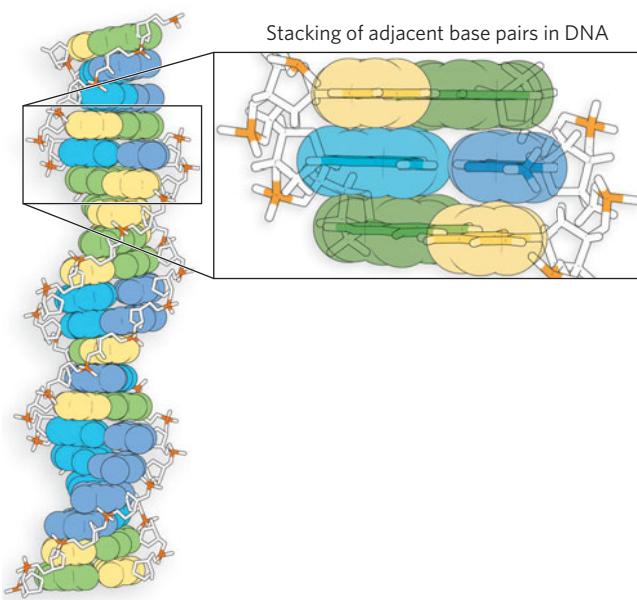


FIGURE 3-18 Hydrophobic interactions in DNA. Stacking of rings in adjacent bases of DNA involves favorable hydrophobic interactions. Molecular orbital overlap between bases stabilizes the double helix. The DNA backbone is shown in white, with phosphate groups in orange. Bases are colored as outlined in Figure 3-3. The spheres indicate van der Waals radii of the atoms in the bases.

Hydrogen Bonds Are a Special Kind of Noncovalent Bond

A **hydrogen bond** is an attractive intermolecular force between two partial electrical charges of opposite polarity. As the name implies, one partner in the bond is a hydrogen, which must be covalently bonded to a strongly electronegative atom such as oxygen, nitrogen, or fluorine; this hydrogen is the hydrogen-bond donor. The electronegative atom attracts the electron cloud from around the hydrogen nucleus and, by decentralizing the cloud, leaves the hydrogen atom with a partial positive charge. This partial charge represents a large charge density that can attract the lone pair of electrons on another, nonhydrogen atom, which becomes the hydrogen-bond acceptor (Figure 3-19). Although other types of atoms can similarly acquire a partial positive charge when bonded to an electronegative element, only hydrogen is small enough to approach another atom or molecule close enough to undergo an energetically significant interaction.

The hydrogen bond is not like a simple attraction between positive and negative charges at two points in

space, but instead has directional preference and some characteristics of a covalent bond. This covalent character is more pronounced when acceptors hydrogen-bond with hydrogens from more-electronegative donors. For this reason, hydrogen bonds can vary in strength from very weak (a bonding energy of 2 kcal/mol) to fairly strong (7 kcal/mol).

The chemical properties of water are mostly due to hydrogen bonds that cause water molecules to have strong attractions for one another, but hydrogen bonds can and do form among many other kinds of molecules as well. In proteins, hydrogen bonds occur between the backbone carbonyl and imino groups to stabilize α helices and β sheets, the basic motifs of protein structure (see Chapter 4). In nucleic acids, hydrogen bonding between complementary pairs of nucleotide bases links one DNA strand to its partner, forming the double helix (see Figure 6-11).

Combined Effects of Weak Chemical Interactions Stabilize Macromolecular Structures

The noncovalent interactions we have discussed (van der Waals forces, hydrophobic interactions, and hydrogen bonds) are substantially weaker than covalent bonds. Just 1 kcal is needed to disrupt a mole of typical van der Waals interactions, but nearly 100 times more energy is required to break an equivalent number of covalent C–C or C–H bonds. Hydrophobic interactions, too, are much more easily disrupted than covalent bonds, although they are significantly strengthened in the presence of polar solvents such as concentrated salt solutions. Hydrogen bonds vary in strength depending on the polarity of the solvent and the alignment of the hydrogen-bonded atoms, but again, they are always weaker than covalent bonds. In aqueous solvent at 25°C, the available thermal energy is typically of the same order of magnitude as the strength of these weak interactions. Furthermore, the interaction of solute and solvent (water) molecules is nearly as favorable as any solute-solute interactions. Consequently, van der Waals forces, hydrophobic interactions, and hydrogen bonds continually break and re-form under physiological conditions.

Although these types of interactions are individually weak relative to covalent bonds, the cumulative effect of many such interactions can be significant. Macromolecules, including DNA, RNA, and proteins, contain so many sites of possible van der Waals or hydrophobic contacts and hydrogen bonding that the combined effect of many small binding forces is

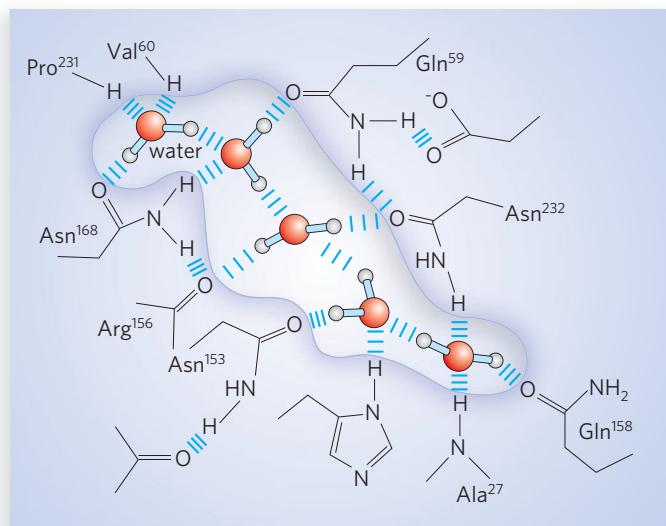


FIGURE 3-19 Hydrogen bonds. Water can donate or accept a hydrogen atom to form a hydrogen bond with another water molecule or another kind of atom. For example, shown here is a region of cytochrome *f* (a photosynthetic protein) in which water molecules form hydrogen bonds with one another and with amino acid residues of the protein. Each hydrogen bond is represented by three blue lines. [Source: Adapted from P. Nicolls, *Cell. Mol. Life Sci.* 57:987, 2000, Fig. 6a (redrawn from information in the PDB and a Kinemage file published by S. E. Martinez et al., *Protein Sci.* 5:1081, 1996).]

largely responsible for their molecular structure. For macromolecules, the most stable structure is usually that in which weak interactions are maximized. This principle determines the folding of a single polypeptide or polynucleotide chain into its three-dimensional shape. Complete unfolding of the structure requires the removal of all these interactions at the same time. And because these contacts are breaking and re-forming rapidly and randomly, such synchronized disruptions are very unlikely. The molecular stability conferred by many weak interactions is therefore much greater than one might expect intuitively, based on a simple summation of many small binding energies.

A special class of weak interactions in macromolecules involves the water molecules that are invariably bound to surface and interior sites by hydrogen bonds. Sometimes these water molecules are so well positioned that they behave as though they are an integral part of the macromolecule. In many cases, bound water molecules are essential to macromolecular function. Certain DNA-binding proteins, for example, use water molecules that are integral to their structure to help recognize specific DNA sequences. In RNA structures, water molecules bridge nucleotide bases that are involved in nontraditional base pairing and in interactions involving three nucleotides (base triples),

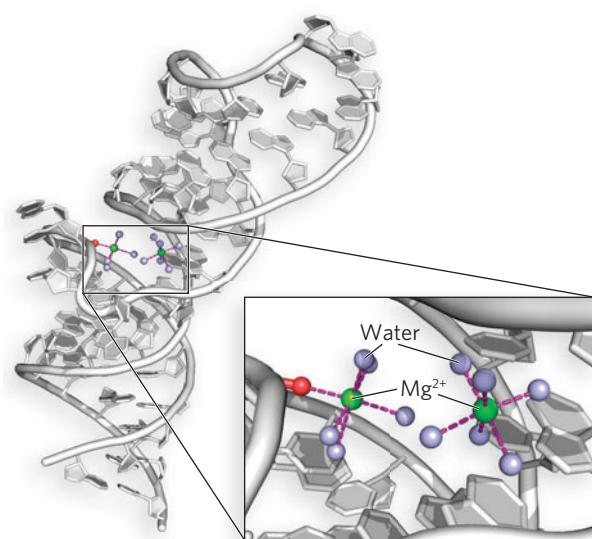


FIGURE 3-20 Ordered water molecules in an RNA molecule. Hydrogen bonding between water molecules and nucleotides is essential to the three-dimensional structure of an RNA molecule. Magnesium ions (green) form hydrogen bonds (dotted lines) with water molecules (blue) and with the RNA (red). [Source: PDB ID 1DUL.]

thereby enabling unique three-dimensional structures to form (Figure 3-20).

Weak Chemical Bonds Also Facilitate Macromolecular Interactions

Interactions among DNA, RNA, and protein molecules, and interactions between these macromolecules and small organic molecules, are also mediated by weak, noncovalent chemical interactions. Such contacts involving information-carrying macromolecules govern how and when cells replicate, repair and recombine their DNA, synthesize RNA and proteins, detect and respond to chemical signals, and conduct all the other activities essential for life.

For example, the noncovalent binding of a protein to another protein, to a small molecule (such as a hormone), or to another macromolecule (such as a nucleic acid) may involve one or more ionic, hydrophobic, and van der Waals interactions, as well as several hydrogen bonds (Figure 3-21). These weak

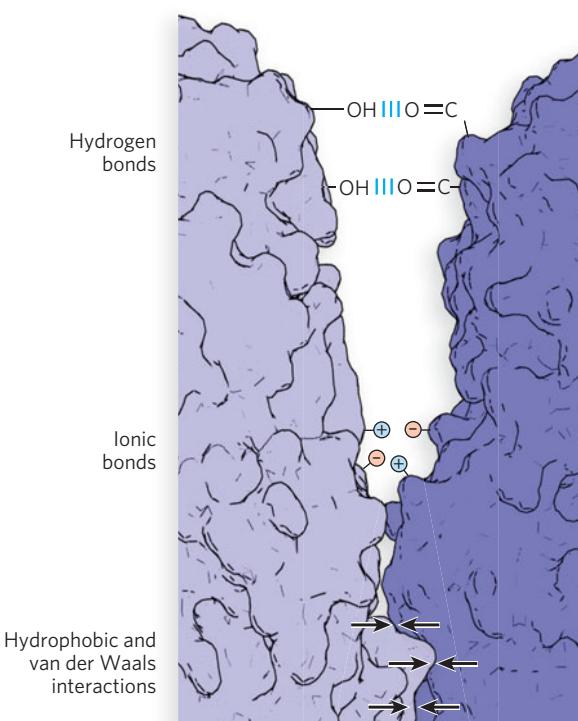


FIGURE 3-21 Stabilization of macromolecular structure by weak interactions. The combined effects of multiple noncovalent forces, including hydrogen bonds, hydrophobic and van der Waals interactions, and weak ionic bonds, allow the specific and stable association of molecules—in this case, two proteins. [Source: Adapted from H. Lodish et al., *Molecular Cell Biology*, 6th ed., W. H. Freeman, Fig. 2-12.]

contacts contribute, overall, to an energetically favorable interaction. It is worth noting, however, that because each hydrogen bond between two groups in a protein or nucleic acid forms at the expense of *two* hydrogen bonds between the same groups and water, the net stabilization is not as great as it might seem. The large size of nucleic acids and proteins provides extensive molecular surfaces with many opportunities for weak interactions with other molecules. The energetic favorability of such interactions reflects the molecular surface complementarity and the resulting large numbers of weak interactions of polar, charged, and hydrophobic groups on the surfaces of these molecules.

SECTION 3.3 SUMMARY

- Weak chemical bonds differ from covalent or ionic bonds in several ways: they involve greater distances between atoms, are easily broken, and are often transient. These properties are useful when short-lived chemical interactions are required in cells. Weak bonds often mediate the interactions of proteins with small molecules, DNA, hormones, or other proteins.
- Van der Waals forces, hydrophobic interactions, and hydrogen bonds are the three most important kinds of weak chemical interactions used by macromolecules to bind other molecules and to form three-dimensional structures. Weak ionic interactions are also used.
- Van der Waals forces occur when two atoms closely approach each other, inducing fluctuating charges that cause a weak, nonspecific, attractive interaction between them. Each type of atom has its own van der Waals radius, the distance at which attractive and repulsive forces with neighboring atoms are balanced.
- Hydrophobic interactions arise from the strong tendency of water to exclude nonpolar groups, forcing these groups into contact with one another. Hydrophobic contacts stabilize protein structures and the stacking of bases in DNA and RNA helices.
- Hydrogen bonds occur between two atoms with partial electrical charges of opposite polarity, one of which is a hydrogen atom. Other types of atoms can also acquire partial positive charge, but only hydrogen can approach another atom or molecule close enough for an energetically useful interaction.

- Although weak chemical interactions have only minor attractive or repulsive effects individually, the cumulative effects can be significant. DNA, RNA, and proteins contain so many sites of possible van der Waals or hydrophobic contacts and hydrogen bonding that the combined effect of many small binding forces is largely responsible for their molecular structure.

3.4 Stereochemistry

The concept of **stereochemistry**, the spatial arrangement of the atoms within a molecule, is critical for understanding the structures and activities of biological molecules. As we shall see, the orientation of the chemical bonds of nucleic acids and proteins influences how these molecules fold in three dimensions, bind to other molecules, and catalyze reactions.

Three-Dimensional Atomic Arrangements Define Molecules

Our understanding of stereochemistry stems from a discovery by the French physicist Dominique Arago in 1811. Arago found that plane-polarized light, which vibrates in just one plane, rotates when it is sent through a piece of quartz crystal. Other scientists subsequently showed that many, but not all, substances share the ability to rotate the plane of polarized light; a substance with this ability is said to be **optically active**. In 1848, 26-year-old chemist and crystallographer Louis Pasteur proposed that if a molecule is not superimposable on its mirror image, it will be optically active; otherwise, it is optically inactive. This bold prediction subsequently proved to be correct. In fact, all objects can be classified as those that can be superimposed on their mirror image, such as golf balls and champagne glasses, and those that cannot, such as hands. Objects that can be superimposed on their mirror image are said to be **achiral**; those that cannot are **chiral** (Figure 3-22). All optically active chemical compounds are chiral, and all optically inactive compounds are achiral.

This is an important idea in molecular biology, because many biologically relevant molecules are chiral. For example, a carbon atom connected to four different groups—such as the α -carbon atom in an amino acid, or the carbons in the pentose sugars of nucleotides—is an asymmetric carbon atom, or **chiral center**. Because carbon points its four single bonds to

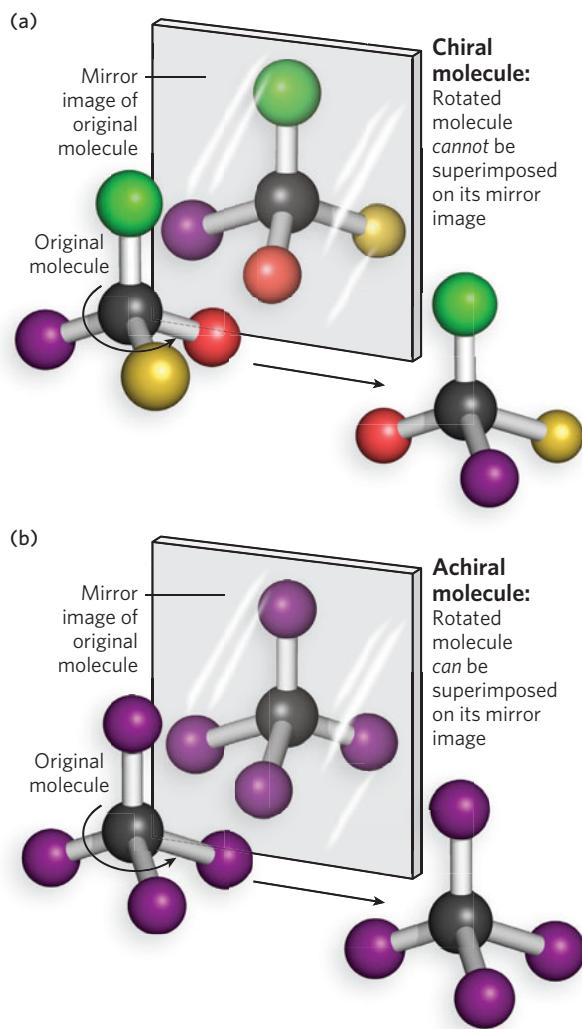


FIGURE 3-22 Chiral versus achiral molecules. (a) When bonded to four different types of atoms, a carbon atom is chiral, because the four atoms can be arranged in two different ways that are not superimposable mirror images. Thus, the two forms of the molecule have the same chemical formula but different chemical behavior. (b) In contrast, a carbon atom bonded to four identical atoms is achiral, because only one configuration is possible and any arrangement of the four bound atoms is superimposable on its mirror image.

the corners of a tetrahedron, any molecule containing a single asymmetric carbon atom cannot be superimposed on its mirror image, analogous to the inability to superimpose a left-hand glove on a right-hand glove. Thus, all nucleic acids and proteins are chiral, because they contain carbon atoms that are chiral centers, as we shall see shortly.

Biological Molecules and Processes Selectively Use One Stereoisomer

Two molecules that have the same chemical and structural formula but differ in the arrangement of their atoms in space (i.e., are not superimposable) are called **stereoisomers**. A pair of stereoisomers that are mirror images of each other are **enantiomers**. All physical and chemical properties of enantiomers are identical, except for two. First, the two enantiomers rotate the plane of polarized light in opposite directions. The one that rotates the light to the right is called the **D** form (dextrorotatory), and the one that rotates the light to the left is called the **L** form (levorotatory). The second distinction is that two enantiomers react at the same rate with any achiral compound, but at different rates with any chiral compound.

Many biological molecules are chiral, and living organisms usually use only one of the two possible enantiomers. Thus, biochemical reactions have evolved to recognize and favor one stereoisomer over the other. For example, sucrose (table sugar) is chiral and rotates the plane of polarized light to the right (**D**-sucrose). **D**-Sucrose can be broken down by digestive enzymes, which are also chiral, at a rate commensurate with its use as an energy source. Because digestive enzymes are chiral, they do not recognize and digest sucrose that has the opposite chirality (**L**-sucrose), even though the two sugar molecules are identical in chemical and structural formulas.

Proteins and Nucleic Acids Are Chiral

Proteins are chiral because nearly all of the common amino acids contain an α carbon bonded to four different groups: a carboxyl group, an amino group, an R group, and a hydrogen atom. (The amino acid glycine is the exception, because its R group is simply a hydrogen atom.) The α carbon is therefore a chiral center, and the tetrahedral arrangement of the bonding orbitals around the α -carbon atom enables the four different functional groups to occupy two different possible spatial arrangements. This means that each amino acid, except for glycine, has two possible stereoisomers. Because these are nonsuperimposable mirror images of each other, the two forms are enantiomers (Figure 3-23a). Like all molecules with a chiral center, amino acids are optically active and rotate plane-polarized light. Nature uses only one of these enantiomeric forms in proteins, and it is the same one for each amino acid—the **L** form. **D**-Amino acids are almost never found in nature; this phenomenon has been cleverly harnessed by researchers to design new

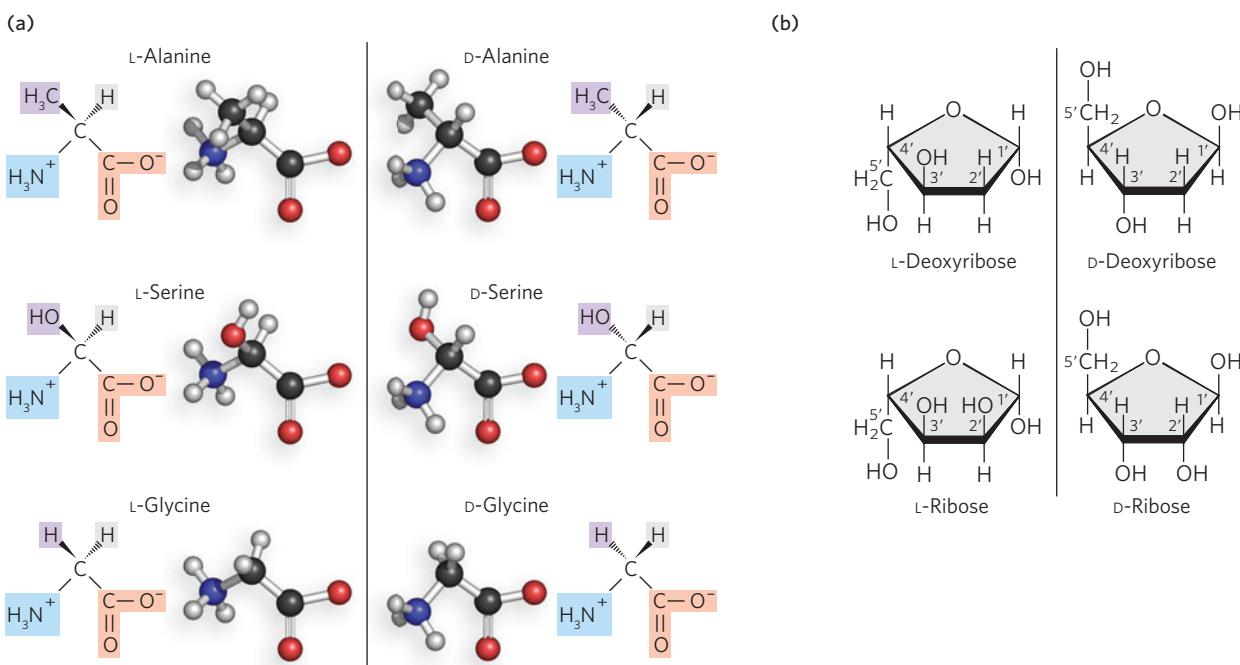


FIGURE 3-23 Enantiomers of amino acids and nucleotides. (a) Each amino acid, except for glycine, has two possible stereoisomers. These nonsuperimposable mirror images of each other (enantiomers) are known as levorotatory (*L*) and dextrorotatory (*D*) forms. Only the

L forms of amino acids are found in natural proteins. (b) All of the carbon atoms in ribose or deoxyribose except for C-5' are chiral centers. Nature uses only the *D*-enantiomeric form of the sugars—*D*-ribose or *D*-deoxyribose—as the building blocks for DNA and RNA.

therapeutics (Highlight 3-1). Note that some amino acids also contain another asymmetric carbon atom, besides the α carbon (see Figure 4-3).

Nucleic acids, too, are chiral molecules. Nucleotides contain several chiral centers in the ribose or deoxyribose ring—all of the carbon atoms except for C-5' are asymmetric (Figure 3-23b). Again, nature uses one enantiomeric form of the sugar—*D*-ribose or *D*-deoxyribose—in the building blocks for DNA and RNA. Enzymes that act on DNA or RNA have evolved to recognize this chiral arrangement of the nucleic acids.

SECTION 3.4 SUMMARY

- Molecules with a structure that cannot be superimposed on its mirror image are chiral.
- All nucleic acids and proteins are chiral, because they contain carbon atoms that are bonded to four different atoms or groups. Because the four bonds of carbon point to the four corners of a tetrahedron, and the four functional groups can be spatially arranged in two different ways, these α -carbon atoms are chiral centers.
- Stereoisomers are pairs of molecules that have the same chemical and structural formulas but are not superimposable on each other, because they differ in the spatial arrangement of their atoms. Biochemical reactions have evolved to recognize and favor one stereoisomer over the other.
- Enantiomers are pairs of stereoisomers that are mirror images of each other. Enantiomers rotate the plane of plane-polarized light in opposite directions; the *D* form rotates it to the right, and the *L* form rotates it to the left.
- Like all molecules with a chiral center, amino acids are optically active and rotate plane-polarized light. Only the *L* enantiomer of each amino acid is found in natural proteins.
- Only the *D* enantiomer of the pentose sugar—*D*-ribose or *D*-deoxyribose—is used in the building blocks for RNA and DNA.

HIGHLIGHT 3-1 MEDICINE

The Behavior of a Peptide Made of D-Amino Acids

L-Peptides capable of binding and blocking the function of important therapeutic targets—such as viral enzymes or receptor proteins—have turned out to be virtually useless for treating patients, because once introduced into the body, they are quickly destroyed. This is because cells and blood serum contain proteases, enzymes that bind to normal, L-amino acid-containing peptides and catalyze their rapid degradation. To get around this problem, Peter Kim, then an MIT professor and Howard Hughes Medical Institute investigator (now president of Merck Research Laboratories), wondered whether unnatural peptides made of D-amino acids would be useful as therapeutics in cells or serum.

Kim and his coworkers (then at the Whitehead Institute at MIT) synthesized a portion of CD8, the human immunodeficiency virus (HIV) protein required for infection, in the unnatural, D-amino acid form, and they used it to identify L-amino acid peptides that bind specifically to this protein (Figure 1). These L-amino acid peptides were then resynthesized, but this time from D-amino acids. Because D- and L-amino acids are mirror images of



Peter Kim [Source: Merck Research Laboratories.]

each other, the new D-amino acid peptides bound to the target CD8 protein of the natural (L) form. Thus, the enantiomeric D-amino acid peptides could be used therapeutically to block activity of the natural viral protein, while avoiding degradation by enzymes that act only on natural peptides. This clever approach was subsequently used to make D-peptide-based drugs that were tested in clinical trials for inhibiting the entry of HIV into cells. Although these peptides were not effective as HIV therapeutics, in part due to difficulties in getting them into infected cells, the strategy of using biologically inert enantiomers as drugs remains attractive.

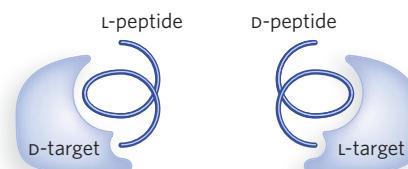


FIGURE 1 Kim's experiment to isolate L-amino acid peptides that bind to unnatural protein. Because of the mirror-image relationship between L- and D-amino acids, L-peptides that bind to the D form of a target protein have the same sequence as D-peptides that bind to the L form of the target protein. [Source: Adapted from T. N. Schumacher et al., *Science* 271: 1854–1857, 1996, Fig. 1.]

3.5 The Role of pH and Ionization

Although pH and the ionization potential of molecules may seem like esoteric topics for biology, they are central to the function of biological molecules. This is because most biological processes occur within a narrow pH range, and cells and their internal compartments carefully regulate pH to suit their needs. For example, foreign molecules that enter cells are degraded in lysosomes, organelles that maintain an interior pH of 4.5, an acidic environment that is optimal for the function of the lysosome's digestive enzymes. The pH of blood, in contrast, must be maintained near neutral pH—neither acidic nor basic. Blood with a pH lower than 7.35 is too

acidic to carry oxygen efficiently, and blood with a pH above 7.45 is too basic (alkaline). In most cases, the activities of proteins and nucleic acids depend on carefully controlled pH, and various mechanisms have evolved for regulating the pH in cells and organelles.

The Hydrogen Ion Concentration of a Solution Is Measured by pH

Many of the macromolecules and chemical reactions of central importance in molecular biology naturally occur in **aqueous solutions**, which are mixtures of molecules dissolved in water. Such solutions can be acidic, neutral, or basic, depending on the concentration of

positively charged water molecules, called hydronium ions (H_3O^+). For example, a solution with $[\text{H}_3\text{O}^+] = 4 \times 10^{-3}$ mol/L is more acidic and less basic than one with $[\text{H}_3\text{O}^+] = 5 \times 10^{-4}$ mol/L. Note that we consider hydronium ions rather than hydrogen ions (H^+), because any H^+ ions (protons) in an aqueous environment quickly bind to a water molecule to form H_3O^+ . Because small numbers such as 4×10^{-3} or 5×10^{-4} are difficult to work with, Sören Sörensen devised another way to express $[\text{H}_3\text{O}^+]$. In 1909, he defined the quantity **pH** as the negative logarithm of the hydronium ion concentration:

$$\text{pH} = -\log [\text{H}_3\text{O}^+] \quad (3-1)$$

It is simple to convert $[\text{H}_3\text{O}^+]$ to pH, and vice versa.

Similarly, we can define **pOH** as $-\log [\text{OH}^-]$. We can easily express the acidity or basicity of an aqueous solution by using pH and pOH; by common practice, however, pH is used rather than pOH. In aqueous solutions at equilibrium, the product $[\text{H}_3\text{O}^+][\text{OH}^-] = 10^{-14}$, and pH + pOH must equal 14. Thus, a solution with a pH of 4.6 has a pOH of 9.4. A solution of pH 7 is neutral, because $[\text{H}_3\text{O}^+] = [\text{OH}^-]$. A solution with a pH lower than 7 is acidic, and $[\text{H}_3\text{O}^+] > [\text{OH}^-]$; one with a pH greater than 7 is basic, and $[\text{H}_3\text{O}^+] < [\text{OH}^-]$. The lower the pH, the more acidic the solution; similarly, the higher the pH, the more basic the solution. Because pH values are on a logarithmic scale, a change in pH of one unit corresponds to a tenfold change in hydronium ion concentration.

KEY CONVENTION

Square brackets are used to indicate concentration, meaning the number of molecules per unit volume, of a particular chemical species. Concentration is often expressed as molarity, abbreviated **M**. 1 M = 1 mol/L (1 mole per liter).

In the laboratory, the pH of an aqueous solution can be determined by using chemical compounds (pH indicators) that change color over a narrow pH range, or with an instrument called a pH meter, which has electrodes that are dipped into the sample solution. Many foods and other common household substances are acidic or basic aqueous solutions. Figure 3-24 lists approximate pH values of some foods and common materials, as well as some body fluids.

Buffers Prevent Dramatic Changes in pH

Most macromolecules found in biological systems have evolved to function at approximately neutral pH,

Concentration of hydrogen ions (M)	Scale	Common examples
Most acidic 10 ⁻¹	pH 1	Battery acid
10 ⁻²	pH 2	Lemons, stomach (hydrochloric) acid
10 ⁻³	pH 3	Soft drinks, apples, cheeses
10 ⁻⁴	pH 4	Bottled water (reverse osmosis)
10 ⁻⁵	pH 5	Coffee, beer, bananas, sugar
10 ⁻⁶	pH 6	Urine, saliva, milk
Neutral 10 ⁻⁷	pH 7	Tap water, blood
10 ⁻⁸	pH 8	Seawater, carrots, cabbage
10 ⁻⁹	pH 9	Baking soda, olive oil, celery
10 ⁻¹⁰	pH 10	Spinach, milk of magnesia
10 ⁻¹¹	pH 11	Ammonia
10 ⁻¹²	pH 12	Soaps
10 ⁻¹³	pH 13	Bleach, oven cleaner
10 ⁻¹⁴	pH 14	Drain cleaner

FIGURE 3-24 The pH scale. Shown here are pH values for some common substances and body fluids.

because most cells and body fluids (other than stomach acid) are neither acidic nor basic. For example, the pH of normal human blood is 7.4, a value that is critical for the health of an individual. If the pH of blood rises to 7.8 or drops to 7.0, serious illness or death can result, because the hemoglobin proteins in red blood cells can no longer bind and release oxygen efficiently. The body is able to maintain the correct pH of blood, despite the constant influx of nutrients with much higher or lower pH, because blood is buffered.

A **buffer solution**, a solution with a pH that does not change very much when H_3O^+ or OH^- ions are added, contains approximately equal amounts of a weak acid—that is, an acid that does not release all of its hydrogen atoms in solution—and its conjugate base. The dissociation of a weak acid (HA) in aqueous solution can be written as follows:



Buffers are present in biological fluids, and they can also be prepared in the laboratory. A typical buffer solution might consist of 0.1 mol of acetic acid (CH_3COOH) and 0.1 mol of sodium acetate (CH_3COONa) dissolved in 1 L of water. Such a solution will have equal concentrations of the weak acid acetic acid (CH_3COOH) and its conjugate base, the acetate ion (CH_3COO^-). This solution acts as a buffer because it neutralizes any H_3O^+ or OH^- that may be added to the solution, up to a certain limit (Figure 3-25). Any added H_3O^+ ions interact with the negatively charged acetate ions and are neutralized, whereas added OH^- ions are neutralized by the acetic acid molecules.

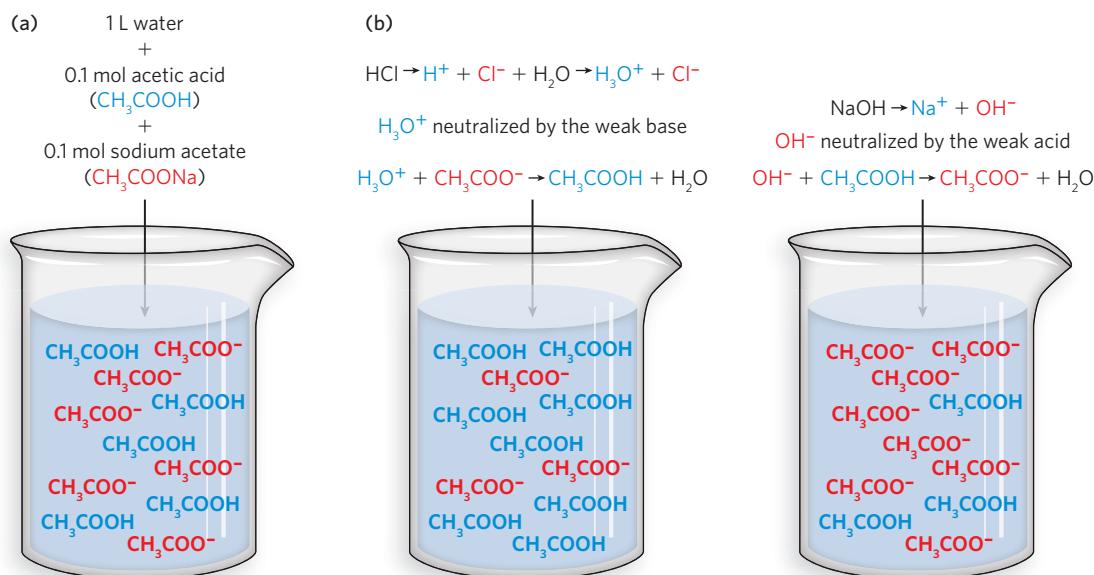


FIGURE 3-25 A typical chemical buffer system. (a) A buffer prepared with 0.1 mol of acetic acid (CH_3COOH) and 0.1 mol of sodium acetate (CH_3COONa) dissolved in 1 L of water has equal concentrations of the weak acid (CH_3COOH) and its conjugate base (CH_3COO^-). (b) When

a strong acid such as hydrochloric acid (left), or a strong base such as sodium hydroxide (right), is added to the solution, an excess of H_3O^+ or OH^- ions are introduced. The solution acts as a buffer by neutralizing the incoming H_3O^+ or OH^- ions, keeping pH constant.

Every buffer solution has two major properties: its pH value and its buffering capacity. The buffer solution has a characteristic value called its **acid dissociation constant (K_a)**, which equals the concentration of conjugate base multiplied by the concentration of H_3O^+ , divided by the concentration of weak acid. We can write this as follows:

$$K_a = [\text{A}^-][\text{H}_3\text{O}^+]/[\text{HA}] \quad (3-3)$$

where HA is the weak acid and A^- is its conjugate base. Just as we use pH to describe the H_3O^+ concentration on a log scale, we can define $pK_a = -\log K_a$. Each acid has a characteristic **p K_a** value that describes the ratio of charged (A^-) to neutral acid (HA) molecules at equilibrium in water. A p K_a is a measure of the tendency for an acid to lose its proton in aqueous solution. The lower the p K_a , the stronger the acid and the stronger its tendency to give up its proton. According to Equation 3-3, the pH of a buffer solution prepared by dissolving equal numbers of molecules (moles) of a weak acid and its conjugate base in water equals the p K_a of the weak acid. If $[\text{HA}] = [\text{A}^-]$, then $[\text{H}_3\text{O}^+] = K_a$ and $\text{pH} = \text{p}K_a$. Buffers work well within a pH range that is within one pH unit above or below their p K_a . Thus, buffer solutions can be prepared for almost any desired pH by selecting the appropriate acid. Molecular biologists often use buffers in the pH range of 6 to 8 to work with biological

molecules that fold and function near physiological (neutral) pH.

Note that a buffer solution has a limit to its ability to neutralize acid or base, beyond which its buffering action is overwhelmed. This **buffer capacity** depends on the total concentrations, rather than the ratio, of HA and A^- . For example, consider a buffer solution Y containing 10 times more molecules of acetic acid and its conjugate base than does buffer solution Z. Both solutions have the same pH (~5), but solution Y has 10 times the buffering capacity of buffer Z because it has 10 times more molecules available to neutralize H_3O^+ and OH^- ions.

The Henderson-Hasselbalch Equation Estimates the pH of a Buffered Solution

A defined relationship exists between the pH of a solution and the concentration of a weak acid dissolved in it. The relationship is defined by the Henderson-Hasselbalch equation:

$$\text{pH} = \text{p}K_a + \log [\text{A}^-]/[\text{HA}] \quad (3-4)$$

In other words, the pH of a solution containing a weak acid equals the p K_a of the acid plus the log of the ratio of base to acid concentrations. Using this equation, it is

possible to calculate the pH of a buffered solution. Or, if the solution pH and the concentration of weak acid are known, the pK_a of the acid can be determined.

This information can be very useful for working with biological samples, such as blood or proteins, where the pH of the solution must be carefully controlled to avoid destruction of the sample. Molecular biologists use the Henderson-Hasselbalch equation to prepare buffered solutions of a specific pH in the laboratory.

SECTION 3.5 SUMMARY

- Aqueous solutions can be acidic, neutral, or basic, depending on the concentration of hydronium ions (H_3O^+) present.
 - The pH value is defined as the negative logarithm of the hydronium ion concentration:
- $$pH = -\log [H_3O^+]$$
- A solution with a pH lower than 7 is acidic, and $[H_3O^+] > [OH^-]$; a solution with a pH greater than 7 is basic, and $[H_3O^+] < [OH^-]$. Because pH values are on a logarithmic scale, a change in pH of one unit corresponds to a tenfold change in hydronium ion concentration.
 - A buffer solution is a solution with a pH that does not change very much when H_3O^+ or OH^- ions are added. It contains approximately equal amounts of a weak acid and its conjugate base.
 - Each acid has a characteristic pK_a value, defined as $pK_a = -\log K_a$. The pK_a describes the ratio of charged (A^-) to neutral acid (HA) molecules at equilibrium in the solution.
 - The relationship between the pH of a solution and the concentration of a weak acid dissolved in it is described by the Henderson-Hasselbalch equation:

$$pH = pK_a + \log [A^-]/[HA]$$

3.6 Chemical Reactions in Biology

Life is possible because molecules in biological systems frequently undergo chemical reactions, enabling organisms to replicate DNA, synthesize RNA and protein molecules, pump small molecules into and out of cells, and use energy in the form of food or light. In this section, we discuss the physical principles governing chemical reactions. We review the fundamental laws of

thermodynamics and the roles of catalysts in accelerating the rates of reactions between biomolecules. Finally, we describe how energy, in the form of high-energy bonds, is harnessed to drive certain chemical reactions that would otherwise occur too rarely or too slowly to be useful to living systems.

The Mechanism and Speed of Chemical Transformation Define Chemical Reactions

Chemical reactions involve the breakage of covalent bonds and the formation of new bonds. Typically, chemical reactions are written with the **reactants**, or starting molecules, on the left and the **products** on the right, connected by an arrow indicating the direction of the reaction. For example, the expression



describes a very common reaction in biology: the breakage of a phosphorus–oxygen bond in the nucleotide adenosine triphosphate (ATP) to produce adenosine diphosphate (ADP) and inorganic phosphate (P_i). A more detailed representation of this reaction can be drawn to indicate, with curved arrows, the direction in which electrons are moving (Figure 3-26).

Reactions of this type involve the attack of a **nucleophile**, a strongly electronegative atom, such as oxygen or nitrogen, on a less electronegative atom, such as phosphorus or carbon. When the nucleophile initiating the reaction is part of a water molecule, as in Equation 3-5, the reaction is known as **hydrolysis**.

Most chemical reactions that take place in biological systems are not spontaneous; if they were, they would be impossible to control, and life as we know it could not exist. Instead, the starting and ending points of chemical reactions are bridged by an energy barrier, called the **activation energy**, that separates the reactants from the products (Figure 3-27). As reactants come together and bonds are breaking and forming, the reacting species progress through a high-energy state called a **transition state**. Once past this state, the reaction proceeds spontaneously because it is energetically favorable. Any chemical reaction can, in principle, proceed in the forward or reverse direction—toward products or reactants. In practice, however, most reactions characteristically tend to proceed more favorably in one direction than the other, due to differences in the relative energetic stability of products versus reactants. In Figure 3-27, the difference in energy between the product and the transition state is greater than the difference in energy between the reactant and the transition state. Hence, at equilibrium, there will be more products than reactants, because the energetic barrier

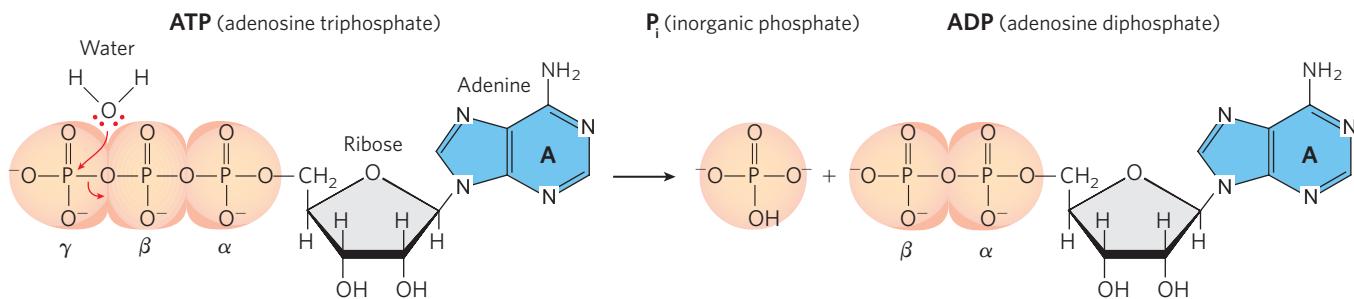


FIGURE 3-26 The hydrolysis of ATP. A phosphorus–oxygen bond in the nucleotide adenosine triphosphate (ATP) reacts with water to produce adenosine diphosphate (ADP) and inorganic phosphate (P_i).

to the reverse reaction is higher. (This is further discussed below.)

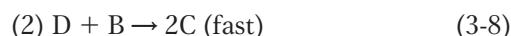
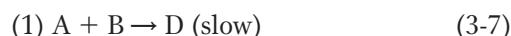
A **reaction mechanism** is the sequence of individual steps that take place during the conversion of reactants to products. It shows which bonds are breaking and forming and which species are **reaction intermediates**—substances that form and exist for an extremely short time before being converted to other intermediates or to reaction products. Understanding the mechanisms of chemical reactions that occur in living systems is important in molecular biology, because it helps us understand how these reactions are used

and controlled during such processes as cell growth and responses to chemical stimuli. For example, the mechanism used by the digestive enzyme chymotrypsin to break the covalent bonds in proteins involves binding of the enzyme to hydrophobic amino acids in the protein, followed by nucleophilic attack by the $-OH$ group of the serine in the enzyme’s catalytic center (Figure 3-28).

Although it is often difficult to determine such reaction mechanisms, understanding the **reaction kinetics**, the rates at which the reaction proceeds in the forward and reverse directions, can provide important clues. Reaction rates are, in part, a function of the concentration of reactants. Chemical reactions require collisions between molecules, and collisions occur more frequently when there are more molecules per unit volume. It is important to note that reaction rates also depend on the individual steps that make up the reaction. Usually, one step is slower than the others, and this slowest step, the **rate-limiting step**, governs the overall reaction rate. For example, suppose the reaction $A + 2B \rightarrow 2C$ is found, by experiment, to occur with a reaction rate proportional to the concentration of A multiplied by the concentration of B. We can write this as:

$$\text{Reaction rate} = k[A][B] \quad (3-6)$$

where k is a value called the **rate constant**, a property of the overall reaction that describes the tendency to react. A possible mechanism for this reaction might be:



Step 1, the rate-limiting step, produces an intermediate, D, that rapidly reacts with B in step 2 to yield product C. This reaction mechanism is not the only one consistent with the observed reaction kinetics, but it provides a hypothesis that can then be tested experimentally.

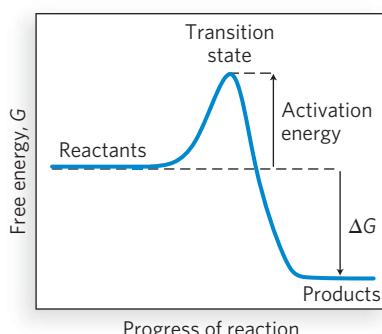


FIGURE 3-27 The activation barrier between reaction substrates and products.

The starting and ending points of chemical reactions are bridged by an energy barrier, called the activation energy, that separates the substrates—the reactants—from the products. If the difference in energy between the products and the transition state is greater than the difference in energy between the reactants and the transition state, then at equilibrium the amount of product will be greater than the amount of reactant because the energetic barrier to the reverse reaction is higher. The difference in free energy (ΔG) between the reactants and products is negative in the forward reaction, indicating that this reaction is energetically favorable. This type of diagram is called a reaction coordinate diagram (see Chapter 5).

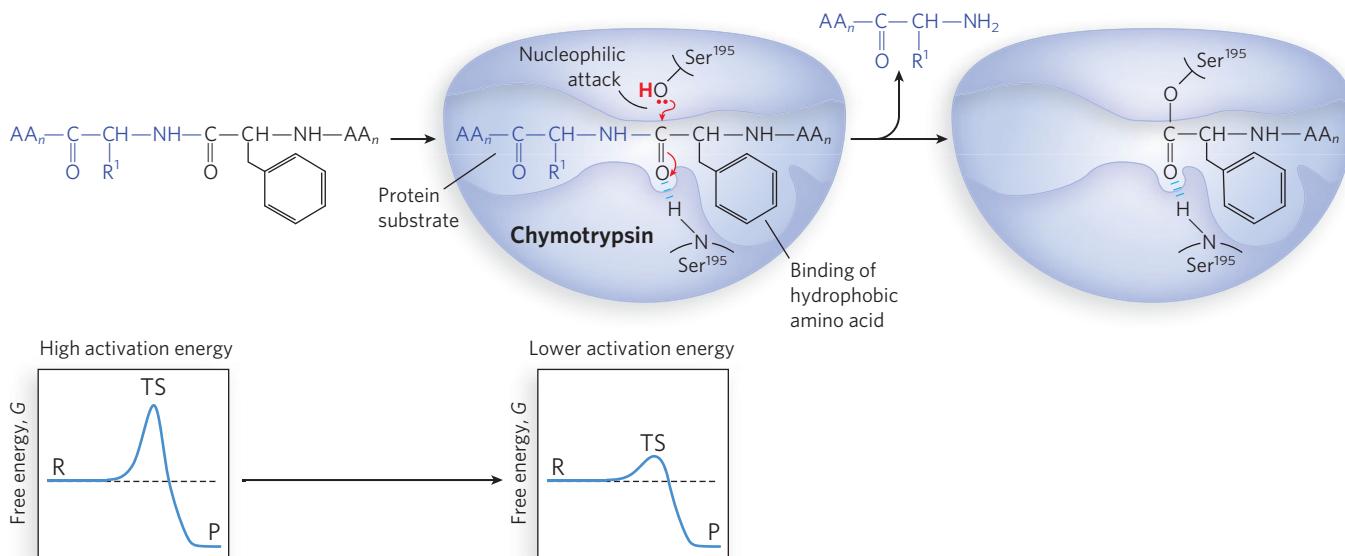


FIGURE 3-28 The catalytic mechanism of chymotrypsin. The mechanism used by the digestive enzyme chymotrypsin to break covalent bonds in proteins involves binding of the enzyme to hydrophobic amino acid residues in the protein substrate, followed by nucleophilic attack by the OH group of

the serine in the enzyme's catalytic center. The reaction coordinate energy diagrams show reactants (R), products (P), and the reaction transition state (TS) for one reaction step (see Figure 3-27).

Many biologically important reactions involve not just two but many individual steps. In the chymotrypsin example mentioned earlier (see Figure 3-28), the enzyme must bind to a protein (substrate) molecule, cleave one of the substrate's peptide bonds to produce a covalent enzyme intermediate, and bind to and remove a proton (H^+) from a water molecule and use the resulting OH^- to cleave the intermediate, thus releasing the product peptide and returning the enzyme to its original state. The identification of so many individual steps can be extremely challenging.

Biological Systems Follow the Laws of Thermodynamics

Living systems demand an almost constant input of energy, and as a result, organisms devote considerable molecular machinery to obtaining and using it. Biology obeys the physical laws of thermodynamics, which are the foundation for understanding energy and its effects on matter. In thermodynamics, a **system** is defined as a container or organism or other portion of the universe that is under study, and the rest of the universe, outside the system of interest, is called the **surroundings**.

The **first law of thermodynamics** states that energy can never be created or destroyed; in other words, the energy of a system is conserved. In a thermodynamic system, all readily occurring (spontaneous) processes take place without the input of additional energy from outside. The first law of thermodynamics cannot predict whether a process is spontaneous, however. For example, heat spontaneously transfers from a warmer object to a cooler one, never the reverse. Yet transfer in either direction is consistent with the first law of thermodynamics, because the total energy of the system remains unchanged in either case. To determine which direction of a process or reaction is spontaneous, we need additional criteria.

According to the **second law of thermodynamics**, all spontaneous processes take place with an increase in disorder, or **entropy**, of the system. Consider, for instance, two vessels of equal volume, one filled with plain water and the other with water containing a purple dye. When a connecting valve is opened, the dye molecules become randomly but equally distributed between the two vessels (Figure 3-29). The number of dye molecules in the two vessels becomes equal because the probability of any other distribution of the molecules is vanishingly small. Thus, the likelihood of all the dye molecules spontaneously remaining in the first vessel,

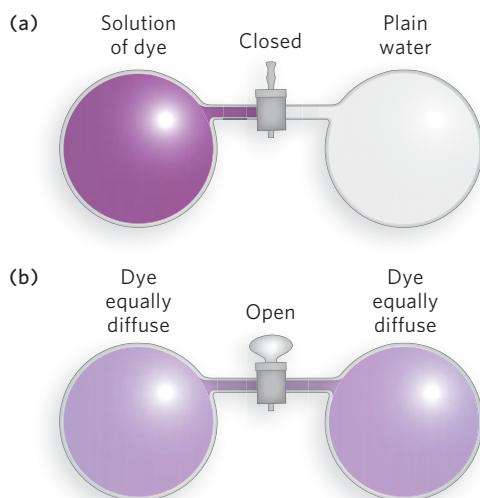
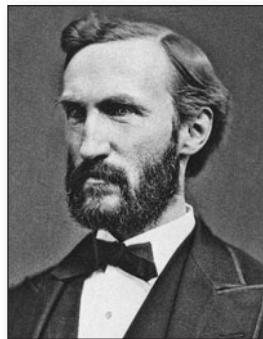


FIGURE 3-29 The spontaneous increase in disorder of a closed system. (a) Two vessels of equal volume are connected by a closed valve, one filled with plain water and the other with water containing a purple dye. (b) When the connecting valve is opened, the dye molecules become randomly but equally distributed between the two vessels over time, maximizing the disorder, or entropy, of the system.

or all the dye molecules moving into the other vessel, is essentially zero, even though the energy of either of these arrangements is no different from that of the evenly distributed molecules. However, the spontaneity of a process, such as a biochemical reaction, cannot be predicted from knowledge of the system's entropy change alone.

Every closed system in the universe tends toward equilibrium, a state in which the forward and reverse reaction rates are exactly balanced, and the approach to equilibrium is accompanied by a transfer of energy from one form to another. The concept of **free energy**

(**G**) provides a useful way to express this energy change. Free energy, denoted by the symbol **G** (for nineteenth-century physicist Josiah Willard Gibbs), is energy that is available to do work. The second law of thermodynamics states that free energy always decreases (i.e., the **free-energy change** (ΔG) is negative) in spontaneous reactions that occur without a change in temperature or pressure; at equilibrium, ΔG is zero. Free energy lost during the



Josiah Willard Gibbs, 1839–1903 [Source: Historical Pictures Service/Stock Montage.]

approach to equilibrium is either converted into heat or used to increase the amount of entropy.

The tendency of a chemical reaction to proceed to completion can be expressed as an equilibrium constant, which is related to the **standard Gibbs free-energy change** (ΔG°) of the reaction by the expression:

$$\Delta G^\circ = -RT \ln K_{\text{eq}} \quad (3-9)$$

where R is the universal gas constant, T is the absolute temperature, and $\ln K_{\text{eq}}$ is the natural logarithm of the equilibrium constant K_{eq} . The standard state (under which K_{eq} is measured) requires that all reactants be present at a concentration of 1 mol/L (1 M). For any compound, the ΔG° of formation (ΔG_f°) is the change in free energy that accompanies the formation of 1 mol of that substance from its component elements in their standard states (the most stable form of each element at 25°C and 100 kPa (kilopascals) of atmospheric pressure). Because K_{eq} can be measured experimentally, this relationship gives us a way of calculating ΔG° —the thermodynamic constant characteristic of each reaction—by plugging in the values of R (1.987 cal/mol • K) and T (298 at 25°C). When K_{eq} is much greater than 1, ΔG° is large and negative; reactions of this nature tend to go to completion. In contrast, when K_{eq} is much less than 1, ΔG° is large and positive; reactions with this property are not spontaneous and require energy to drive them to completion.

Catalysts Increase the Rates of Biological Reactions

As noted above, only when the ΔG of a chemical reaction is negative does the thermodynamic equilibrium favor the reaction. However, most biologically important chemical reactions, such as those required to form nucleic acids and proteins and to carry out many other cellular activities, have ΔG values that are positive. Such reactions do not occur spontaneously at significant rates under physiological conditions. Virtually every chemical reaction in a cell occurs at a significant rate only because of the action of enzymes, which, like all **catalysts**, are molecules that dramatically increase the rate of specific chemical reactions without being consumed in the process.

Catalysts function by lowering the activation energy for a particular reaction without affecting the reaction equilibrium. Given that a catalyst changes the reaction's rate but not its equilibrium, it must change the rate of the reverse reaction to the same extent as the rate of the forward reaction. Catalysts can do this because they enable the reaction to proceed by a

different mechanism than that of the uncatalyzed reaction. For example, enzymes bind to the transition state of the reactants by providing a molecular surface complementary to its shape and charge. Because of this favorable interaction, binding of an enzyme stabilizes the transition state, reducing the activation energy for the reaction and thus greatly enhancing the reaction rate. Additional contributions to catalysis occur when reacting molecules—substrates—bind to an enzyme in an orientation that favors the reaction and when chemical groups in the enzyme bind metal ions or protons that participate in the reaction. As a consequence of these effects, enzymes often increase reaction rates 10^{12} -fold or more above the rate of the uncatalyzed reaction.

Most cellular enzymes are proteins, though some RNA molecules also have catalytic activity. In general, each enzyme catalyzes a specific reaction, and each reaction in a cell is catalyzed by a different enzyme. Thus, many thousands of enzymes are required in each cell. Because enzymes are exquisitely capable of discriminating between reactants, and because they are subject to various regulatory mechanisms, cells can enhance reaction rates (or not) selectively. Such selectivity is critical for the effective control of cellular processes. By enabling specific reactions to occur at particular times and locations within a cell or organism, enzymes determine how chemicals and energy are channeled into biological activities. Enzyme function is described in detail in Chapter 5.

Energy Is Stored and Released by Making and Breaking Phosphodiester Bonds

The formation and breakdown of adenosine triphosphate (ATP) (and in some cases guanosine triphosphate, GTP) links the molecule-making and molecule-degrading pathways of cellular metabolism. The formation of this critical energy-storing molecule from inorganic phosphate and adenosine diphosphate (ADP), by the creation of a pyrophosphate linkage, is coupled to some of the steps of degradative metabolism and, in plants, to photosynthesis.

Because the hydrolysis of the pyrophosphate bond is exothermic under physiological conditions, energy is released when ATP is hydrolyzed to ADP. In turn, the free energy stored in the phosphodiester bonds of ATP is used to drive biosynthetic reactions of metabolism. Although energy is required for bond-breaking in ATP, the products of the reaction (ADP and phosphate) form highly favorable interactions

with water. Thus, hydration of the breakdown products of ATP more than makes up for the input energy necessary to break the bond in the first place, resulting in an overall energetically favorable process. Almost as soon as it is formed in concert with a coupled degradative reaction, ATP is consumed by enzymes to provide the energy necessary to propel another reaction to completion. In this way, ATP functions as a transient vehicle of intracellular energy transfer (Highlight 3-2).

ATP consists of an adenine nucleoside (base + ribose) and three phosphate groups. The phosphate groups, starting with the one directly bonded to ribose, are referred to as the alpha (α), beta (β), and gamma (γ) phosphates, respectively. Like other such high-energy molecules, ATP contains bonds—in this case, the phosphodiester bonds between the phosphate groups—that undergo breakdown by water (hydrolysis) to release significant free energy (see Figure 3-26). The second and third (β and γ) phosphate groups of ATP are unusually rich in chemical energy: the net change in energy (ΔG) upon hydrolysis of ATP to produce ADP and inorganic phosphate (P_i) is -12 kcal/mol inside a living cell. This large negative change in free energy makes the breakdown of ATP thermodynamically favorable and hence valuable for chemically storing energy that can be used to do work. The stored energy is captured when the cleaved phosphoryl group is transferred to another small molecule or protein as part of a metabolic pathway. Many enzymes harness ATP hydrolysis in this way to perform the work of the cell.

Note that a single hydrolytic reaction produces one P_i and ADP, which can be broken down further to yield a second P_i and adenosine monophosphate (AMP). But ATP can also be hydrolyzed to AMP directly, with the release of inorganic diphosphate, or pyrophosphate (PP_i). This latter reaction is effectively irreversible in the cellular environment, because enzymes called pyrophosphatases rapidly hydrolyze pyrophosphate into two phosphates. As a result, it would be very difficult to accumulate sufficient concentrations of pyrophosphate in the cell to drive the reaction in the reverse direction. Thus, the hydrolysis of ATP to AMP can be used to drive coupled processes in one direction.

For example, DNA and RNA are synthesized from nucleoside triphosphate precursors by phosphodiester bond formation that involves the release of PP_i . The required free energy of bond formation comes in part from the concomitant splitting of the high-energy pyrophosphate group (by pyrophosphatase) as it is

HIGHLIGHT 3-2 EVOLUTION

ATP: The Critical Molecule of Energy Exchange in All Cells

The universal role of ATP in cellular reactions has led many molecular biologists and chemists to question its origin and the reasons for its emergence as the universal mediator of energy exchange in cells. Experiments performed by Juan Oro, Stanley Miller, and Harold Urey in the 1950s and 1960s showed that adenine bases can be produced by heating concentrated hydrogen cyanide and ammonia, leading to speculation that adenine came into wide biological use in part because it arose very early in the evolution of life. The yields of adenine in the laboratory experiments ranged from ~1% to upward of 20%, depending on the reaction conditions; however, the likelihood of concentrated cyanide and ammonia existing in the environment of early Earth is uncertain. Adenine has been found in meteorites, though, providing evidence that it is produced naturally

in space. Some researchers have speculated that various inorganic clays could have helped sequester adenine and foster its reaction with ribose to form adenosine nucleosides (see Chapter 1, How We Know).



Juan Oro, 1923–2004 [Source: Courtesy of University of Houston.]

The uncatalyzed phosphorylation of nucleosides to nucleotides (i.e., of adenosine nucleoside to AMP, ADP, and ATP) has been observed under hot, dry conditions in the lab. To date, however, such proposed prebiotic reactions have not been found to proceed efficiently. More research is required to determine whether other synthetic conditions could suggest more plausible pathways for prebiotic accumulation of nucleotides.

released. Interestingly, when there is no pyrophosphatase around, such as in a laboratory test tube, robust DNA and RNA synthesis can still occur in this reaction. This is because the thermodynamics of phosphate bond formation and cleavage is only part of the story; base stacking and base-pair formation in polynucleotides are energetically favorable and therefore contribute some of the push toward polynucleotide synthesis. During protein synthesis in cells, amino acids are activated for peptide bond formation by linkage to AMP with the release of PP_i. Again, the splitting of the pyrophosphate group helps drive the reaction irreversibly toward formation of the activated amino acid-AMP (aminoacyl adenylate), the substrate for protein synthesis.

In addition to phosphorus–oxygen bonds, phosphorus–nitrogen and sulfur–carbon bonds also release significant free energy on hydrolysis, and these bonds are found in other important classes of energy-storing compounds that are used to drive reactions in biology. For example, the attachment of a phosphate group to the oxygen atom of a carboxyl group in the reaction noted above creates the high-energy acyl bond in the aminoacyl adenylate substrates for protein synthesis. The

high-energy sulfur–carbon (thioester) bond in acetyl-coenzyme A is the primary source of energy for fatty acid biosynthesis. The free energy released by hydrolysis of high-energy bonds, then, ranges in value, and its utility comes from the coupling of the released energy to another reaction to drive it forward. Thus, the coupling of biosynthetic reactions having a positive (unfavorable) ΔG value with the breakage of high-energy bonds that have a negative ΔG of greater absolute value ensures that the equilibrium favors synthesis of the biomolecule over its breakdown (Figure 3.30).

The overall free-energy change associated with a series of linked reactions determines whether a reaction in the group will occur. Reactions with small positive ΔG values, which in isolation would be unfavorable and would not take place, are often embedded in metabolic pathways in which they precede reactions having large negative ΔG values. It is critical to bear in mind that no single biochemical reaction, and no single pathway, takes place in isolation. Instead, the overall equilibria of the huge network of pathways within a cell are constantly adjusting to changing concentrations of substrates as the cell grows and responds to its environment.

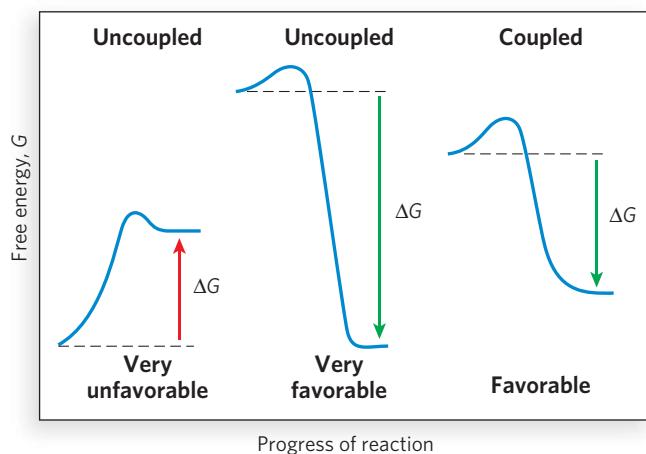


FIGURE 3-30 Energy coupling in biological processes. The first reaction coordinate diagram shows an energetically unfavorable biosynthetic reaction. The second shows a favorable reaction such as the hydrolysis of ATP. By coupling biosynthetic reactions having a positive (unfavorable) ΔG value with the breakage of high-energy bonds having a negative ΔG of greater absolute value, the overall reaction has a negative ΔG . At equilibrium, the coupled reaction favors synthesis of the biomolecule over its breakdown.

SECTION 3.6 SUMMARY

- Chemical reactions involve the breakage of covalent bonds and formation of new bonds. They are written with reactants on the left and products on the right, connected by half arrows (\rightleftharpoons) indicating that the reaction can proceed in either direction.
- The starting and ending points of chemical reactions are bridged by activation energy, a barrier that separates reactants from products. As reactants come together and bonds are breaking and forming, the reacting species progress through a high-energy transition state. Once past this state, the reaction

proceeds spontaneously because it is energetically favorable.

- Reaction rates depend on the individual steps in the reaction. Usually, one step is slower than the others, and this rate-limiting step governs the overall reaction rate.
- Chemical reactions obey the laws of thermodynamics. The first law states that energy can never be created or destroyed. The second law states that all spontaneous processes take place with an increase in disorder, or entropy, of the system.
- Free energy, G , is energy that can do work. According to the second law of thermodynamics, free energy always decreases (ΔG is negative) in spontaneous reactions that occur without a temperature or pressure change. The tendency of a chemical reaction to proceed to completion is expressed by an equilibrium constant, related to the standard free-energy change (ΔG°) of the reaction by the expression:

$$\Delta G^\circ = -RT \ln K_{eq}$$

- Virtually every chemical reaction in a cell occurs at a significant rate only because enzymes dramatically increase the rate of chemical reactions without being consumed in the process.
- Breakdown of the phosphodiester bonds between the phosphates of adenosine triphosphate (ATP) by reaction with water (hydrolysis) produces significant free energy, which can be used to change the structure or binding properties of enzymes and thus assist in catalyzing other cellular reactions. The phosphoryl groups are transferred from ATP to other metabolites or proteins in a coupled reaction, yielding new high-energy phosphate bonds that can be hydrolyzed to provide the free energy for further reactions.

How We Know

Single Hydrogen Atoms Are Speed Bumps in Enzyme-Catalyzed Reactions

Cha Y., C.J. Murray, and J.P. Klinman. 1989. Hydrogen tunneling in enzyme reactions. *Science* 243:1325–1330.

Hammes-Schiffer, S., and S.J. Benkovic. 2006. Relating protein motion to catalysis. *Annu. Rev. Biochem.* 75:519–541.

Understanding what limits the rates of biochemical reactions, and how enzymes speed them up, has long fascinated scientists. Thanks to a phenomenon called the kinetic isotope effect, researchers can deduce how single atoms affect reaction rates. Kinetic isotope effects are observed when different isotopes of an atom (such as of hydrogen or carbon), incorporated into a reactant, alter the rate of a chemical reaction. Substituting one isotope for another in a chemical bond that is broken or formed in the rate-limiting step will significantly change the observed reaction rate. This is exactly what happened when Judith Klinman and her colleagues initially studied the conversion of benzyl alcohol to benzaldehyde, a reaction catalyzed by yeast alcohol dehydrogenase (Figure 1). In Klinman's experiments, the rate constant measured for the substrate containing hydrogen (^1H) differed from the constants measured for the substrates containing deuterium (^2H) or tritium (^3H). The magnitude of

these effects indicated that the transfer of a hydrogen atom was the rate-limiting, or slowest, step of the reaction, and hydrogen transfer is the part of the reaction influenced by the enzyme.

In experiments of this type, the isotope-dependent rate changes are largest when the relative difference between isotope masses is maximized. This is because the effect results from changes in vibrational frequencies of the chemical bonds involved in the reaction. A deuterium atom (D) has twice the mass of a hydrogen atom, and a C—D bond reacts 6 to 10 times more slowly than the corresponding C—H bond, which provides an easily measurable difference.

These initial findings for alcohol dehydrogenase, and later for additional enzymes, led Klinman and other researchers to conclude that many enzymes enhance chemical reaction rates by speeding up the movement of hydrogen atoms in a quantum-mechanical process known as tunneling.

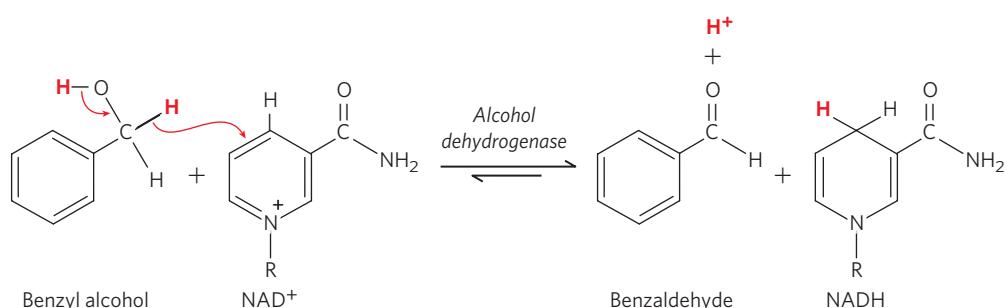


FIGURE 1 Alcohol dehydrogenase, an enzyme essential for metabolizing ethanol and other alcohols, catalyzes the conversion of an alcohol to an aldehyde. The reaction

uses a molecule called a cofactor (in this case, nicotinamide adenine dinucleotide, or NAD^+) as a proton acceptor.

Peptide Bonds Are (Mostly) Flat

Edison, A.S. 2001. Linus Pauling and the planar peptide bond. *Nat. Struct. Mol. Biol.* 8:201–202.

MacArthur, M.W., and J.M. Thornton. 1996. Deviations from planarity of the peptide bond in peptides and proteins. *J. Mol. Biol.* 264:1180–1195.

Pauling, L., R.B. Corey, and H.R. Branson. 1951. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. USA* 37:205–211.



Janet Thornton [Source: Courtesy of Janet Thornton.]

More than 50 years ago, Linus Pauling realized that the planar nature (flatness) of peptide bonds was an important constraint on polypeptides, leading him to the prediction of key elements of protein structure: the α helix and the β -pleated sheet (described in Chapter 4). But are peptide bonds truly flat? Two dominant resonance structures of the N–C bond, as measured in small molecules by spectroscopic methods, result in \sim 40% double-bond character, supporting the idea that peptide bonds and their covalently attached atoms lie in a plane (see Figure 3-13). But Pauling was working in the absence of any high-resolution protein structures, so the planarity of peptide bonds in real proteins couldn't be tested.

Today, the availability of thousands of protein and peptide structures makes it possible to conduct statistical surveys of peptide bonds in natural proteins. Janet Thornton and her colleagues showed that many such structures contain deviations from planar peptide bonds. Using a subset of available high-resolution protein structures, they estimated the energies of peptide bond rotation (Figure 2). This work revealed a small but statistically significant trend away from absolute planarity. Furthermore, previous experimental studies of

small peptides had shown that nonplanar peptide bonds do occur in both cyclic and linear peptides. Pauling realized this, of course! As a brilliant chemist, he wrote about the calculated low energetic barrier to small rotations about the peptide bond, which provided proteins with some flexibility—the extent depending on the structural environment of a particular segment of the polypeptide chain. Thus, theory, calculation, and experimentation all led to the same conclusion: the peptide bond is (mostly) planar.

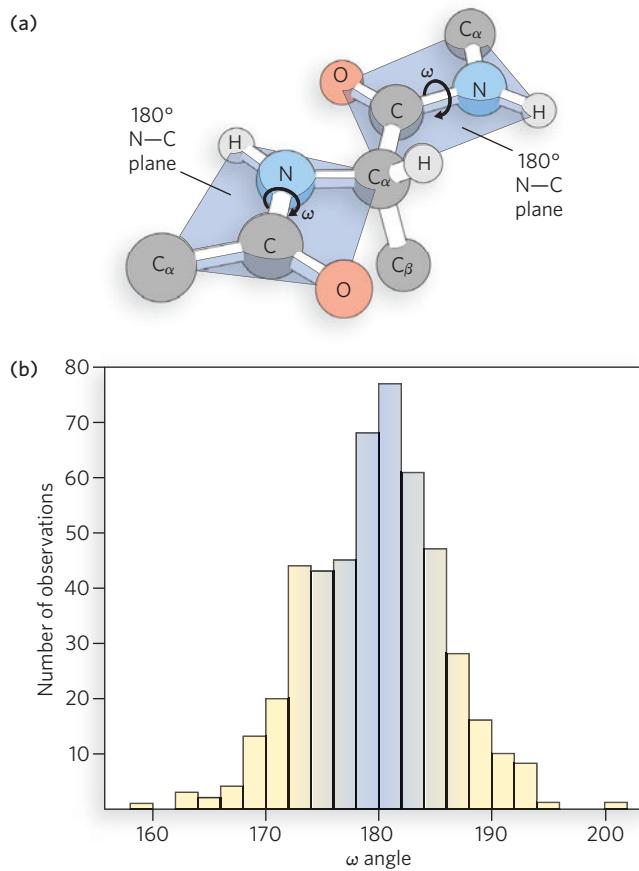


FIGURE 2 Peptide bond rotational energies and distribution of angular frequencies. (a) The angle ω represents the rotation of a bonded atom about the peptide bond; 180° is planar, because the bonded atoms point to opposite corners of a rectangle. (b) The histogram represents the angular frequency distribution of 237,807 ω values from coiled regions of 3,938 high-resolution protein structures in the January 2001 release of the Protein Data Bank. [Source: Adapted from M. W. MacArthur and J. M. Thornton, *J. Mol. Biol.* 264:1180–1195, 1996.]

Key Terms

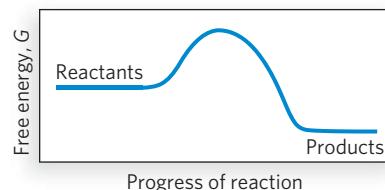
deoxyribonucleic acid (DNA), p. 62
 ribonucleic acid (RNA), p. 62
 nucleotide, p. 62
 nucleoside, p. 62
 deoxyribonucleotide, p. 62
 adenine (A), p. 62
 cytosine (C), p. 62
 guanine (G), p. 62
 thymine (T), p. 62
 ribonucleotide, p. 62
 uracil (U), p. 64
 amino acid, p. 64
 alpha carbon atom (α carbon, or C_α), p. 64

chemical bond, p. 68
 mole, p. 72
 van der Waals interaction, p. 74
 hydrophobic interaction, p. 75
 hydrogen bond, p. 76
 achiral, p. 78
 chiral, p. 78
 pH, p. 82
 buffer solution, p. 82
 pK_a , p. 83
 activation energy, p. 84
 transition state, p. 84
 reaction mechanism, p. 85
 reaction intermediate, p. 85

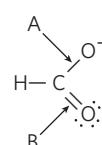
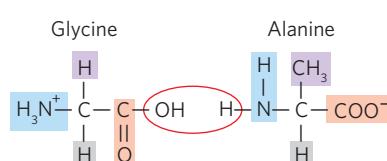
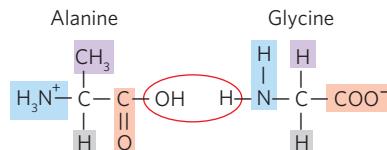
reaction kinetics, p. 85
 rate constant, p. 85
 first law of thermodynamics, p. 86
 second law of thermodynamics, p. 86
 entropy, p. 86
 free energy (G), p. 87
 standard Gibbs free-energy change (ΔG°), p. 87
 catalyst, p. 87

Problems

- Consider the O–O and the O=O bonds. Is the O=O bond stronger or weaker? Are the oxygen atoms in the O=O bond closer together or farther apart than in the O–O bond?
- Do two enantiomers of a chemical have the same density? The same melting point? If the chemical is an acid, do they have the same pK_a ?
- Which of the following statements about a catalyst is correct?
 - It can change the equilibrium constant of a chemical reaction.
 - It speeds up the rate of the forward but not the reverse reaction.
 - It is used up in the course of a reaction.
 - It lowers the activation energy for a reaction.
- A solution with pH 7 is 100 times more basic than a solution with a pH of what value?
- Amino acids are joined by peptide bonds, whose formation is accompanied by the loss of water. Is the dipeptide alanylglycine the same as the dipeptide glycylalanine? Why or why not? (Note that peptides are always written with the amino-terminal residue on the left.)
- For the reaction profile shown below, the activation energy is larger when the reaction proceeds in which direction?



- A flask contains 10 mL of salt water. If 10 mL of distilled water is added to the flask, does the number of moles of sodium chloride increase by 50%, decrease by 50%, or remain unchanged?
- Which law of thermodynamics explains why living things require the input of energy to maintain their ordered structure?
- One of the two resonance structures for a formate ion is shown below. Which carbon–oxygen bond, A or B, is longer?



- The activation energy for a chemical reaction can be determined in which of the following ways?
 - Measuring product amounts.
 - Measuring rates.
 - Calculating energy of bond hydrolysis.
 - Calculating change-in-entropy values.

11. What is the pH of the solutions with the following hydrogen ion concentrations?
- $1.75 \times 10^{-5} \text{ M}$
 - $6.50 \times 10^{-10} \text{ M}$
 - $1.0 \times 10^{-4} \text{ M}$
 - $1.50 \times 10^{-5} \text{ M}$

12. What is the hydrogen ion concentration of the solutions with the following pH values?
- 3.82
 - 6.53
 - 11.11

13. Calculate the pH of dilute solutions that contain the following molar ratios of acetate to acetic acid ($\text{p}K_a = 4.70$).
- 2:1
 - 1:3
 - 5:1
 - 1:1
 - 1:10

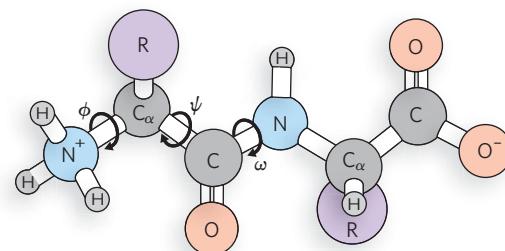
14. A buffer contains 0.01 mol of lactic acid ($\text{p}K_a = 3.60$) and 0.05 mol of sodium lactate per liter.
- Calculate the pH of the buffer.
 - Calculate the change in pH after 5 mL of 0.5 M HCl is added to a liter of the buffer.
 - Calculate the change in pH after the same quantity of this acid is added to a liter of pure water.

15. An unknown compound is thought to have a carboxyl group with a $\text{p}K_a = 2.0$ and a second ionizable group with a $\text{p}K_a$ between 5 and 8. When 75 mL of 0.1 M NaOH was added to 100 mL of a 0.1 M solution of this compound at pH 2.0, the pH increased to 6.72. Calculate the $\text{p}K_a$ of the second group.

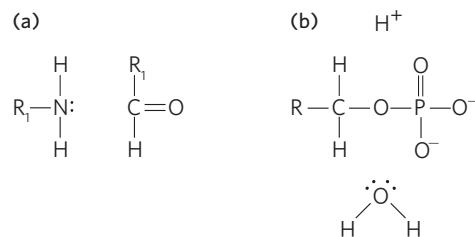
16. The base thymine (see Figure 3-1) contains a six-membered ring. From your understanding of bond

structures, is this ring flat/planar or bent? Explain your reasoning.

17. Free rotation is possible around single bonds, but not around double bonds. The repeating structure of the backbone of a polypeptide is shown below, as it is usually drawn. The bond torsion angles, describing rotation about these bonds, are labeled phi (ϕ), psi (ψ), and omega (ω). In reality, significant rotation can occur about just two of these bonds. For which bond is free rotation most restricted, and why?



18. The two sets of reactants shown below represent the starting points for (a) amide formation and (b) phosphoryl group transfer. In each panel, draw the curved arrows needed to indicate the first step in each reaction. Do not draw any additional intermediates or steps.



Additional Reading

Chemical Building Blocks of Nucleic Acids and Proteins

Adams, R.L., J.T. Knowler, and D.P. Leader. 2009. *The Biochemistry of the Nucleic Acids*, 11th ed. New York: Academic Press.

Saenger, W. 1984. *Principles of Nucleic Acid Structure*. New York: Springer-Verlag.

Chemical Bonds and Weak Chemical Interactions

Pauling, L. 1960. *The Nature of the Chemical Bond*. Ithaca, NY: Cornell University Press. A classic text covering the details of chemical bonding and the properties of molecules.

Pauling, L. 1988. *General Chemistry*. New York: Springer-Verlag. This text, also a classic, provides a great general introduction to the principles of chemistry.

Chemical Reactions in Biology

Fersht, A. 2005. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*. New York: Macmillan. A discussion of enzyme catalytic mechanisms from the standpoint of protein structure and folding.

Jencks, W.P. 1987. *Catalysis in Chemistry and Enzymology*. Mineola, NY: Courier Dover Publications. A clear and cogent discussion of catalytic mechanisms and experimental approaches to understanding how enzymes work.

Protein Structure



Steve Mayo [Source: Courtesy of Caltech.]

characteristic polypeptide fold that is held together by zinc ions. After many attempts, student Bassil Dahiyat finally generated a sequence called FSD1 that was predicted to form a zinc finger fold without requiring any zinc. He synthesized this peptide in the laboratory and late that evening analyzed it by circular dichroism, a method that measures the amount of secondary structure in a protein. We had made many unsuccessful attempts at protein design by this time, so we were very familiar with the CD [circular dichroism] spectra of unfolded proteins! At about midnight, Bassil called me at home and said, "Steve, you've got to see this spectrum!" On my home computer over an incredibly slow Internet connection, I watched as a gorgeous spectrum with exactly the shape expected for a folded protein came up on my screen. We realized at that moment that we had achieved something many had considered impossible. When we later solved the molecular structure of the peptide using NMR spectroscopy, the peptide had exactly the structure we had predicted.

—Steve Mayo, on his discovery of the first successful method for computational protein design

4.1 Primary Structure	97
4.2 Secondary Structure	103
4.3 Tertiary and Quaternary Structures	107
4.4 Protein Folding	115
4.5 Determining the Atomic Structure of Proteins	121

The beauty of the DNA double helix is indisputable, but to a trained eye, protein structures are even more compelling. Proteins have wonderfully complex architectures, sculpted over time to perform their tasks to near perfection. The fact that a protein adopts a unique conformation is amazing: despite the astronomical number of ways in which even a small protein could possibly fold, it folds into a single shape. The instructions for the unique shape of a protein are contained entirely within the linear amino acid sequence. Exactly how the folding instructions are encoded is still not understood; it remains the holy grail of the protein-folding field, given that the conformation of proteins is essential to their proper function.

Part of the explanation of how proteins fold lies in their reaction to an aqueous environment. Most proteins reside in the cell's aqueous cytoplasm, yet many amino acids are hydrophobic, or water-fearing. Hydrophobic residues, scattered throughout the length of a protein, tend to gather together, thus helping to fold the protein. In this way, proteins form highly compact molecules with hydrophobic interiors. The polar amino acids are oriented toward the outer surface where they may interact with water. The final, overall protein structure is held together by weak noncovalent forces, which include hydrophobic interactions, hydrogen bonds, ionic interactions, and van der Waals forces (see Chapter 3). As a consequence, proteins are only marginally stable and tend to unfold quite easily.

One might wonder why protein structures did not evolve to be more stable. In fact, thermophilic organisms—those that live at near-boiling temperatures—have very stable proteins. Why didn't evolution select for high stability in proteins for organisms living at lower temperatures? Interestingly, studies of proteins from thermophilic organisms provide an explanation: many proteins isolated from thermophiles are simply not active at 20°C to 40°C and require high temperatures for optimal activity. Thus, conformational flexibility must be important to the function of many proteins, and too much stability may compromise that flexibility.

Protein structure is commonly defined in terms of four hierarchical levels (Figure 4-1). Primary structure is essentially the sequence of amino acid residues. Secondary structure includes particularly stable hydrogen-bonded arrangements of amino acid residues that give rise to regular, repeating patterns. Tertiary structure includes all aspects of the three-dimensional folding pattern of the protein. And, in proteins that have two or more subunits, quaternary structure describes how the various subunits are arranged in space.

In this chapter, we explore how proteins are constructed, starting with the features of the peptide bond, which links amino acids together. Then we look at how weak forces mold protein chains into shape and discover that despite the bewildering array of different structures, all proteins contain only a few types of secondary structural elements. We'll also see that

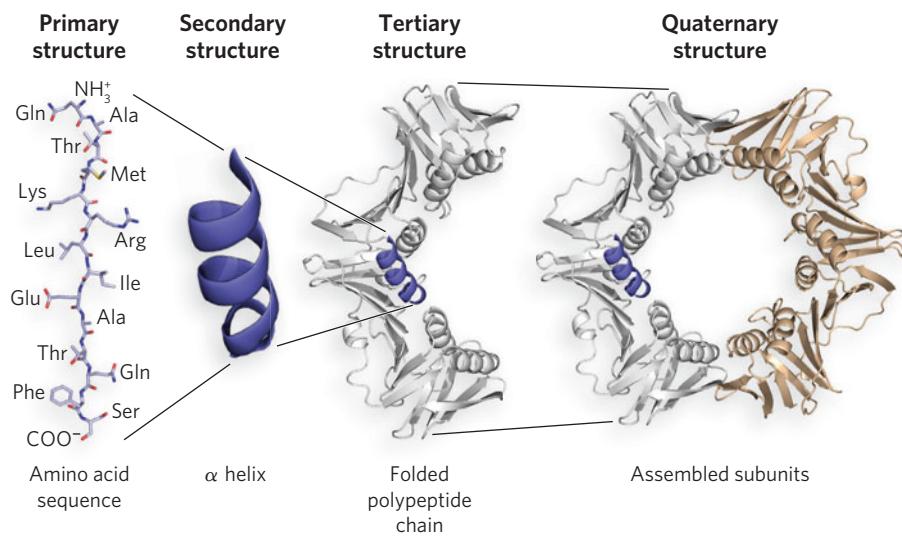


FIGURE 4-1 Levels of structure in proteins. The primary structure consists of a sequence of amino acids linked together by peptide bonds. The resulting linear polypeptide can be coiled into units of secondary structure, such as an α helix. The helix and other secondary structural elements fold

together and define the polypeptide's tertiary structure. The folded polypeptide shown here is one of the subunits that make up the quaternary structure of a multisubunit protein, the dimeric *Escherichia coli* β processivity factor, which is involved in DNA replication. [Source: PDB ID 2POL.]

there are some common ways in which these elements are stitched together to generate a diversity of folded proteins. A discussion of the two methods currently used to solve the atomic structure of proteins completes the chapter.

4.1 Primary Structure

The **primary structure** of a protein is the sequence of amino acids that make up the polypeptide chain. Many proteins range in size from 100 to 1,000 amino acid residues, although there are many examples of proteins that fall outside this range. In this section, we first examine the properties of the amino acids and take a close look at how amino acids are linked together, then examine how protein sequences hold information about their evolutionary heritage. We should note first that before a protein can be studied, it must be purified away from all other cellular proteins. Protein purification typically takes several fractionation steps. Particularly powerful techniques used to purify and analyze proteins include column chromatography and polyacrylamide gel electrophoresis, as summarized in Highlight 4-1.

KEY CONVENTION

The terms *peptide*, *polypeptide*, and *protein* are often used interchangeably. However, as generally defined, a peptide usually consists of a very short segment of 2 to 4 amino acids. A polypeptide usually consists of fewer than 100 amino acids, and “polypeptide chain” can refer to a polypeptide of any size. A protein is a large macromolecule that can be composed of one or more polypeptide chains.

Amino Acids Are Categorized by Chemical Properties

All amino acids have a central carbon atom, designated C_α (the α carbon), which is bonded to a hydrogen, an amino group, a carboxyl group, and a side chain called an R group (Figure 4-2). The **R group** distinguishes one amino acid from another and ranges from a simple hydrogen atom (in glycine) to relatively complex arrangements of carbon, hydrogen, nitrogen, oxygen, and sulfur. Side chains can be assorted into four groups according to their polarity and charge. The 20 most common amino acids found in proteins are shown in Figure 4-3 and Table 4-1; however, other, much less common amino acids sometimes occur in protein sequences. Amino acids are often abbreviated using a

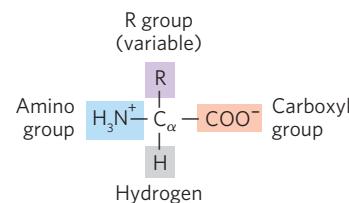


FIGURE 4-2 The general structure of an amino acid. The R group, or side chain, attached to the α carbon (the central carbon as shown here) is different in each amino acid. The name “amino acid” derives from the presence of both an amino group and a carboxylic acid group on the α carbon.

three-letter name or a one-letter symbol. Some amino acids have an R group that is ionizable, which may give it a positive charge when protonated and a neutral charge when unprotonated, or a negative charge when protonated and a negative charge when unprotonated. When the pK_a of the side-chain R group is lower than the pH of its surroundings, the group will be unprotonated (see Chapter 3).

Nonpolar, Aliphatic R Groups Aliphatic side chains are those composed only of hydrocarbon chains ($-\text{CH}_2-$), which are nonpolar and quite hydrophobic. Methionine, with a nonpolar thioether group, is also included here. These residues tend to cluster inside proteins and stabilize the structure through hydrophobic interactions. Glycine is also nonpolar, but having only a single hydrogen atom as its side chain, it contributes little to hydrophobic interactions. Proline has an aliphatic side chain, too, but more important is its rigid cyclic structure, which constrains and limits its possible conformations.

Polar, Uncharged R Groups Polar, uncharged R groups can interact extensively with water or with atoms in other side chains through hydrogen bonds. Recall from Chapter 3 that hydrogen bonds are interactions between a donor hydrogen atom that is covalently bonded to an electronegative atom, and an acceptor atom that usually has a lone pair of electrons. Examples of donor groups are the hydroxyl groups of serine and threonine and the sulphydryl group of cysteine. Asparagine and glutamine contain an amide group that can act as donor or acceptor. Two Cys residues brought in close proximity may be oxidized to form a **disulfide bond** (see How We Know).

Polar, Charged R Groups Three amino acids carry a positive charge at pH 7.0 (i.e., they are basic). Lysine contains a side-chain amino group, arginine has a

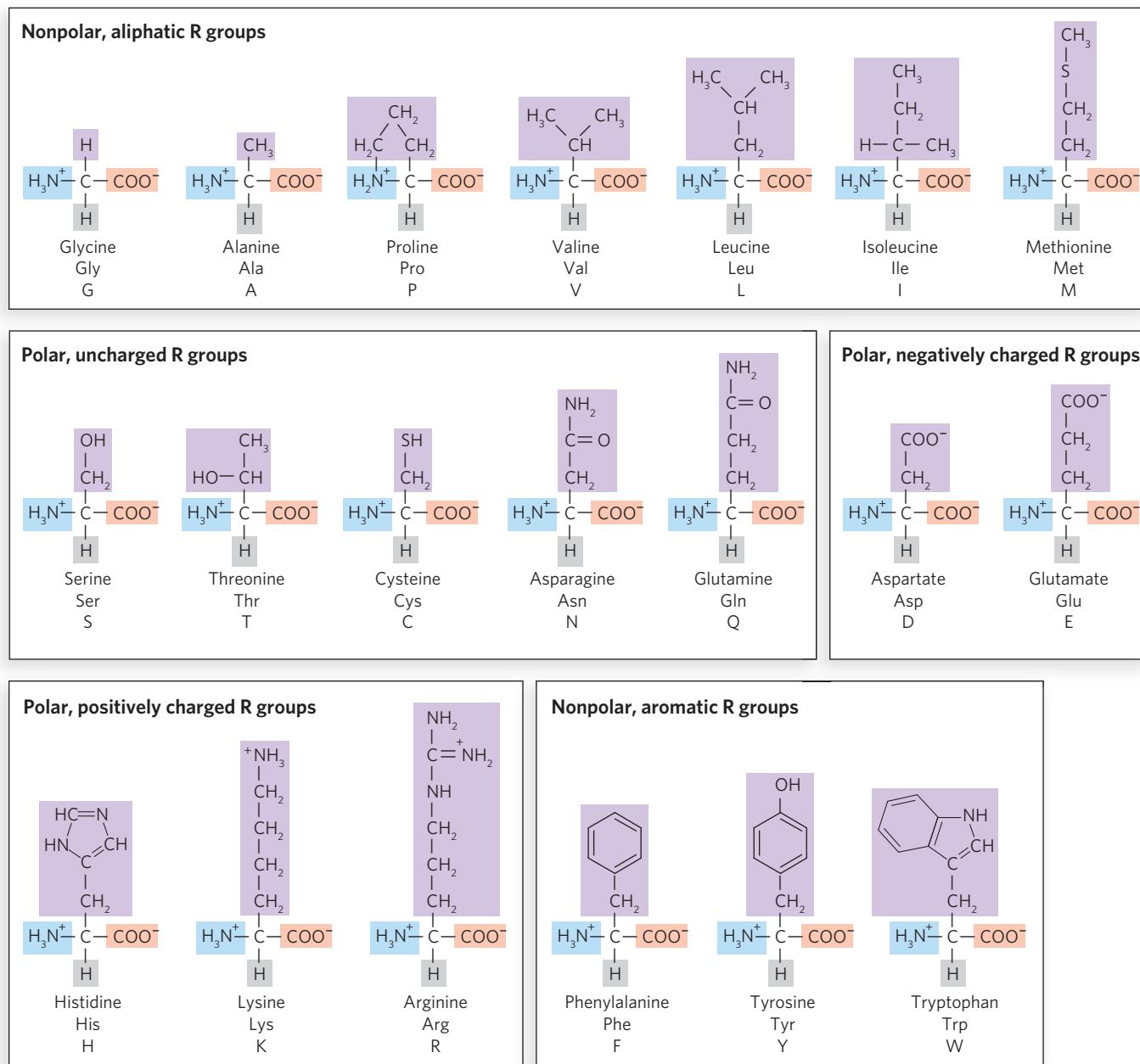


FIGURE 4-3 The 20 common amino acids. The structural formulas show the state of ionization at pH 7.0. Portions shaded pink, blue, or gray are common to all the amino acids;

the purple portions are the R groups. The R group of histidine is shown uncharged, but its pK_a is such that it carries a small but significant positive charge at pH 7.0.

guanidinium group, and histidine contains an imidazole group. The side chains of two amino acids, aspartate and glutamate, contain a carboxyl group and therefore carry a negative charge at pH 7.0 (i.e., they are acidic). Charged side chains can form hydrogen bonds and can form ionic interactions with amino acids of opposite charge.

Nonpolar, Aromatic R Groups Phenylalanine, tyrosine, and tryptophan contain aromatic side chains and therefore are hydrophobic. Phenylalanine is the most hydrophobic among them, whereas the tyrosine hydroxyl group and the tryptophan nitrogen can form hydrogen bonds and thus impart some polarity to these residues.

Table 4-1 The 20 Common Amino Acids

Name	Abbreviation	Symbol	M_r^*	pK _a values		
				pK ₁ (-COOH)	pK ₂ (-NH ₃ ⁺)	pK _R (R group)
Alanine	Ala	A	89	2.34	9.69	—
Arginine	Arg	R	174	2.17	9.04	12.48
Asparagine	Asn	N	132	2.02	8.80	—
Aspartate	Asp	D	133	1.88	9.60	3.65
Cysteine	Cys	C	121	1.96	10.28	8.18
Glutamine	Gln	Q	146	2.17	9.13	—
Glutamate	Glu	E	147	2.19	9.67	4.25
Glycine	Gly	G	75	2.34	9.60	—
Histidine	His	H	155	1.82	9.17	6.00
Isoleucine	Ile	I	131	2.36	9.68	—
Leucine	Leu	L	131	2.36	9.60	—
Lysine	Lys	K	146	2.18	8.95	10.53
Methionine	Met	M	149	2.28	9.21	—
Phenylalanine	Phe	F	165	1.83	9.13	—
Proline	Pro	P	115	1.99	10.96	—
Serine	Ser	S	105	2.21	9.15	—
Threonine	Thr	T	119	2.11	9.62	—
Tryptophan	Trp	W	204	2.38	9.39	—
Tyrosine	Tyr	Y	181	2.20	9.11	10.07
Valine	Val	V	117	2.32	9.62	—

* M_r values reflect the structures shown in Figure 4-3. The elements of water (M_r 18) are deleted during peptide bond formation, when the amino acid is incorporated into a polypeptide.

Amino Acids Are Connected in a Polypeptide Chain

The covalent link between two adjacent amino acids is called a **peptide bond**, and the result of many such linkages is known as a **polypeptide chain**. The peptide bond is formed by condensation of the α -carbon carboxyl group of one amino acid with the α -carbon amino group of another. Therefore, the linear sequence of a polypeptide chain has an **amino terminus**, or **N-terminus**, and a **carboxyl terminus**, or **C-terminus**.

KEY CONVENTION

When an amino acid sequence is given, it is written and read from the N-terminus to the C-terminus, left to right.

The C_α atoms of two adjacent amino acids in a polypeptide chain are separated by three covalent bonds: $C_\alpha-C-N-C_\alpha$. These bonds connect all the residues of a polypeptide chain and constitute the polypeptide “backbone.” Single bonds between atoms typically

allow free rotation, but not so for the peptide bond. Linus Pauling and Robert Corey's analysis of dipeptides and tripeptides by x-ray crystallography revealed that the atoms of a peptide bond lie in the same plane. Another key observation was that the C–N bond length (1.32 Å; 1 Å is 1×10^{-10} m) is significantly shorter than a single C–N bond (1.49 Å) and approaches the length of a C=N double bond (1.27 Å). These observations are explained by resonance, the sharing of electrons between the carboxyl oxygen and amide nitrogen, creating partial double bonds (Figure 4-4a; see also Chapter 3, How We Know).

Atoms are not free to rotate about a double bond. The partial double bond gives rise to two possible configurations, referred to as the cis and trans isomers. In peptide bonds, the trans isomer is favored about 1,000:1 over the cis isomer. The trans isomer of a peptide bond is one in which the two C_α atoms of adjacent amino acids lie on opposite sides of the peptide bond, as do the carbonyl oxygen and the amide hydrogen (Figure 4-4b). The double-bond character of the peptide bond explains why the atoms in a peptide bond lie in the same plane. Therefore, a chain of amino acid residues can be

HIGHLIGHT 4-1 A CLOSER LOOK

Purification of Proteins by Column Chromatography and SDS-PAGE

To study the structure of a protein, the researcher must first purify it from all other proteins in the cell. First, cells are lysed and particulate matter is removed by centrifugation, to yield a “crude extract.” The crude extract is then fractionated to separate the proteins and isolate the one that is of particular interest, a process known as **chromatography**. One of the most powerful chromatographic techniques is **column chromatography**, in which the protein mixture is applied to a column containing a resin, or matrix, that interacts differently with the various proteins (Figure 1). After the protein solution is applied, a buffer is passed through the column to thoroughly wash away any proteins that do not bind to the matrix. Then another buffer is applied that causes proteins to dissociate from the matrix; the proteins are carried out in the buffer flow, a process referred to as “elution” of proteins from the column. The proteins come off the column at different times, depending on how they interact with the resin. The column matrix and “elution buffer” are chosen so that different proteins dissociate from the matrix at different times. The eluted proteins are collected in a fraction collector, which gradually moves test tubes under the column, thus keeping the proteins that elute at different times separate from one another.

Several types of resin are used in column chromatography, which separate proteins based on different properties. Proteins can be sorted by charge in **ion-exchange chromatography**, in which the resin contains either cation groups (in a process called anion exchange) or anion groups (in cation exchange). Proteins are usually eluted from the column with an increasing gradient of salt solution, and their release depends on the nature of charged amino acid residues on their surface. Proteins are separated by size in **gel-exclusion chromatography**. The resin is composed of hollow beads with pores of a particular size; large proteins move around the beads and so elute earlier than smaller proteins that can enter the resin pores and thus take a longer path through the column. In **affinity chromatography**, proteins are sorted by the type of ligand they bind. A selected ligand is

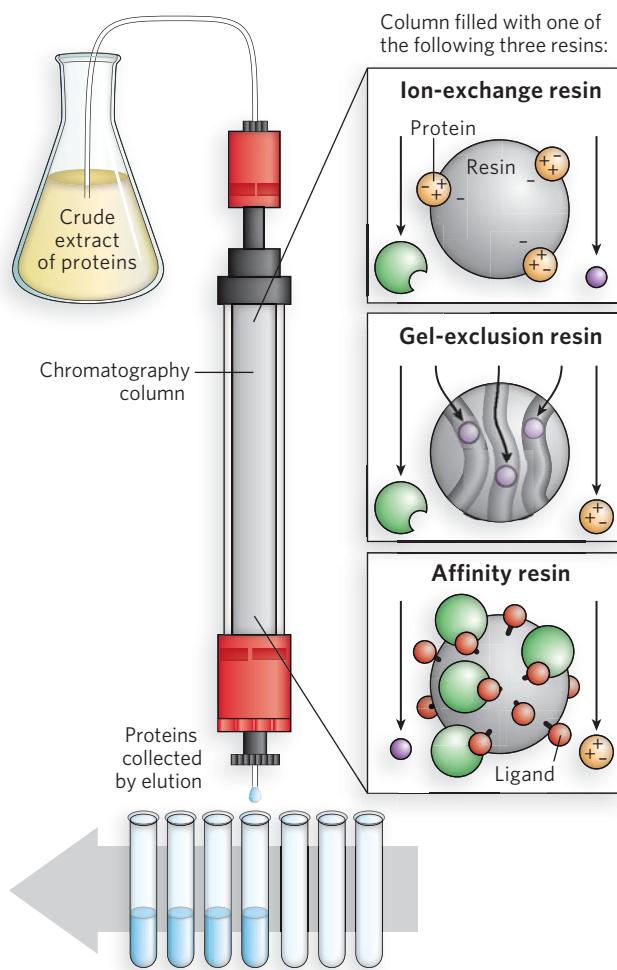


FIGURE 1 Column chromatography is performed in a glass or plastic tube containing one type of fractionating resin (matrix). The protein mixture is applied to the top of the column, and as buffer flows through, different proteins bind to the matrix according to the properties selected by the particular resin. These properties are typically the size or charge of the protein or the specific ligand to which the protein binds. Proteins are then dissociated from the column by eluting with a buffer that releases them at different times, and fractions are collected to keep the eluted proteins separate.

covalently coupled to the column resin, and the protein mixture is applied. Elution can be performed with a salt solution but is often done with a solution of the ligand itself, which binds to the active site of the protein, releasing it from the resin-bound ligand. Because ligand binding can

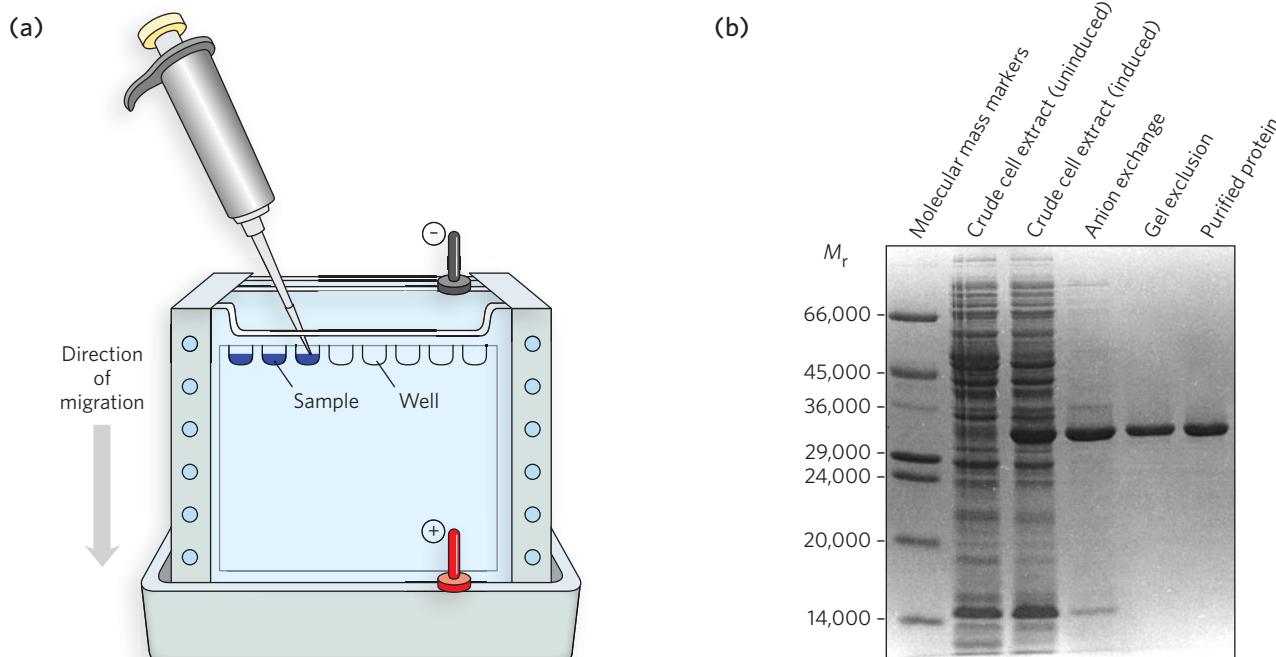


FIGURE 2 (a) In SDS-PAGE, the cross-linked gel is contained in a device to which an electric current can be applied, directing the proteins to migrate through the gel matrix. (b) A Coomassie Blue-stained SDS-PAGE gel,

tracking the gradual purification of glycine N-methyltransferase. [Source: (b) H. Ogawa et al., *Biochem. J.* 327:407–412, 1997.]

be very specific to a protein, this technique is often highly selective for the protein of interest.

After column chromatography, the fractions are analyzed for protein activity and visualized by **sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE)**. In this technique, a polyacrylamide gel is poured between two plates of glass in the presence of a cross-linking reagent, which quickly solidifies the gel (Figure 2a). The cross-linked gel acts like a sieve to sort proteins by size. The protein samples are treated with SDS, a negatively charged detergent that binds proteins and denatures them, giving all proteins in the sample a similar shape. And because SDS binds most proteins in relation to their size, it also gives all proteins a similar charge-to-mass ratio. Therefore, mixtures of proteins treated in this way separate in SDS-PAGE according to their relative mass.

The treated sample is applied to the top of the gel (which also contains SDS), followed by

application of an electric current, which pulls the charged proteins through the gel matrix. The gel is removed from the glass “sandwich” and soaked in an acidic buffer to precipitate the proteins, then it is treated with a dye that selectively binds to proteins. A common dye for this purpose is Coomassie Blue. Figure 2b shows a Coomassie Blue-stained SDS-PAGE gel containing protein samples taken at different stages of a protein purification. The rightmost lane in the gel shows only the subunits of the pure protein, glycine N-methyltransferase; samples taken earlier in the purification procedure show additional proteins. As described above, in SDS-PAGE, proteins separate according to their molecular mass. Proteins of known molecular mass are typically applied to one lane of the gel to serve as “molecular mass markers” (as in the leftmost lane in Figure 2b), which allows the researcher to estimate the mass of other proteins or standards in the gel.

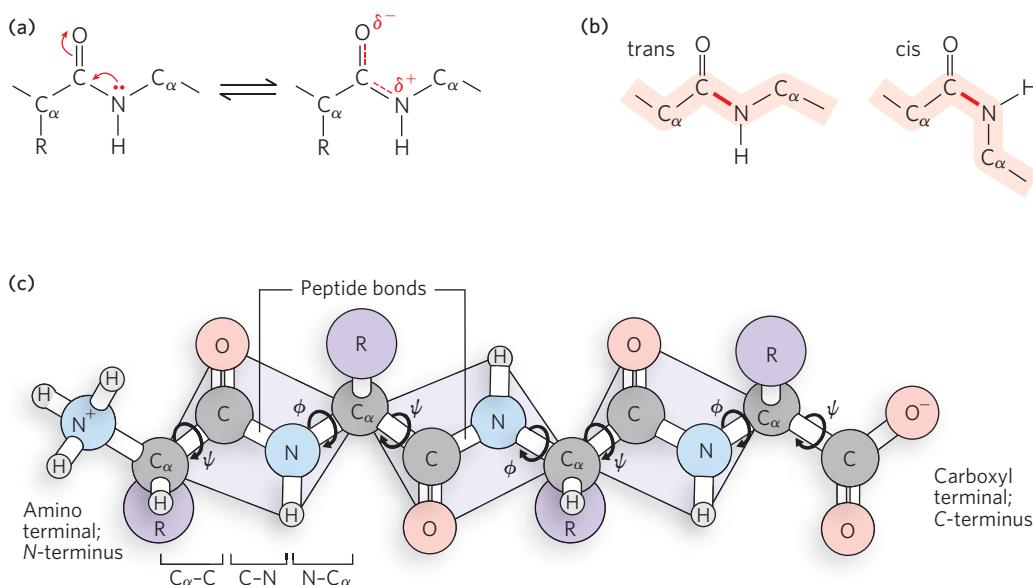


FIGURE 4-4 Peptide backbone atoms. (a) Resonance of the peptide bond gives it a partial double-bond character. (b) The cis and trans isomers of a peptide bond. The bonds in most proteins are trans. The peptide backbone is shaded

in orange and the peptide bond is in red. (c) The three bonds that separate sequential α carbons in a polypeptide chain lie in a plane. The N—C_α and C_α—C bonds can rotate, with torsion angles designated ϕ and ψ .

envisioned as a series of connected planes (Figure 4-4c). The C_α—C and N—C_α bonds are free to rotate. However, the angles between these bonds are constrained in a protein. These angles are referred to as torsion angles (or dihedral angles): ϕ (phi) for the N—C_α bond and ψ (psi) for the C_α—C bond.

KEY CONVENTION

Rotation around a double-bonded pair of atoms is restricted, placing the other atoms that adjoin them in one plane. Two atoms or groups adjoining the double-bonded atoms can lie either in cis (Latin for “same side”) or in trans (“other side”). The two forms are isomers because there is no difference between them other than their configuration. The amide hydrogen and carbonyl oxygen can be used to specify the cis and trans isomers of the peptide bond, as can the C_α atoms of adjacent amino acid residues. For example, in the trans isomer, the C_α atoms of adjacent amino acids lie on opposite sides of the peptide bond that joins them.

In reality, rotational movements are restricted, because the size of a bulky side chain may preclude a close approach to nearby atoms in the polypeptide

backbone. This “steric clash” between an amino acid side chain and neighboring atoms limits ϕ and ψ and thus the permissible orientations of one peptide-bond plane relative to another. G. N. Ramachandran developed a way to represent graphically the allowed values of ϕ and ψ for each amino acid. The **Ramachandran plot** for alanine is shown in Figure 4-5. The plots for most other amino acids look quite similar, with two exceptions. Glycine, which has a hydrogen-atom side chain, has a broader range of allowed angles, and the cyclic structure of proline greatly restricts its allowed range of conformations. Conformations deemed possible are those that involve little or no interference between atoms, based on known van der Waals radii and bond angles.

Evolutionary Relationships Can Be Determined from Primary Sequence Comparisons

As organisms evolve and diverge to form different species, their genetic material, at first, remains almost the same, but it differs increasingly as time passes. Therefore, proteins’ amino acid sequences can be used to explore evolution. The premise is simple. If two organisms are closely related, the primary sequences of their proteins should be similar, but they will diverge as the evolutionary

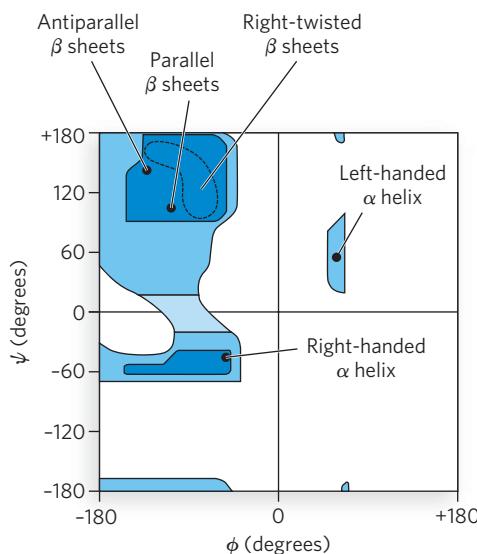


FIGURE 4-5 A Ramachandran plot: torsion angles between amino acids.

The conformations of peptides are defined by the values of ψ and ϕ for each amino acid residue. Allowable conformations are those that involve little or no steric hindrance between atoms of the amino acid side chain and nearby atoms of the peptide backbone. Shown here is the Ramachandran plot for Ala residues. Easily allowed conformations are in dark blue; medium blue signifies bond conformations that approach unfavorable values; light blue, conformations that are allowed if some flexibility is permitted in the torsion angles. Unshaded regions indicate conformations that are not allowed. With the exception of Gly and Pro residues, the plots for all other amino acid residues are very similar to this plot for alanine. The range of allowed ϕ and ψ values is characteristic for each type of secondary structure, as shown. Secondary structural elements are discussed in Section 4.2. [Source: Adapted from T. E. Creighton, *Proteins*, p. 166. © 1984 by W. H. Freeman and Company.]

distance between the organisms—that is, the time since they arose from a common ancestor—increases. The wealth of whole-genome sequences, from bacteria to humans, can be used to trace evolutionary lineages.

Amino acid substitutions, occurring through mutations, are not always random, and this opens up any analysis to interpretation. Some proteins have more amino acid variation among species than others, indicating that proteins evolve at different rates. At some positions in the primary structure, the need to maintain protein function limits amino acid substitutions to a few that can be tolerated. In other words, amino acid residues essential for the protein's activities are conserved over evolutionary time. Residues that are less important to function vary more

over time and among species, and these residues provide the information needed to trace evolution.

Protein sequences are superior to DNA sequences for exploring evolutionary relationships. DNA has only four different nucleotide building blocks, and a random alignment of unrelated sequences would produce matches about 25% of the time. In contrast, the 20 common amino acids used in proteins greatly lower the probability of such uninformative alignments. An example of how protein sequences can be used to trace evolutionary origins is presented in How We Know. Genomics, proteomics, and the use of sequences to study the molecular evolution of cells are discussed in detail in Chapter 8.

SECTION 4.1 SUMMARY

- The primary structure of a protein is its sequence of amino acids, along with any disulfide linkages between Cys residues.
- An amino acid consists of an amino group and a carboxyl group with a central carbon atom (C_α) between them. Also connected to C_α are a side chain and a hydrogen atom.
- There are 20 common amino acids, with characteristic side chains that differ in their chemical properties. Side chains can be charged or uncharged, polar or nonpolar, aliphatic or aromatic.
- Amino acid residues in a protein are linked by peptide bonds. The atoms of a peptide bond lie in one plane, due to the partial double bond between the carbonyl and amide groups, giving rise to cis and trans isomers. The trans isomer of the peptide bond is the most common in proteins.
- The planar configuration of the peptide bond limits how close the R groups of adjoining amino acids can approach one another. This leads to preferred, or allowed, torsion angles of the single bonds that connect the C_α atom to the carbonyl carbon ($C_\alpha-C$) and amide nitrogen ($N-C_\alpha$): angles ψ (psi) and ϕ (phi), respectively.
- Protein sequences reveal evolutionary relationships. The more similar the primary sequence between two proteins, the more recently they diverged from a common ancestor.

4.2 Secondary Structure

Secondary structure refers to regularly repeating elements within a protein, in which hydrogen bonds form between polar atoms in the backbone chain. These hydrogen-bonded structures allow the intrinsically

polar polypeptide chain to traverse the nonpolar interior of a protein. The main secondary structures are the α helix, typically 10 to 15 residues long, and the β sheet, composed of individual segments (called β strands) of 3 to 10 residues. A typical protein consists of about one-third α helix and one-third β sheet, although there are plenty of exceptions to this general rule, including proteins that have only one of these types of secondary structure. The portion of a protein that has neither α helices nor β sheets is composed of loops and turns that allow secondary structural elements to reverse direction back and forth to form a folded, globular protein. Here we describe the structure and properties of α helices and β sheets and briefly discuss the structure of reverse turns, which allow secondary structures to fold.

The α Helix Is a Common Form of Secondary Protein Structure

The α helix was originally predicted by Pauling and Corey in 1951, based on x-ray studies of keratin by William Astbury in the 1930s. The α helix contains 3.6 amino acid residues per turn (Figure 4-6a). One

full turn of the α helix is 5.4 Å (1.5 Å per residue) long, and the R groups protrude outward from the helix. The hydrogen on the amide nitrogen forms a hydrogen bond with the carbonyl oxygen of the fourth residue toward the N-terminus, which makes about one helical turn. The α helix forms a right-handed spiral, which, moving away from an observer looking down the spiral, corresponds to a clockwise rotation. You can determine the chirality of a spiral (i.e., whether right- or left-handed) using your hands (Figure 4-6b). With your fingers making a fist and your thumbs sticking out, a left-handed spiral would appear to curve in the same direction as the fingers on your left hand, in a counterclockwise rotation, as the spiral projects in the direction of your thumb. A right-handed spiral, such as the α helix, curves in the same direction as the fingers of your right hand, as the spiral projects in the direction of your right thumb.

All the hydrogen bonds of an α helix point in the same direction, and this sets up an electric dipole that gives a partial positive charge to the N-terminus and a partial negative charge to the C-terminus of the helix. Because the last four residues at either end of an α helix are not fully hydrogen-bonded, the dipole charges are

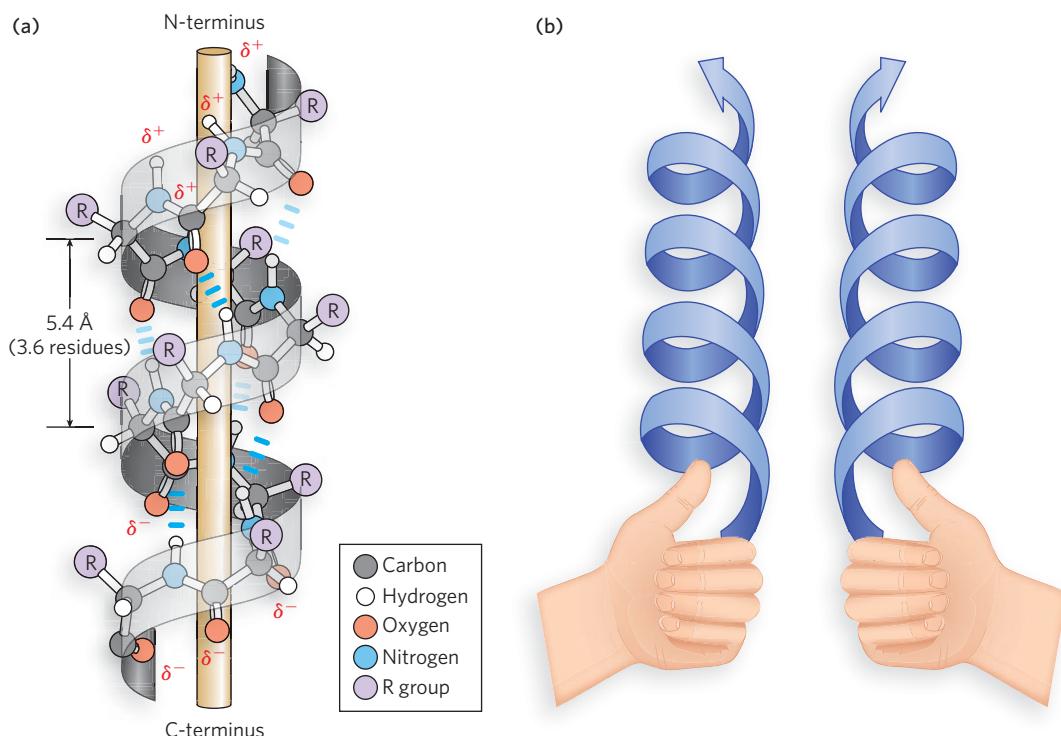


FIGURE 4-6 The structure of the α helix. (a) Peptide bonds are parallel to the long axis of the helix; intrachain hydrogen bonds are shown. The electric dipole of the helix, established through intrachain hydrogen bonds, propagates to the amino

and carbonyl constituents of each peptide bond. The partial charges of the electric dipole are indicated by δ^+ and δ^- . (b) An easy way to distinguish left- and right-handed helices (see text).

spread out on these residues. For this reason, the conformations at the ends of an α helix are often irregular or form a more strained version of the α helix with less favorable torsion angles.

Some general guidelines enable us to predict from a protein sequence the sections where an α helix will form. Consecutive stretches of amino acid residues with long or bulky R groups can't approach one another closely enough to form the tightly packed α helix. Also, polar side chains can hydrogen-bond to the peptide backbone, thereby destabilizing the helix. For this reason, serine, asparagine, aspartate, and threonine are found less frequently in α helices than most other amino acids. In addition, consecutive like-charged R groups repel one another in the close confines of the α helix. Glycine, due to its conformational flexibility, is also infrequently found in α helices. Finally, proline is infrequent in α helices because its cyclic structure lacks an amide hydrogen-bond donor and restricts N–C _{α} bond rotation. Proline is often referred to as a helix-breaking residue. The relative frequency of the 20 amino acids in different types of secondary structure is shown in [Figure 4-7](#).

Some configurations of amino acid residues can stabilize the helix. For example, side chains spaced four residues apart are stacked upon one another in

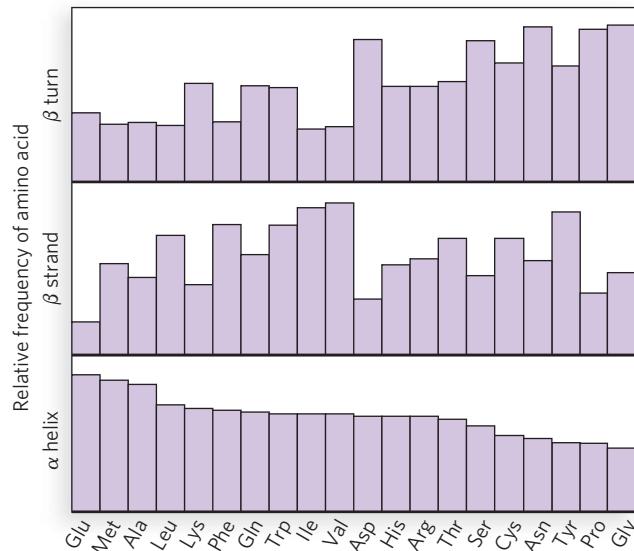


FIGURE 4-7 The relative frequency of amino acids in secondary structural elements. The plot shows the observed frequency of the 20 common amino acids in three types of secondary structure. [Source: Adapted from G. Zubay, *Biochemistry*, Macmillan, 1988, p. 34.]

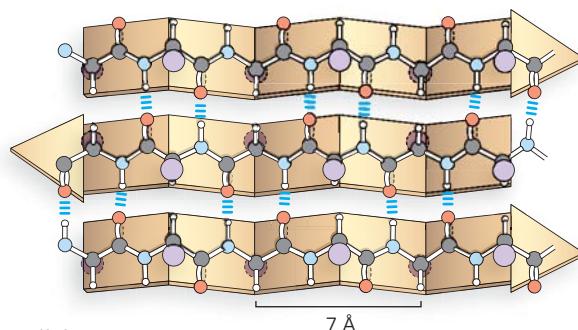
the helix. Oppositely charged side chains that are close together can form an ion pair, which stabilizes the helix. Likewise, aromatic side chains with this four-residue spacing can form hydrophobic interactions that stabilize the helix. Amino acids with a charge opposite to the partial charge of the helix dipole are sometimes located at the ends of a helix, which adds stability.

The diameter of the α helix is about 12 Å, similar to the width of the major groove in DNA (see Figure 1-3). For this reason, α helices are often found in proteins that bind to DNA.

The β Sheet Is Composed of Long, Extended Strands of Amino Acids

The **β sheet** consists of at least two β strands, and frequently it contains numerous such strands ([Figure 4-8](#)). Often, the many β strands that compose a β sheet are covalently connected in a single polypeptide. The β sheet, like the α helix, is formed by hydrogen bonds

(a) Antiparallel



(b) Parallel

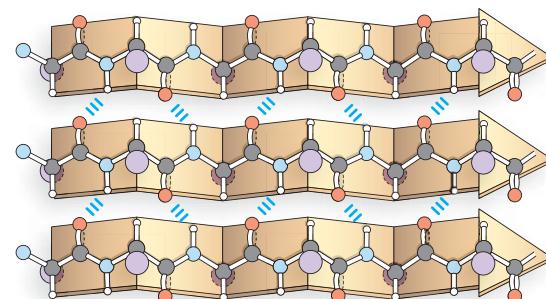


FIGURE 4-8 The structure of the β sheet. R groups extend out from the β sheet, emphasizing the pleated shape. Hydrogen bonds between adjacent β strands are also shown. (a) In an antiparallel β sheet, the N- to C-terminal orientation of the β strands alternates (shown by the arrowheads). (b) In a parallel β sheet, the β strands align in the same direction.

between backbone amide and carbonyl groups, but unlike the α helix, the β sheet structure cannot form from one β strand. Instead, all hydrogen bonds are formed between the backbones of two different β strands. The peptide bonds in a β sheet are arranged in a remarkably extended form, with a distance of 3.5 Å per residue. The R groups of adjacent amino acid residues in a strand lie on opposite sides of the sheet, and this alternating geometry prevents the interaction of R groups of adjacent residues. This sets up a zigzag pattern and, along with the side-to-side arrangement of the strand segments, resembles a series of pleats. Thus, the β sheet is often referred to as a “ β -pleated sheet.”

The strands of a β sheet may be close together in a polypeptide sequence, but they can also be far apart, separated by other secondary structures in the same polypeptide chain. The formation of β sheets between two different polypeptides is also common. When the β strands are oriented in the opposite N- to C-terminal directions, the structure is known as an **antiparallel β sheet**; when they run in the same direction, it is called a **parallel β sheet** (see Figure 4-8). Sheets can also be composed of a mixture of parallel and antiparallel strands. Antiparallel strands form nearly straight hydrogen bonds and are thought to be slightly more stable than parallel sheet structures. The β sheet structure can readily accommodate large aromatic residues, such as Tyr, Trp, and Phe residues. In addition, proline, which is unfavored in the α helix, is often found in β sheets, especially in the “edge” strands, perhaps to prevent association between proteins that are not meant to bind one another. The most common sheet structures are antiparallel, followed by mixed sheets, then purely parallel sheets. Because alternating R groups in β sheets are on opposite sides of the sheet structure, hydrophobic residues that alternate with polar side chains can yield a sheet structure that acts as a boundary between greasy and watery environments.

Reverse Turns Allow Secondary Structures to Fold

The size of α helices and β strands is limited by the diameter of a globular protein, and these structures must repeatedly reverse direction back and forth to form a properly folded protein. Approximately one-third of a polypeptide chain is composed of **reverse turns**, or loops, where secondary structural elements reverse themselves. Sometimes these turns are large and irregular, but many reverse turns are small and precise, and are called β turns. The β turn makes a com-

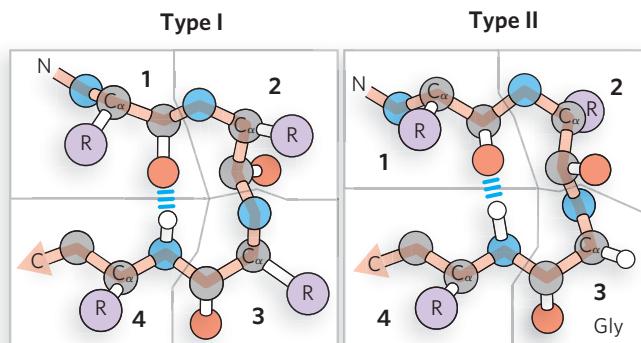


FIGURE 4-9 The linking of secondary structural elements by β turns. Different types of β turn are categorized based on torsion angles of residues 2 and 3. Type I and type II β turns are most common. Type I β turns have torsion angles of about -60° (ϕ) and -30° (ψ) at residue 2, and about -90° (ϕ) and 0° (ψ) at residue 3. Type II β turns have torsion angles of about -60° (ϕ) and 120° (ψ) at residue 2, and 80° (ϕ) and 0° (ψ) at residue 3. Note the hydrogen bond between residues 1 and 4. Individual amino acid residues are outlined, and the peptide backbone is shaded in orange. For clarity, only bonding atoms are shown.

plete chain reversal using only four residues, in which the backbone carbonyl oxygen of the first residue forms a hydrogen bond with the amide hydrogen of the fourth residue. The second and third residues form no inter-residue hydrogen bonds, and this typically places β turns on the surface of the protein, where the backbone can hydrogen-bond to water.

There are eight types of β turns, classified according to the torsion angles between the inner two amino acid residues. Two major types of β turns, type I and type II, are illustrated in Figure 4-9. Often, β turns contain a Pro residue at position 2 and Gly residue at position 3. Glycine is prevalent in β turns because its R group (hydrogen) allows it to accommodate many conformations that are not possible for other amino acids (see Figure 4-3). Although the conformation of a Pro residue is highly restricted compared with other amino acids, the imino nitrogen of proline can readily assume a cis configuration, a form that is especially suitable for tight turns. The type I β turn is the more prevalent of the two, occurring two to three times more often than type II β turns.

A much less common reverse turn is the γ turn, consisting of only three amino acids. The backbone carbonyl and amide groups of the first and third amino acid residues form a hydrogen bond, and the second amino acid of the γ turn is not involved in inter-residue hydrogen bonding.

SECTION 4.2 SUMMARY

- Secondary structure is a regularly repeating element that has hydrogen bonds between atoms of the peptide backbone.
- The α helix contains 3.6 amino acid residues per turn, and its internal hydrogen bonds set up a charged dipole for the entire helix.
- The compact structure of the α helix is unfavorable for certain combinations of residues, such as consecutive residues that are like-charged or bulky; the confined geometry of the proline ring can distort the helix, and proline is sometimes referred to as a helix breaker.
- The β sheet is formed by hydrogen bonding between two β strands of a polypeptide chain (or of separate polypeptide chains) and can have a parallel or antiparallel configuration.
- The extended structure of a β sheet can accommodate most amino acid residues, including bulky aromatic amino acids and proline.
- Secondary structural elements in a polypeptide are connected by reverse turns, which may consist of long, unstructured stretches of amino acids or tighter, precise structures of three or four amino acids (γ turns and β turns).

4.3 Tertiary and Quaternary Structures

Despite the use of only a few types of regular secondary structure, proteins exhibit a diverse spectrum of three-dimensional shapes, honed by evolution to perform their particular roles in the cell. And despite the great diversity of protein structure and function, general principles of protein shape apply. We concern ourselves here mainly with globular proteins in the aqueous environment of the cell. Globular proteins are roughly spherical, but contain sufficient irregularities to yield a surface area that is about twice that of a perfect sphere of equivalent volume. These nooks and crannies often form active sites or protein-protein interaction surfaces and are essential to protein function.

Knowledge of the tertiary and quaternary structures of a protein at atomic resolution offers a wealth of information and greatly helps our understanding of how the protein functions. Further, with a structure in hand, the biochemist can direct amino acid substitutions to defined architectural positions to test hypotheses about function. Atomic resolution of proteins that have medical relevance also enables the design of drugs aimed specifically at an active site.

Tertiary and Quaternary Structures Can Be Represented in Different Ways

The **tertiary structure** of a protein is defined by the three-dimensional orientation of all the different secondary structures and the turns and loops that connect them. The **quaternary structure** of a protein is defined by the connections between two or more polypeptide chains. A common quaternary structure is the association of two identical subunits. A protein consisting of two subunits is called a **dimer**.

The tertiary and quaternary structures of the *E. coli* DNA polymerase β subunit, a dimer of two identical polypeptides that encircles the bacterial DNA, is shown in Figure 4-10. The β subunit is part of the replication apparatus that duplicates the genome, and it holds the apparatus to the DNA (see Chapter 11). The figure shows six different representations of the β dimer. Each representation emphasizes one or more of the many structural features of a complete protein, and no single diagram can represent them all. In most of the representations, two colors differentiate the two identical polypeptide chains that comprise the complete protein. Figure 4-10a shows the van der Waals radius for each atom, but because there are thousands of atoms in the structure, only the surface of the protein is visible. Figure 4-10b shows the atoms as sticks, allowing us to view the inside, but it is still somewhat bewildering to view the protein all at once. Typically, this representation is used to study the structure of just a small section of a protein. Figure 4-10c shows a thick-line trace of only the α carbons in the backbone, which simplifies the structure considerably and gives a view of the overall architecture. Figure 4-10d is a ribbon diagram, in which β sheets are shown as broad arrows pointing in the N- to C-terminal direction, α helices as coils, and loops and turns as narrow tubes, with the polypeptides in different colors. The ribbon representation summarizes the secondary structural elements and overall architecture of the protein. Figure 4-10e is also a ribbon diagram, with the α helices and β sheets colored differently. Finally, Figure 4-10f is an electrostatic surface representation, showing the surface charges (red for acidic and blue for basic). The location of DNA-binding sites in proteins is often revealed by a basic patch in the electrostatic surface representation (note the blue area inside the β dimer).

Domains Are Independent Folding Units within the Protein

For a protein with 150 to 200 residues ($M_r \sim 20,000$), the polypeptide chain usually folds into two folding

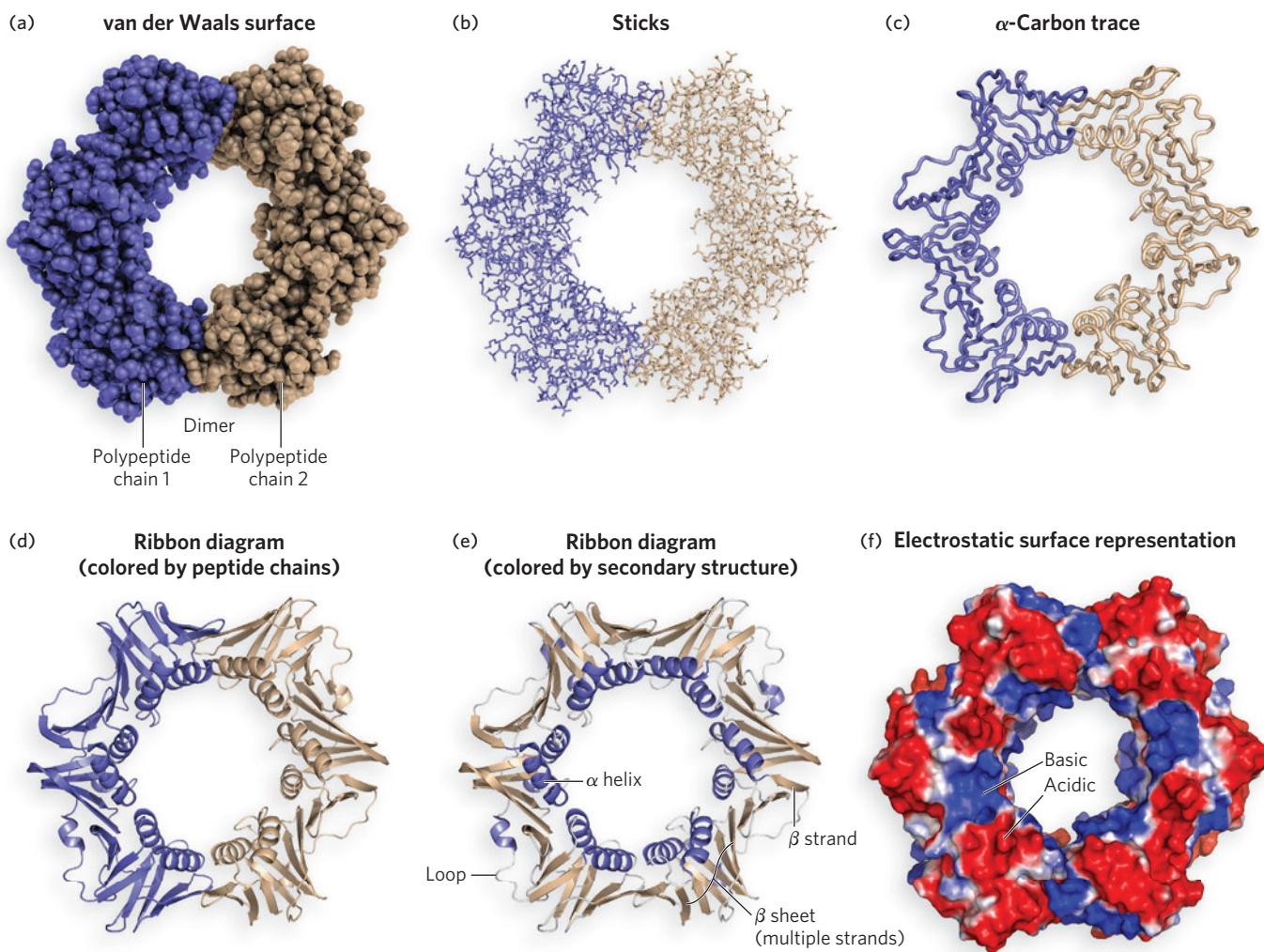


FIGURE 4-10 Different representations of tertiary and quaternary structure of the *E. coli* DNA polymerase β subunit. This protein dimer surrounds the *E. coli* DNA (not shown here). See the text for details of the six types of representation. [Source: PDB ID 2POL.]

units known as **domains**. The larger the protein, the more domains it usually contains. The secondary structures that comprise a domain are typically adjacent to one another in the primary sequence, although this is not always the case. A protein with two or more domains may perform a single function in the cell, but sometimes, different domains within one protein have different functions. Although domains are independent folding units, various domains in the same protein often interact.

Domains, then, can have independent functions. For example, the zinc finger domain is often used to bind DNA. (See Chapter 19 for more on zinc-binding domains.) An individual domain can also catalyze a reaction, such as a nuclease activity. There are also proteins that utilize several domains to perform a single

function. Structures of proteins containing one, two, three, and four domains are shown in Figure 4-11. Myoglobin (M_r 16,700) is a small, oxygen-binding protein that folds into a single domain. An example of a two-domain protein is γ crystallin (M_r 21,500), a component in the lens of the eye. DnaA (M_r 52,000), a bacterial protein involved in the initiation of DNA synthesis, contains three domains. A proteolytic fragment of *E. coli* DNA polymerase I (M_r 68,000), an enzyme that synthesizes new DNA strands, consists of four domains with two separate enzymatic activities. One domain comprises an exonuclease, which acts to proofread the product of the DNA polymerase and remove any mistakes. The other three domains cooperate to form the DNA polymerase. (DNA polymerase structure and function are described in Chapter 11.)

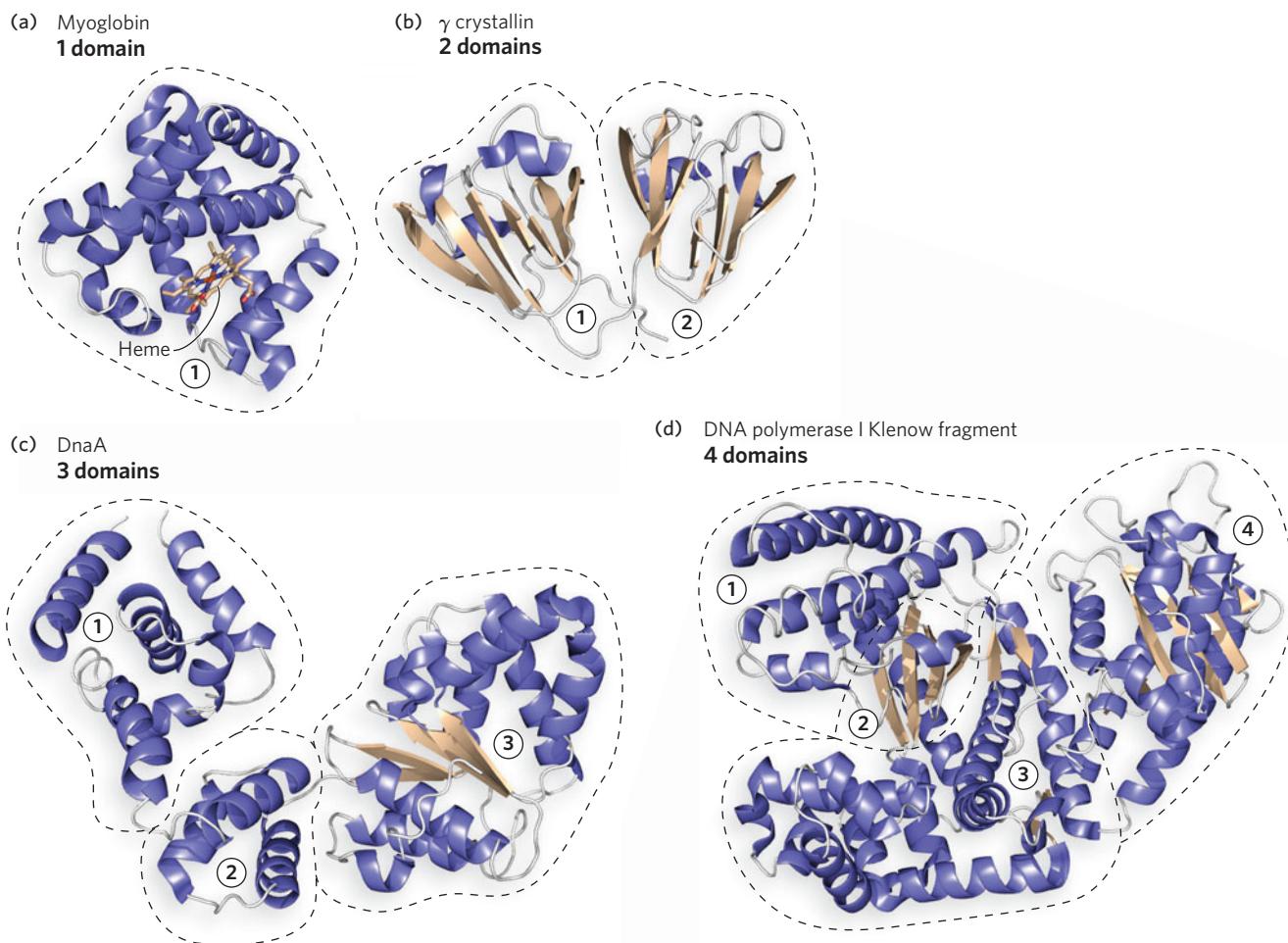


FIGURE 4-11 Domains. Proteins fold into one or more domains, depending on their size. (a) One domain: sperm whale myoglobin. (b) Two domains: human γ crystallin. (c) Three domains: *Aquifex aeolicus* DnaA. (d) Four domains: *E. coli* Pol I Klenow fragment; the DNA polymerase activity

is contained in three domains in the left two-thirds of the diagram, and the exonuclease activity is in the domain on the right side. [Sources: (a) PDB ID 2BLH. (b) PDB ID 1H4A. (c) PDB ID 2HCB. (d) PDB ID 1D8Y.]

Supersecondary Structural Elements Are Building Blocks of Domains

Particularly stable and common arrangements of multiple secondary structural elements are called **supersecondary structures**, also referred to as **motifs** or **folds**. Supersecondary structures are linked together to form sections of domains, or even whole domains. Some supersecondary structures are formed only of β sheets, and others have α helices and/or β sheets.

β Sheet Supersecondary Structure The smallest β sheet supersecondary structure is the **β hairpin**, in which two antiparallel β strands are connected, often by a β or γ turn (Figure 4-12a). In proteins that contain this motif, the β hairpin can be found alone or as a repeated struc-

ture that forms a larger, antiparallel β sheet. Whether they are parallel or antiparallel, β sheets tend to follow a right-handed twist (Figure 4-12b). When the strands of a β sheet contain hydrophobic R groups as every second residue, the groups lie on the same side of the sheet, thus facilitating layer formation in the folded state. For example, a β sheet with a hydrophobic surface may pack against the hydrophobic side of another sheet. A β sheet of eight or more strands, and with one surface that is hydrophobic, can form a cylinder in which the first β strand hydrogen-bonds with the last β strand. This structure, referred to as a **β barrel**, sequesters the hydrophobic side chains inside the cylinder (Figure 4-12c). In proteins with a β barrel, the barrel forms a complete domain. The simplified diagram below each supersecondary structure in Figure 4-12 shows the chain-folding

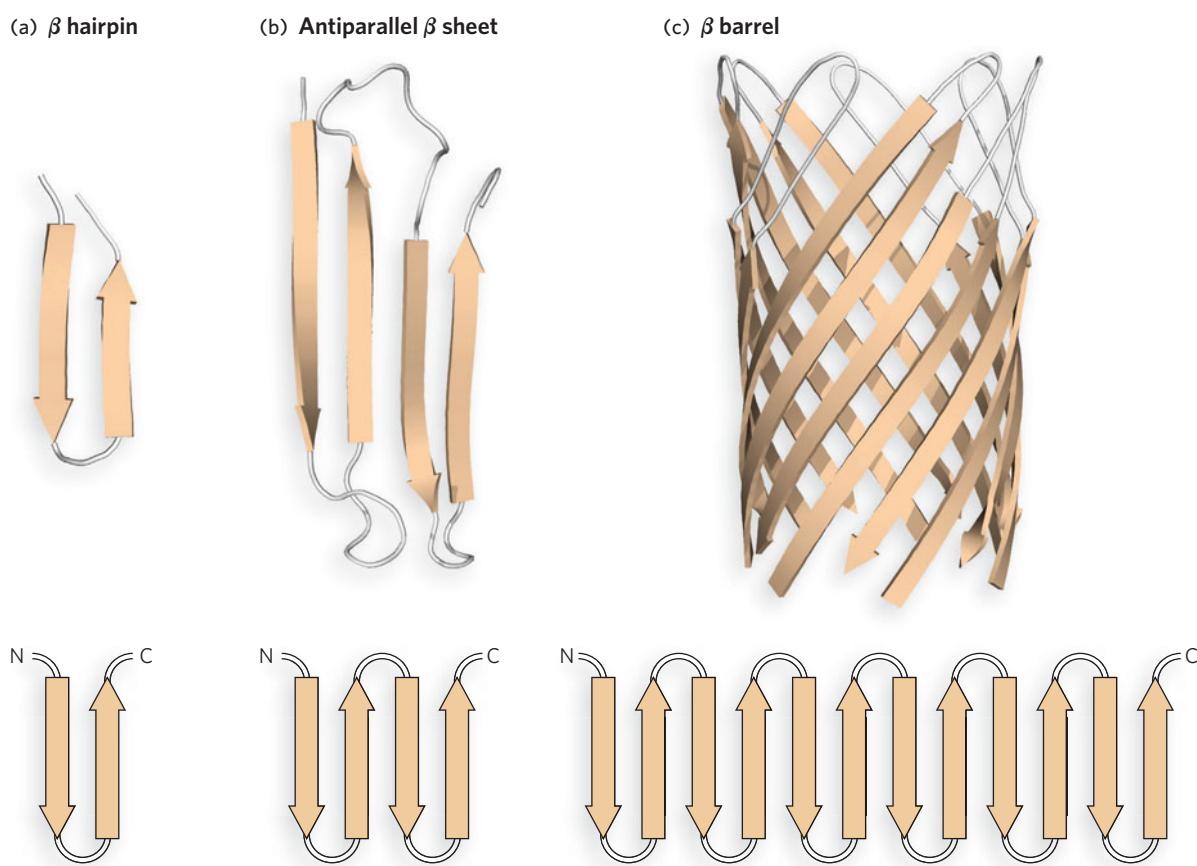


FIGURE 4-12 Supersecondary structures of antiparallel β sheets. (a) The β hairpin is an antiparallel β sheet composed of two β strands adjacent in the primary structure; the strands are often connected by β turns. (b) The β strands of antiparallel and parallel β sheets tend to have a right-handed twist (an antiparallel sheet is shown). (c) The β barrel is formed by the twist in a β sheet with a connection between the first and last strands. This example is a single domain of α -hemolysin (a pore-forming toxin that kills a cell by creating a hole in its membrane)

from the bacterium *Staphylococcus aureus*. The diagram below each structure shows the folding topology of the polypeptide chain. The β strands in a β barrel can have several different chain topologies; shown in (c) is an “up-and-down barrel,” reflecting the chain topology. In a simple β barrel, the strands would be connected as shown in the topology diagram, but the bottom strands of the particular β barrel shown here are not connected because the barrel is associated with additional domains (not shown). [Sources: (b, top) PDB ID 1LSH. (c, top) PDB ID 7AHL.]

pattern of the structure in two dimensions and is known as a **chain topology diagram**.

The **Greek key motif**, named for a design on Greek pottery, is another common β motif, consisting of four antiparallel β strands. One example is found in the OB-domain, which mediates DNA binding in many different types of proteins (Figure 4-13).

α Helix and/or β Sheet Supersecondary Structure In parallel sheets, the connection between adjacent strands requires a much longer linker. A basic unit of parallel β sheets is the **β - α - β motif**, which consists of two parallel β strands connected by an α helix. The β - α - β motifs can be stitched together by the α helix

linker in two different ways. In one, the α helix linker connects two β strands that are adjacent and hydrogen-bonded to each other (Figure 4-14a). This linkage forms a very common domain architecture called an **α / β barrel**, consisting of eight β strands surrounded on the outside by eight α helices (Figure 4-14b). The active site of an α / β barrel is almost always found on the loops at one end of the barrel.

The second way that β - α - β motifs are joined by an α helix is shown in Figure 4-15. This arrangement does not allow β sheets of adjacent β - α - β motifs to hydrogen-bond and therefore prevents circularization of the β - α - β motifs. The result is a domain with a central parallel β sheet that contains α helices on both sides of the

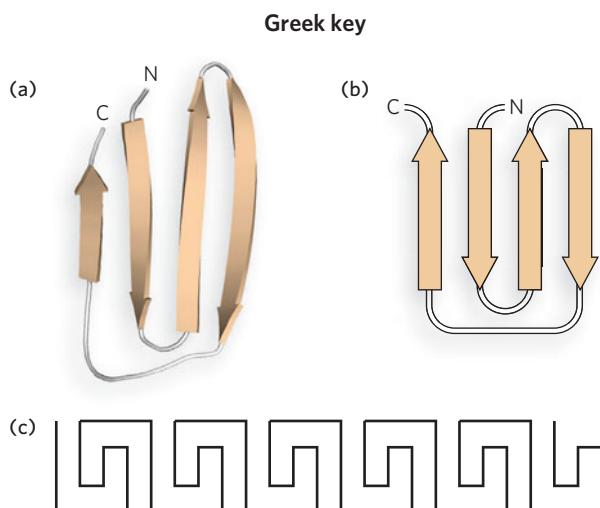


FIGURE 4-13 The Greek key motif. (a) A portion of the β subunit of *E. coli* DNA polymerase III adopts the Greek key conformation. (b) A chain topology diagram of the Greek key motif. (c) An example of the design found on Greek pottery from which this motif derives its name. [Source: (a) PDB ID 2POL.]

sheets. This architecture is commonly observed in ATP-binding or GTP-binding proteins and is sometimes referred to as a **Rossmann fold**. The active site in these proteins is usually located on the loops at the junction formed by the β sheets that are not directly connected by an α helix (see How We Know).

Several motifs have only α helices. One supersecondary structure that uses two α helices is the **helix-turn-helix motif**, sometimes found in proteins that bind specific DNA sequences (Figure 4-16a). It is commonly found in bacterial transcription factors (proteins that regulate gene expression), as well as in some eukaryotic transcription factors. In the **coiled-coil motif**, two α helices pack against each other at an angle of 18° and gently twist around one another in a left-handed supercoil (Figure 4-16b). The two α helices interact through hydrophobic contacts along the sides of each helix that form the coiled-coil interface.

Another common α helix supersecondary structure is the **four-helix bundle** (Figure 4-16c). This motif consists of four α helices; the way they interact depends on the protein. Sometimes the bundle is formed by antiparallel helices, and other times by parallel helices or a mixture of the two. Some four-helix bundles are even formed by dimerization of two different subunits, each of which contributes two α helices to the motif.

The interaction between α helices in a protein, whether in a four-helix bundle or not, occurs through hydrophobic surfaces that face one another. A useful way to visualize the alignment of residues along the edges of

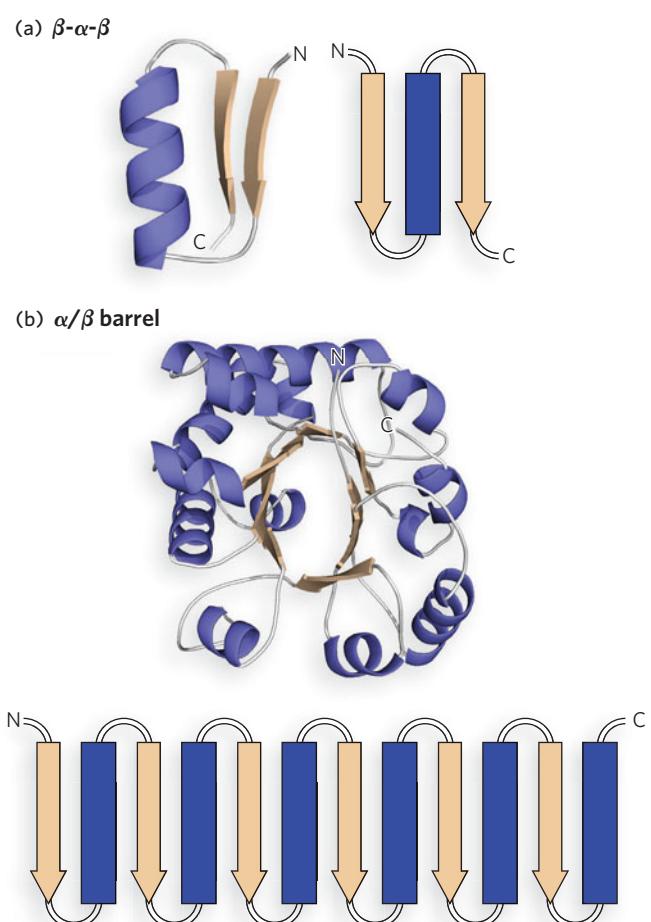


FIGURE 4-14 The $\beta\text{-}\alpha\text{-}\beta$ motif. (a) The $\beta\text{-}\alpha\text{-}\beta$ motif contains two parallel β strands connected by an α helix. (b) Four $\beta\text{-}\alpha\text{-}\beta$ motifs, interconnected by α helices, underlie the α/β barrel structure. This α/β barrel is from triosephosphate isomerase of *Trypanosoma brucei*. The topology diagram in (b) shows the general arrangement of β strands and α helices; the actual structure of the α/β barrel contains eight β strands connected by α helices. [Source: (b, top) PDB ID 4TIM.]

an α helix is the helical wheel, a two-dimensional representation of the residues in an α helix (Figure 4-17a). The α helix has a nonintegral number of residues (3.6) per turn, but a nearly integral number of residues (7.2) in two turns. For this reason, the helical wheel representation consists of seven positions along two turns of α helix, designated by letters *a* through *g*.

Residues that are spaced every two turns, or seven residues, lie on approximately the same side of the helix. This seven-residue spacing pattern of hydrophobic residues in an α helix is sometimes referred to as a heptad repeat, or a leucine repeat, because the residue occupying the *d* position is often leucine. Coiled-coils usually contain two heptad repeats, one set within the

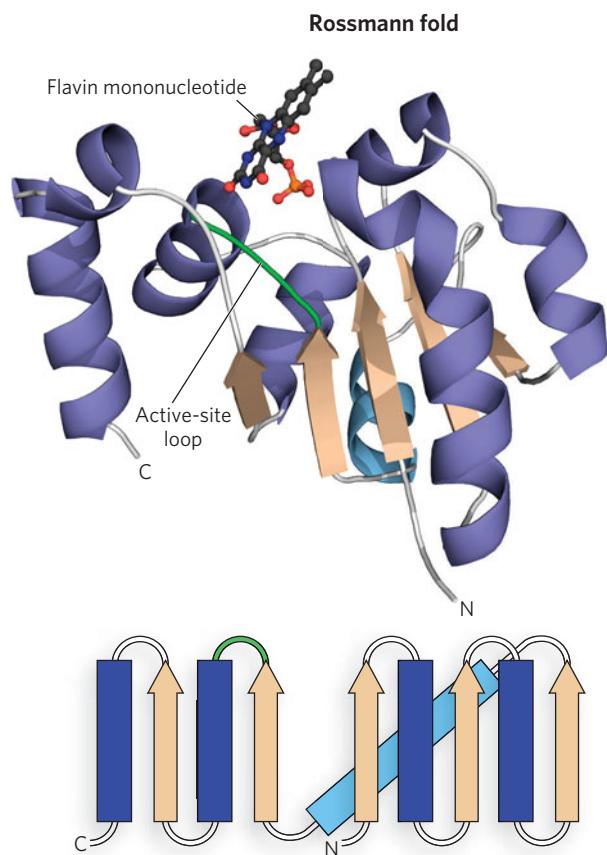


FIGURE 4-15 The nucleotide-binding Rossmann fold. The Rossmann fold, a common nucleotide-binding motif, is minimally composed of a β - α - β - α - β motif. This example is from a section of EpiD, a decarboxylase enzyme from the bacterium *Staphylococcus epidermidis*. The β - α - β motifs are connected as shown in the chain topology diagram at the bottom of the figure. In EpiD, the active site loop (shown in green) binds a flavin mononucleotide. [Source: PDB ID 1G5Q.]

other, with *a*-to-*a* and *d*-to-*d* spacing that is in contact with another helix with a similar arrangement of hydrophobic residues, as illustrated by the α -helical wheel diagrams in Figure 4-17b. This four-and-three hydrophobic repeat pattern is a hallmark of coiled-coils and forms a helix that is hydrophobic on one side and hydrophilic on the rest of the surface, which is referred to as an **amphipathic helix** (Figure 4-17c).

Some proteins, such as keratin, are long, extended, fibrous proteins in which the main structural element is a very long coiled-coil. However, there are many examples of globular proteins with short coiled-coils that mediate dimer formation. For example, the common **leucine zipper motif** in some eukaryotic transcription factors consists of four to five heptad repeats of leucine. These globular proteins dimerize by forming coiled-coils, and in some cases form dimers with one another. Dimerization

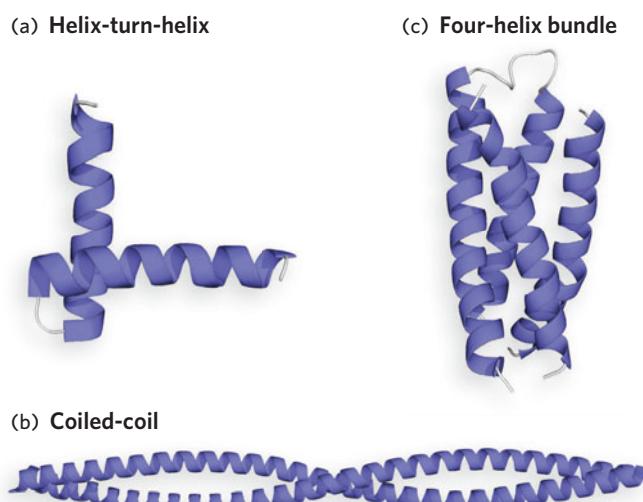


FIGURE 4-16 Supersecondary structures of α helices.

(a) The helix-turn-helix motif is often used in DNA-binding proteins. (b) The coiled-coil motif consists of two α helices twisted in a left-handed supercoil. (c) The four-helix bundle is formed by the interaction of four α helices. [Sources: (b) PDB ID 1D7M. (c) PDB ID 1UM3.]

requires the leucine heptad repeat and is important to the ability of these transcription factors to bind DNA.

Quaternary Structures Range from Simple to Complex

A protein composed of multiple polypeptide chains is referred to as an **oligomer**, or multimer, and the individual polypeptide chains are referred to as subunits, or **protomers**. An oligomer with identical subunits is referred to as a **homooligomer**, while an oligomer with nonidentical subunits is a **heterooligomer**. The quaternary structures of three proteins are shown in Figure 4-18. An example of a homodimer (i.e., with two identical subunits) is *E. coli* cAMP receptor protein (CRP), also called catabolite gene activation protein (CAP) (M_r , 22,000). Each CRP subunit has two domains; one domain binds DNA and the other binds the cyclic nucleotide cAMP. In the homodimer, the CRP protomers interact through a coiled-coil, and the two DNA-binding domains are oriented adjacent to each other (Figure 4-18a). An example of a homotrimer is eukaryotic PCNA, in which the three subunits (subunit M_r 29,000) are joined by an intermolecular β sheet and by helix-helix packing (Figure 4-18b). PCNA functions like the *E. coli* β subunit, anchoring the replicating apparatus to DNA. Hemoglobin is a well-studied example of a heterooligomer (tetramer M_r 64,500). Hemoglobin contains two α protomers and two β protomers (Figure 4-18c).

Some oligomers are made up of numerous subunits. One example is the ribosome, a large, multiprotein

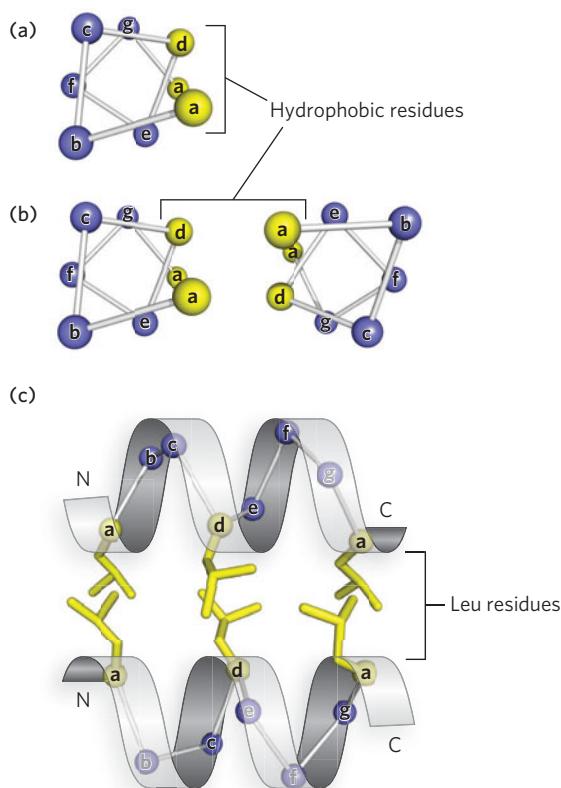


FIGURE 4-17 The helical wheel and heptad repeats.

(a) The seven residues that comprise two turns of the α helix can be represented by a helical wheel and are labeled *a* through *g*. (b) Hydrophobic residues in positions *a* and *d* lie on the same side of the helix, creating an amphipathic helix, and form the interface with a second helix to form a coiled-coil. (c) The packing of *a* and *d* residues in two α helices in a coiled-coil. Here, the hydrophobic residues are shown as leucines.

oligomer that contains many nonidentical subunits, often as a single copy, as well as functional RNAs—all of which, together, form a molecular machine that translates a nucleic acid sequence into a protein sequence. The GroEL chaperonin is another example of a large, machine-like oligomer (see Section 4.4).

Why do cells assemble such large oligomers from multiple subunits, rather than producing a large protein as a single, multidomain polypeptide chain? There are several reasons. For one, the folding of the many different domains in a single, very large polypeptide chain may be problematic. In addition, if one domain did not fold properly, the entire protein would lack function, so the investment of energy in making it would be wasted. A multisubunit composition avoids these problems. If a protein subunit misfolds, it will not be included in the oligomer, but at least only the investment of cellular resources to make one domain will be wasted. In fact, a similar argument can be made even if the entire machinery is folded correctly. If one subunit subsequently denatures (becomes unfolded) or becomes inactive for any reason, it can be replaced by another subunit. The accuracy of translation is also an issue for very large proteins. During protein synthesis, approximately one mistake occurs in every 100,000 peptide-bond joining events. Therefore, a large protein of M_r exceeding 10^6 may accumulate mistakes, perhaps becoming inactive, whereas smaller, individual subunits that do not fold properly can simply be discarded. Finally, the architecture of many large oligomers includes multiple copies of some subunits. More DNA would be required to encode a single large protein than is needed to encode multiple copies of individual subunits.

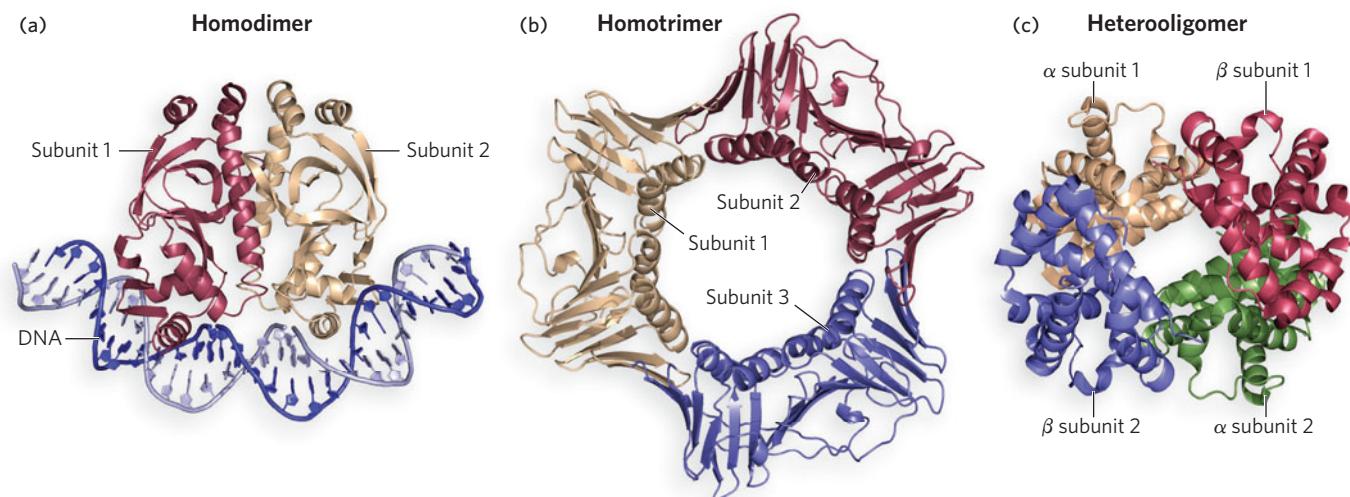


FIGURE 4-18 The quaternary structure of three proteins.

Protomers (subunits) are represented by different colors.
(a) Homodimeric *E. coli* CRP protein, as a complex with DNA.

(b) Human PCNA, a homotrimer. (c) Human hemoglobin, a heterooligomer of two α chains and two β chains. [Sources: (a) PDB ID 1CGP. (b) PDB ID 1AXC. (c) PDB ID 1GZX.]

HIGHLIGHT 4-2 A CLOSER LOOK

Protein Structure Databases

Each protein structure contains thousands of atoms arranged in three-dimensional space. When a scientist determines a protein structure, the x , y , and z coordinates for each atom are stored in a database called the **Protein Data Bank (PDB)** as a PDB file. To view the atomic structure of a protein on a computer screen, the PDB file is imported into a computer program that displays the atomic model. There are many types of computer programs written just for this purpose, including PyMol, which is easily used on desktop and laptop computers.

Proteins can be classified according to their secondary structural elements, supersecondary structural motifs, and sequence homologies, and there are several databases that classify protein structures. One of the oldest is the **Structural Classification of Proteins (SCOP) database**. The highest level of classification in the SCOP database divides proteins into one of four classes: all α helix, all β sheet, α/β (where α and β segments are interspersed), and $\alpha + \beta$ (where α and β segments are somewhat segregated). For instance, myoglobin has α helices but no β sheet and therefore is placed in the α class. Likewise, proteins consisting only of a

β barrel are in the β class. DnaA and Pol I have α helices mixed in with β sheets and thus are in the α/β class. In the circular β and PCNA clamps, the β sheet and α helices are somewhat segregated, placing these proteins in the $\alpha + \beta$ class.

Each class contains tens to hundreds of distinct substructure folding arrangements. Some substructures are very common, and others are found in only one protein. The number of unique folding motifs in proteins is far lower than the number of proteins—perhaps fewer than 1,000 different folds. As new protein structures are elucidated, the proportion of those containing a new motif has been declining. Below the levels of class and fold, which are purely structural, categorization in the SCOP database is based on evolutionary relationships.

The SCOP database is curated manually, with the objective of placing proteins in the correct evolutionary framework on the basis of conserved structural features. Two similar enterprises, the CATH (Class, Architecture, Topology, and Homologous superfamily) and FSSP (Fold Classification Based on Structure-Structure alignment of Proteins) databases, make use of more automated methods and can provide additional information.

Protein Structures Help Explain Protein Evolution

Proteins with a similar primary sequence and function usually share a common evolutionary heritage and are said to be in the same **protein family**. For example, the globin family consists of many different proteins with structural and sequence similarity to myoglobin (e.g., myoglobin and the α and β subunits of hemoglobin have the same folding pattern). However, in many cases, although the primary sequence does not show an evolutionary relationship, the protein structures are similar, indicating that they are related through a common ancestor. This is because the three-dimensional structure of a protein is more highly conserved than the primary sequence.

Protein families with little sequence similarity but with the same supersecondary structural motif and functional similarities are referred to as **superfamilies**. An evolutionary relationship among the families of a superfamily is considered probable, even though time and functional distinctions—resulting from different adaptive pressures—may have erased many telltale sequence relationships. A protein family may be widespread in all three domains of life—the Bacteria, Archaea, and Eukarya—

suggesting a very ancient origin. Other families may be present in only a small group of organisms, indicating that the protein structure arose more recently. Tracing the natural history of structural motifs, using systematic structural classification databases, provides a powerful complement to sequence analysis (Highlight 4-2).

SECTION 4.3 SUMMARY

- The tertiary structure of a protein is its three-dimensional structure, consisting of all of its secondary structural elements and loops.
- A protein domain is an independent folding unit within a protein and typically consists of up to 150 residues.
- Supersecondary structural elements, also called motifs, are arrangements of multiple secondary structural elements commonly found in proteins. Supersecondary structures are the building blocks associated with particular functions.
- The quaternary structure of a protein includes all the connections between two or more polypeptides and can range from a simple homodimer to a large, multiprotein assembly such as a ribosome.

- The three-dimensional structure of proteins that evolved from a common ancestor is more conserved than their primary sequence. Therefore, protein structures are very useful in determining evolutionary heritage.

4.4. Protein Folding

The structures of several hundred proteins have been solved, and in all cases the polypeptide backbone folds to adopt a particular conformation, a process known as **protein folding**. Decades have passed since the classic studies of Christian Anfinsen showing that the amino acid sequence determines the folding pattern of a protein. He showed that ribonuclease that had been completely unfolded in a denaturing solution could rapidly fold into a biologically active protein after removal of the denaturant. Ribonuclease contains eight Cys residues that form four intrachain disulfide bonds. Amazingly, all four disulfide links formed in the correct places, even though there are 105 possible combinations. The results suggested that other, weak interactions direct protein folding and precisely position the Cys residues prior to disulfide bond formation.

Not all proteins renature as easily as ribonuclease, and the exact process by which most proteins fold is still unknown. Furthermore, some proteins require the assistance of other proteins for proper folding. Even when correctly folded, a protein structure is constantly in flux, and all of its atoms and structural elements vibrate rapidly. This inherent flexibility is essential to protein function—to achieve, for instance, a different thermodynamic state when substrate is converted to product in an enzyme reaction. As a further example, some proteins are triggered by the binding of a substance known as an allosteric effector to adopt another conformational state that has substantially different activity (see Chapter 5).

Predicting Protein Folding Is a Goal of Computational Biology

The primary sequence holds the instructions for protein folding, so theoretically, researchers should be able to predict a protein's tertiary structure from its sequence alone. However, this is not yet possible—but not for want of trying. The protein folding “code” is complex, and the problem is complicated by the small difference in free energy between the folded and unfolded states of a protein. Although hydrogen bonds are strong, hydrogen bonding is not very different in the folded and unfolded states, because water is plentiful in the cell and can also hydrogen-bond with the protein. The main force driving

protein folding is therefore derived from van der Waals contacts and hydrophobic interactions, but the free-energy difference between folded and unfolded protein states is still only about 5 to 15 kcal/mol. This small energy difference makes it difficult to understand and quantify the protein folding “code” that determines three-dimensional structure from primary sequence.

Inspection of many protein structures has provided some guidelines by which proteins fold (see Section 4.5). Hydrophobic residues avoid water by becoming buried, and this drives the polar peptide backbone into the interior, where it must form secondary structures with internal hydrogen bonds. The interior of a protein is amazingly well packed: about 75% of atoms in the interior of a protein are packed together as close as their van der Waals radii, the theoretical limit. Not even crystals of free amino acids are more closely packed. Polar residues of a protein are usually at the surface, where the side chains can interact with water. These guidelines are used when predicting the extent to which different residues are buried in protein structures (Figure 4-19).

The exceptions to this general rule are easily explained. The hydrophobic side chain of proline is sometimes found on the surface, but Pro residues are useful for making sharp turns, which are usually on the surface. Some Cys residues form disulfide bridges that are hydrophobic and therefore easily buried. Also, some residues that do not at first seem hydrophobic contribute greatly to hydrophobic forces. For example, lysine, a polar, charged amino acid, has a long hydrocarbon side chain that is often buried and contributes to hydrophobic interactions, while only its charged amino group protrudes to the surface.

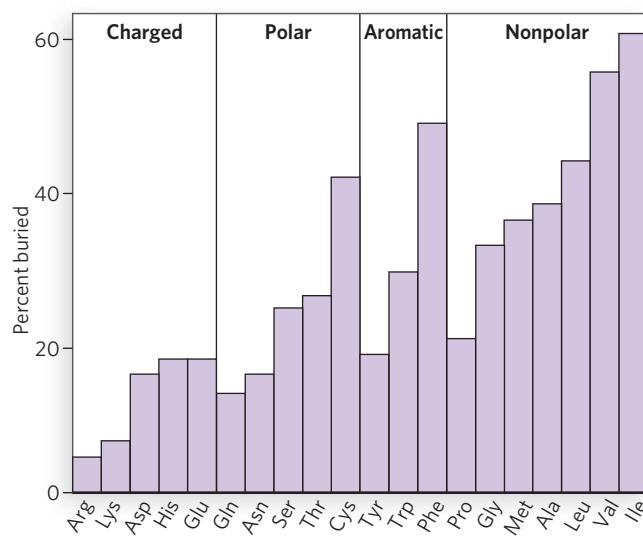


FIGURE 4-19 Buried residues. This plot shows the relative extent to which the 20 different amino acid residues are found in buried positions in globular proteins. [Source: Adapted from P. Y. Chou and G. D. Fasman, *Biochemistry* 13:222–245, 1974.]

Computational biologists have developed powerful algorithms to predict protein structure from amino acid sequences. The computations make use of the allowed values of torsion angles for different amino acids, their polar, ionic, and hydrophobic features, and the large body of structural information contained in the protein structure database. In fact, there is an international competition each year, called CASP (Critical Assessment of Techniques for Structure Prediction), that evaluates different protein structure prediction algorithms. To make this competition possible, research groups that have determined a protein structure but have not yet published it provide only the sequence of the protein to the CASP competition. Then computational biologists apply their algorithms to these sequences and arrive at structural models. When the experimentally determined structures are published, the degree to which the computed models agree with the experimentally determined structures is evaluated, and the scientist with the most accurate algorithm wins. The CASP competition offers awards in several areas, including secondary, tertiary, and de novo structure prediction.

Structure-based protein design is another goal of computational biology. The scientist starts with a “target” structure and then computationally designs a sequence that adopts the target folding pattern. Most of the algorithms developed for protein design focus on the redesign of a preexisting protein core. The complete de novo design of a protein is a difficult computational problem, given the vast number of possible conformations each amino acid residue can adopt.

A striking advance in de novo protein design was made by Steve Mayo's group. An algorithm they had developed was applied to compute a sequence that folds into the structure of a zinc finger domain (Figure 4-20a). As we noted earlier in the chapter, zinc finger domains consist of a β - α - β motif, composed of about 30 amino acids. Despite the small size, the zinc finger domain contains sheet, helix, and turn structures, and the zinc atom is not required for proper folding. For a target structure, Mayo and colleagues used a 28-residue zinc finger domain in a protein called Zif268. There are 1.9×10^{27} possible combinations of 28-residue proteins, given the allowed torsion angles of the 20 common amino acids. The enormous size of the computational space can be appreciated by the fact that just one molecule of each of these peptides, lumped together, would amount to 11.6 metric tons! The sequence to emerge from 90 hours of computing time gave a novel protein sequence that was unrelated to any known sequence in the database, yet was predicted to have a folding pattern similar to that of the target protein (compare the two in Figure 4-20a,b). The structure of the protein having the computed sequence, FSD-1 (full sequence design-1), was then experimentally determined by nuclear magnetic resonance

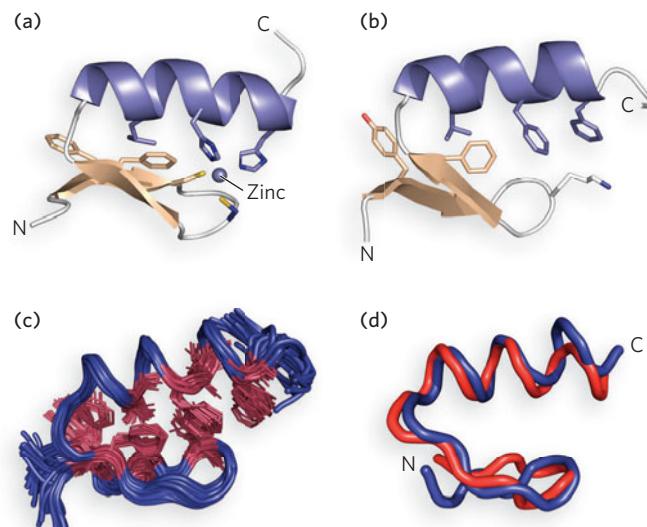


FIGURE 4-20 A protein structure designed by computation. (a) The structure of a zinc-binding domain from a DNA-binding protein (Zif268). (b) The predicted structure of the FSD-1 sequence, derived entirely by computation. (c) The NMR structure of FSD-1. (d) Superimposition of the Zif268 zinc-binding domain (red) and FSD-1 (blue). Only residues 3 through 26 are shown. [Source: B. I. Dahiyat and S. L. Mayo, *Science* 278:82–87, 1997.]

(NMR) (Figure 4-20c); the actual FSD-1 structure corresponded amazingly well with the target protein (Figure 4-20d). The overall deviation in backbone atoms of the target and FSD-1 structures was only 0.98 Å over residues 8 through 26. The fact that a completely different sequence can serve the same function supports the idea that once evolution arrives at a particular solution, it often uses this solution repeatedly in other proteins, rather than coming up with an entirely new one.

Polypeptides Fold through a Molten Globule Intermediate

In 1968, Cyrus Levinthal reasoned that a protein should not be capable of randomly folding into a unique conformation in our lifetime. For example, starting from a 100-residue polypeptide in a random conformation, and assuming that each amino acid residue can have 10 different conformations, 10^{100} different conformations for the polypeptide are possible. If the protein folds randomly, by trying out all possible backbone conformations, and if each conformation is sampled in the shortest time possible ($\sim 10^{-13}$ seconds, the time scale of a single molecular vibration), it would still take about 10^{77} years to sample all possible conformations! Yet, *E. coli* makes a biologically active protein of 100 amino acids in about 5 seconds at 37°C. This apparent contradiction is referred to as Levinthal's paradox, and the astronomical disparity between calculation and

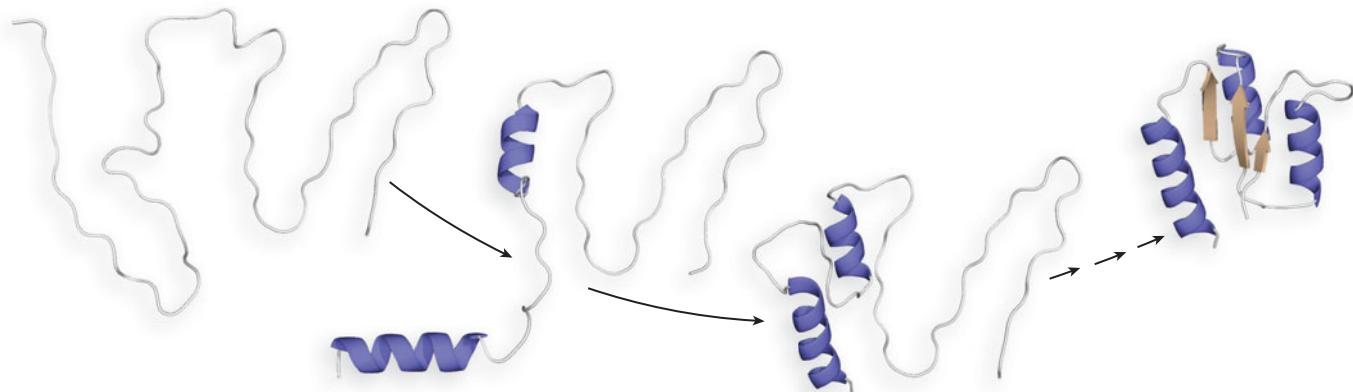


FIGURE 4-21 A hierarchical model of protein folding. In this model, which has been demonstrated for a few small proteins, some secondary structural elements form locally, and they nucleate the folding of the rest of the protein.

observation reveals that protein folding is far from random: it must follow an ordered path that side-steps most of the possible intermediate conformations.

Intensive studies indicate the different ways that intermediate conformations could be side-stepped. We discuss here two pathways that are supported by experimental data. A **hierarchical model** of folding proposes that local regions of secondary structure form first, followed by longer-range interactions (e.g., between two α helices), and that this process continues until complete domains form and the entire polypeptide is folded (Figure 4-21).

In the **molten globule model**, the hydrophobic residues of a polypeptide chain rapidly group together and collapse the chain into a condensed, partially ordered, “molten globule state.” With the protein condensed into a ball, the number of possible conformations is drastically limited to those that can occur within the confines of the molten globule. The molten globule state is not as compact as the final state; only about half of the hydrophobic residues are buried, and those that remain on the surface must find their way to the core. As hydrophobic residues are buried, the polar backbone atoms must form hydrogen-bonded secondary structures. The molten globule is therefore an ensemble of different partially ordered segments that shift and churn through multiple conformations, searching for the compact state of the native—completely folded—structure. As subdomains with tertiary structure begin to develop, the variety of different conformations decreases until most members of the population finally attain the native structure.

Most proteins probably fold by a process that incorporates features of both models. Instead of following a single pathway, a population of identical polypeptides may take several routes to the same end point. Thermodynamically, the folding process can be viewed as a kind of free-energy funnel (Figure 4-22). Unfolded forms

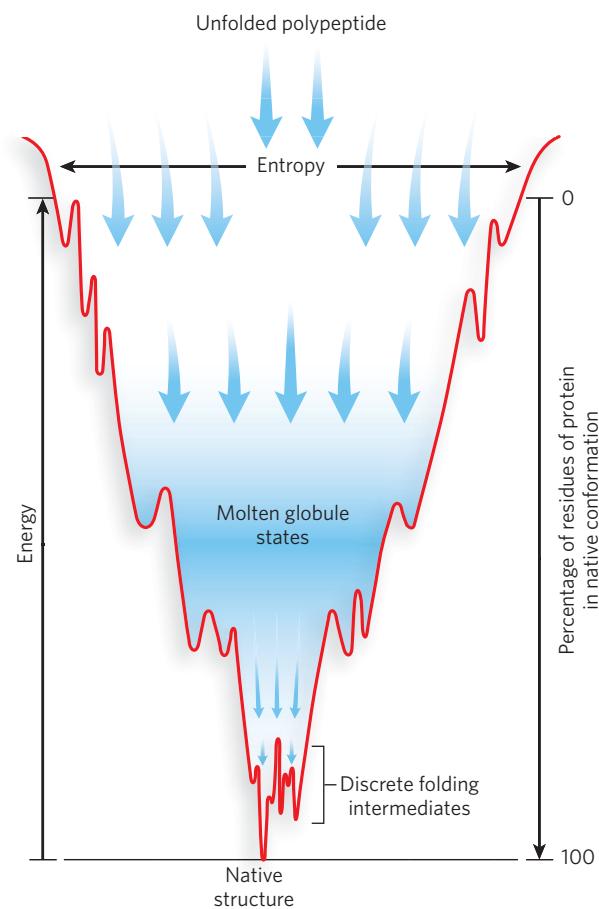
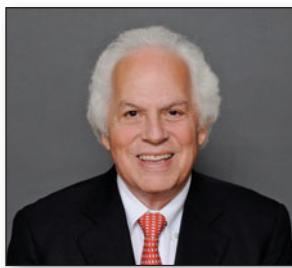


FIGURE 4-22 Protein folding as a free-energy funnel.

The top of the funnel represents the unfolded state of a protein, the highest entropy state. As it folds, the protein progresses to lower free energy. The more compact the protein gets, the more rapidly it arrives at the properly folded and lowest free-energy state, at the bottom of the funnel. [Source: Adapted from P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, *Science* 267:1619, 1995.]

HIGHLIGHT 4-3 MEDICINE

Prion-Based Misfolding Diseases



Stanley Prusiner [Source:
Courtesy of Stanley Prusiner.]

A misfolded protein seems to be the cause of several rare, degenerative brain diseases in mammals. Perhaps the best known is bovine spongiform encephalopathy (BSE), also known as mad cow disease. An outbreak of BSE made international headlines in the spring of 1996.

Related diseases are kuru and Creutzfeldt-Jakob disease in humans, scrapie in sheep, and chronic wasting disease in deer and elk. These diseases are referred to as **spongiform encephalopathies** because the diseased brain frequently becomes riddled with holes (Figure 1). Symptoms include dementia and loss of coordination, and the diseases are usually fatal.

In the 1960s, investigators found that preparations of the disease-causing agents seemed to lack nucleic acids, suggesting that the agent was a protein. Initially, the idea seemed heretical. All disease-causing agents known up to that time—viruses, bacteria, fungi, and so on—contained nucleic acids, and their virulence was related to genetic reproduction and propagation. However, four decades of investigation, pursued most notably by Stanley Prusiner, provided evidence that spongiform encephalopathies are different.

The infectious agent has been traced to a single protein (M_r 28,000), dubbed **prion** (from proteinaceous infectious *only*; analogous to “virion”).

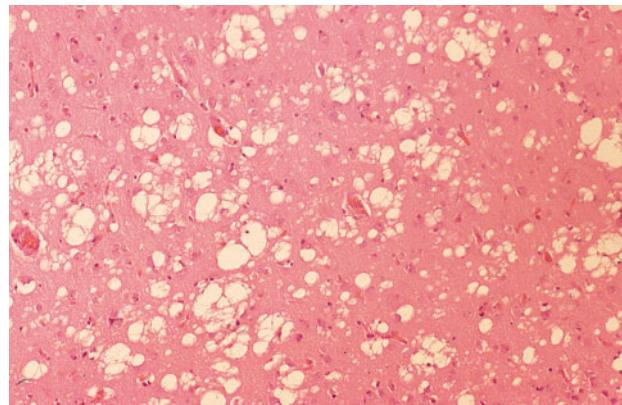


FIGURE 1 A stained section of cerebral cortex from a patient with Creutzfeldt-Jakob disease shows spongiform degeneration (holes in the tissue), the most characteristic neurohistological feature. [Source: Stephen J. DeArmond.]

The infected tissue contains abnormal spots with densely packed protein fibers called plaques. Fibers in these plaques are resistant to protease, and thus the protein is referred to as PrP (protease-resistant protein). The normal prion protein, called PrP^C (C for cellular), is found throughout the body in healthy animals, but it does not form plaques. Its role in the mammalian brain (or any tissue) is not known, but it may have a molecular signaling function. Strains of mice lacking the *Prnp* gene that encodes PrP (and thus lacking the protein itself) suffer no obvious ill effects. Illness occurs only when the PrP protein produced misfolds, or converts, to an altered conformation, called PrP^{Sc} (Sc for scrapie). The misfolded PrP^{Sc} aggregates to

rapidly collapse to more compact forms, after which the number of conformational possibilities decreases. These rapid first stages quickly narrow the funnel. Small depressions along the sides of the free-energy funnel represent semistable intermediates that briefly slow the folding process. At the bottom of the funnel, the ensemble of folding intermediates has been reduced to the single conformation of the final, native state of the protein.

Defects in protein folding may be the molecular basis for a wide range of human genetic disorders. For example, cystic fibrosis is caused by the misfolding

of a chloride channel protein called cystic fibrosis transmembrane conductance regulator (CFTR). Many disease-related mutations in collagen are also caused by defective folding. In addition, defective protein folding causes the prion-related diseases of the brain (Highlight 4-3).

Thermodynamic stability is not evenly distributed over a protein. For example, a protein may have two stable domains joined by a segment with lower structural stability. The regions of low stability may allow the protein to alter its conformation between two (or

form fibers that presumably lead to the plaques associated with spongiform encephalopathy. Although researchers do not completely understand how the misfolded state of PrP is propagated during infection, it is commonly thought that the interaction of PrP^{Sc} with PrP^{C} converts the latter to PrP^{Sc} , initiating a domino effect in which more and more of the brain protein converts to the disease-causing form.

Spongiform encephalopathy can be inherited, can occur spontaneously, or can be transmitted through infection. Most cases are spontaneous. Infectious transmission accounts for fewer than 1% of cases and occurs through intimate contact with diseased tissue. Inherited forms of prion diseases, which account for 10% to 15% of cases, are due to a variety of point mutations in the *Prnp* gene, each of which is believed to make the spontaneous conversion of PrP^{C} to PrP^{Sc} more likely. A detailed understanding of prion diseases awaits new information on how prions affect brain function.

The structure of the C-terminal region of normal PrP^{C} is known (Figure 2); it contains three α helices and two β strands. In contrast, circular dichroism indicates that fibers of PrP^{Sc} contain a core of tightly packed β sheets. The exact structure of PrP^{Sc} is unknown, because it is an aggregate and difficult to work with, but it is thought to contain several β sheets—accounting for the prevalence of β sheet structure in the PrP^{Sc} aggregate that is inferred from CD measurements. Future insights about the structure and function of PrP^{Sc} will provide treatment strategies for this disease.

more) states. As we shall see in Chapter 5, variations in the stability of regions within a given protein are often essential to protein function.

Chaperones and Chaperonins Can Facilitate Protein Folding

Not all proteins fold spontaneously. The folding of many proteins is facilitated by **chaperones**, specialized proteins that bind improperly folded polypeptides and facilitate correct folding pathways or provide microenvironments

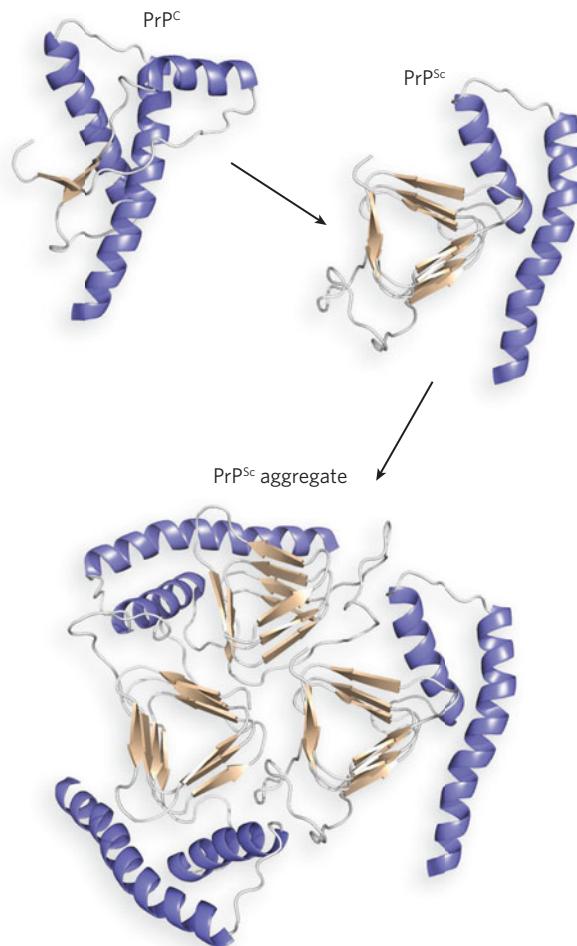
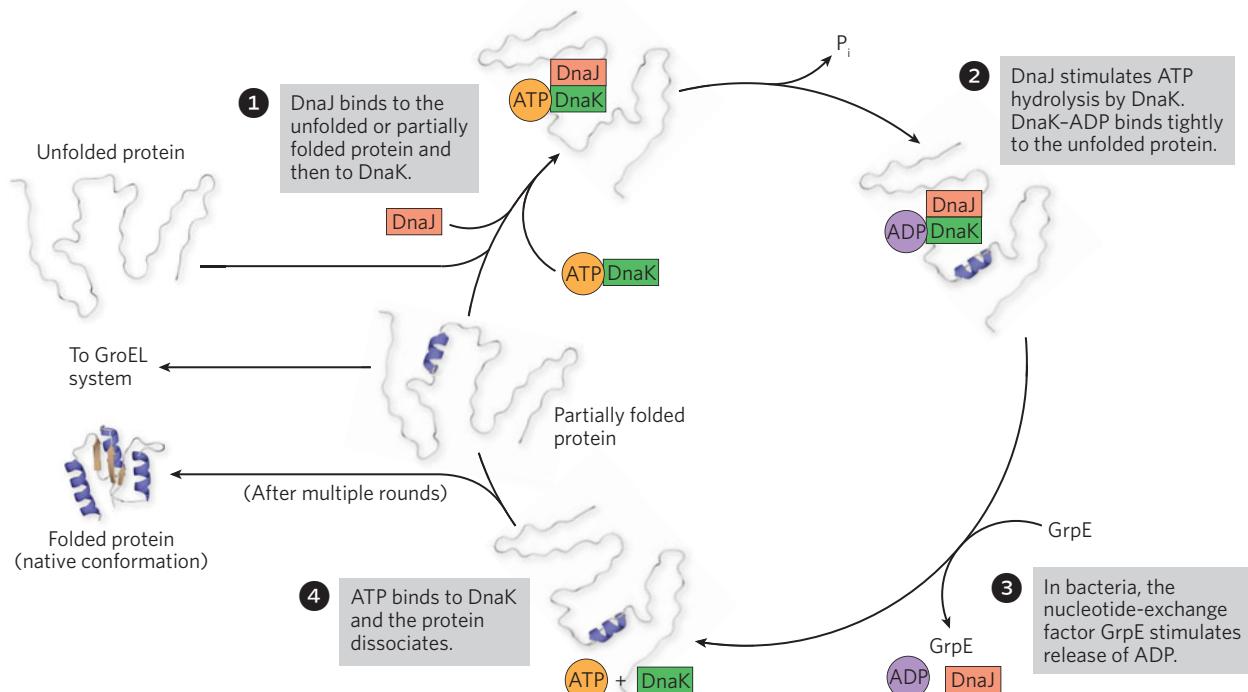


FIGURE 2 The globular domain of human PrP^{C} monomer (top, left) contains three α helices and two short β strands. To its right is a proposed structure of the corresponding region of PrP^{Sc} , containing many more β strands that may promote aggregation (bottom). [Sources: (top left) PDB 1QLX. (top right and bottom) Adapted from S. B. Prusiner, *Sci. Am.* 291(1):86–93, 2004.]

in which folding can occur. Two classes of molecular chaperones have been well studied, and both are conserved across species, from bacteria to humans. The first class, a family of proteins called **Hsp70** (heat-shock proteins of M_r 70,000), become more abundant in cells stressed by elevated temperature. Hsp70 proteins bind to unfolded regions that are rich in hydrophobic residues, preventing aggregation, and thus protecting denatured proteins and newly forming proteins that are not yet folded. Some chaperones also facilitate the assembly of subunits in oligomeric proteins.

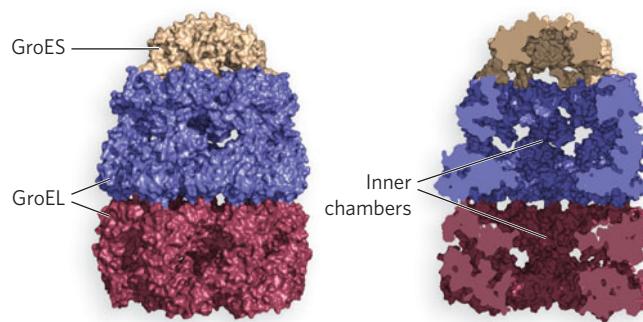
**FIGURE 4-23 Chaperone-assisted protein folding.**

Chaperones bind and release polypeptides by a cyclic pathway, shown here for *E. coli* chaperone proteins DnaK and DnaJ, homologs of the eukaryotic chaperones Hsp70 and Hsp40. The chaperones mainly prevent the aggregation of unfolded polypeptides. Some polypeptides released at the

end of a cycle are in a native conformation; the rest are rebound by DnaK or are directed to the chaperonin system (GroEL/GroES; see Figure 4-24). In bacteria, GrpE interacts transiently with DnaK late in the cycle (step 3), promoting dissociation of ADP and possibly DnaJ. No eukaryotic analog of GrpE is known.

Hsp70 proteins bind and release polypeptides in a cycle that involves several other proteins (including the class Hsp40) and ATP hydrolysis. Figure 4-23 diagrams chaperone-assisted folding for the DnaK and DnaJ chaperones of *E. coli*, functional equivalents (homologs) of the eukaryotic Hsp70 and Hsp40 proteins.

Chaperonins, the second class of chaperones, are elaborate protein complexes required for the folding of some cellular proteins. In *E. coli*, an estimated 10% to 15% of cellular proteins require the resident chaperonin system—GroEL/GroES—for folding under normal conditions. Up to 30% of proteins require assistance when the cell is heat-stressed. Unfolded proteins are bound in pockets in the GroEL complex, and the pockets are capped transiently by GroES (Figure 4-24). GroEL undergoes substantial conformational changes, coupled with ATP hydrolysis and the binding and release of GroES, which together promote folding of the bound polypeptide. Although the structure of the GroEL/GroES chaperonin is known, many details of its mechanism of action remain unresolved.

**FIGURE 4-24 Chaperonin-assisted protein folding.** The *E. coli* chaperonins GroEL and GroES. Each GroEL complex consists of two chambers formed by two heptameric rings (each subunit M_r , 57,000). GroES, also a heptamer (each subunit M_r , 10,000), blocks one of the GroEL pockets. Surface (left) and cut-away (right) images of the GroEL/GroES complex are shown. [Source: PDB ID 1AON.]

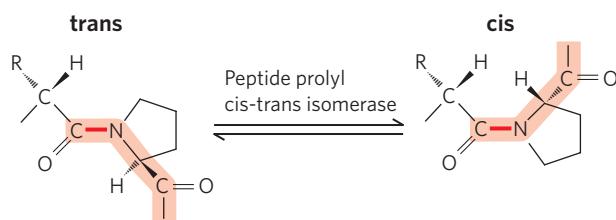


FIGURE 4-25 The cis and trans isomers of proline in a peptide bond. Most peptide bonds (>99.95%) are in the trans configuration, but about 6% of bonds involving Pro residues are in the cis configuration; many of these occur in β turns.

Protein Isomerases Assist in the Folding of Some Proteins

The folding of certain proteins requires enzymes that catalyze isomerization reactions. **Protein disulfide isomerase** is a widely distributed enzyme that catalyzes the interchange, or shuffling, of disulfide bonds until the bonds of a protein's final conformation have formed. Among its functions, protein disulfide isomerase catalyzes the elimination of folding intermediates with inappropriate disulfide cross-links. The cis and trans isomers of peptide bonds are rapidly interchangeable (see Section 4.1). However, the cyclic structure of proline makes interconversion between cis and trans isomers a slow process. **Peptide prolyl cis-trans isomerase** catalyzes the interconversion of the cis and trans isomers of proline peptide bonds (Figure 4-25). Most peptide bonds in proteins are in the trans isomeric form, but the cis isomer of proline is often found in tight turns between secondary structural elements.

SECTION 4.4 SUMMARY

- The polypeptide chain of a protein folds into a unique conformation, and the instructions for this folding are inherent in its primary sequence.
- The folding pattern of a protein is hard to predict from the primary sequence, because the forces that stabilize the folded state are weak and cannot be recognized from the amino acid sequence.
- In a protein's folded state, hydrophobic residues are usually found in the interior of the protein, and polar residues often localize to the surface.
- The protein-folding pathway is not random. The folding of some proteins is thought to proceed through a condensed molten globule state that limits folding to conformations that are compatible with a compact volume.

- Protein folding is sometimes assisted by chaperones and chaperonins. Chaperones are proteins of the Hsp70 class that bind unfolded proteins and use cycles of ATP-binding and hydrolysis to help the proteins refold. Chaperonins are complex multisubunit structures that engulf the protein in a chamber during the refolding process.

- Protein folding is also assisted by isomerases. Protein disulfide isomerase catalyzes the breakage and re-formation of disulfide bonds, and peptide prolyl cis-trans isomerase facilitates interchange between the cis and trans isomers of Pro residues.

4.5 Determining the Atomic Structure of Proteins

There are very few methods to deduce a protein's tertiary structure. Proteins are too small to allow resolution of structural details with visible light. The lower limit of visible light has a wavelength of about 400 nm (400×10^{-9} m) and therefore cannot resolve objects of a size less than about half this wavelength (200 nm, or 2,000 Å). Even huge ribosomes, with a radius of 18 nm, are not visible in a light microscope. The electron microscope has high resolving power, but at the high-energy wavelengths needed for atomic resolution, the electron beam rapidly destroys the sample. True atomic resolution requires a wavelength of ~1.5 Å, about the length of an atomic bond. X rays fall within this range, and they provide atomic resolution. Nuclear magnetic resonance (NMR) operates in an entirely different way and is the only other method that can reveal protein structures at the atomic level.

Most Protein Structures Are Solved by X-Ray Crystallography

More than 90% of the protein structures in the Protein Data Bank, a repository of information on protein structures, have been determined by **x-ray crystallography** (see Highlight 4-2). There is no theoretical limit to the size of protein that can be analyzed by this method. Although the process requires specialized equipment and sophisticated computer programs, the basic principles are not difficult.

Amplifying Diffracted X Rays X-ray crystallography illuminates a protein crystal with an x-ray beam, and the diffracted x rays are collected for analysis (Figure 4-26a). Diffracted x rays travel in every direction

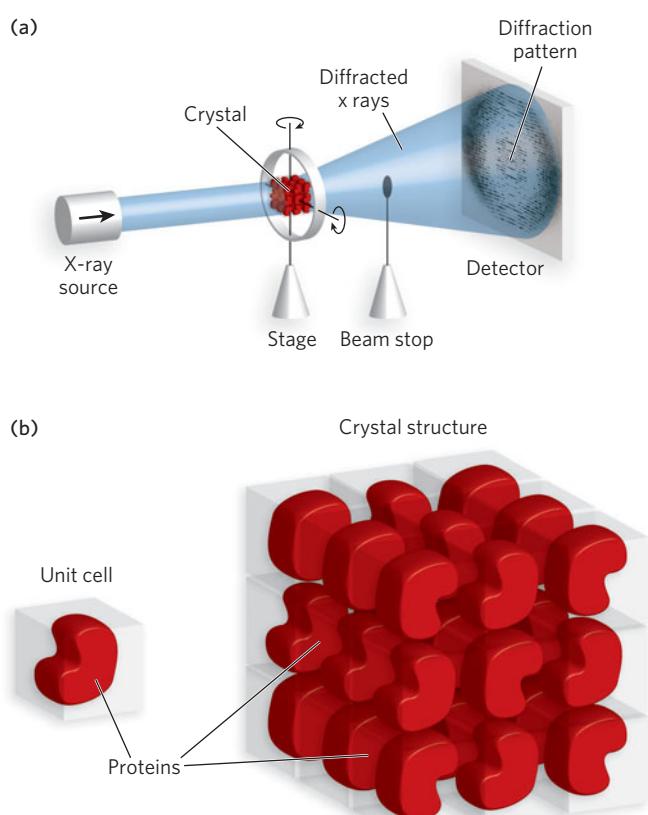


FIGURE 4-26 Protein crystals and diffraction patterns. (a) Protein crystals produce a diffraction pattern when x rays are passed through them. A crystal is rotated in all directions, and diffracted x rays are collected by a detector. X rays that pass through the crystal without diffraction are blocked from hitting the detector by a beam stop. (b) The unit cell, or repeating unit in the lattice of a protein crystal, may contain one or more protein molecules.

and thus normally create a blur on a detector. But in a crystal, trillions of protein molecules are aligned in a regular lattice, and therefore some of the diffracted x rays combine and add up in a process called **constructive interference**, forming a **reflection spot** on a film or detector. Each reflection spot in a **diffraction pattern** is produced by the summation of diffracted x rays from every atom in the **unit cell**, the smallest regularly repeating unit in the crystal (Figure 4-26b). The unit cell can be as small as a single protein molecule, but often consists of two or more identical protein molecules.

The physical basis of a crystal diffraction pattern was determined in 1913 by William Henry Bragg and his son William Lawrence Bragg. They made an analogy between diffracted x rays and light rays reflected from a mirror or grating. In x-ray diffraction, the layers of the grating are created by the different atoms in each unit cell. The only emitted x rays that constructively inter-

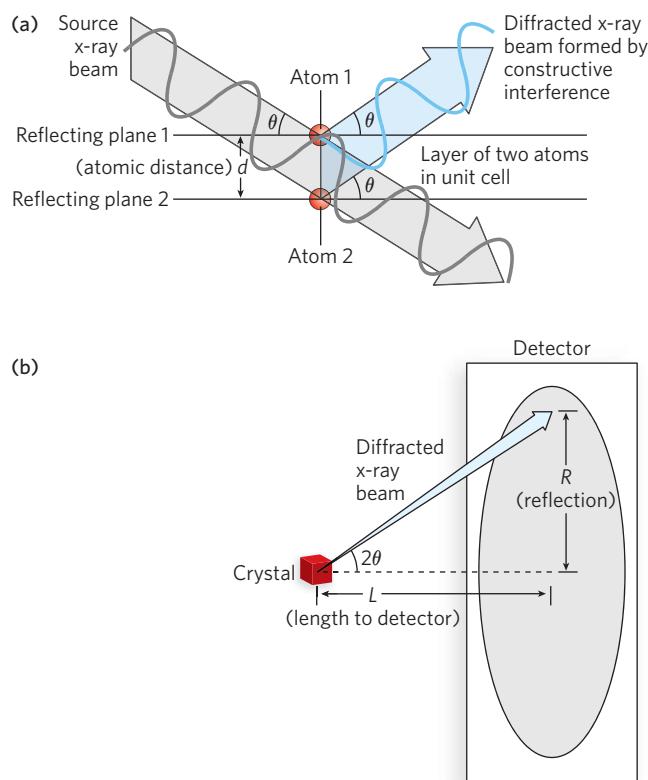


FIGURE 4-27 Determining the distance between two atoms in a crystal. (a) Diffracted x rays that constructively interfere produce reflection spots at the detector. In a unit cell, each reflection plane is a different atom. Just two reflection planes are shown, for simplicity. (b) One diffracted x-ray beam from a crystal. Reflections are related to reflecting planes (atoms) in the crystal by Bragg's law ($\lambda = 2d \sin \theta$). R is the distance of a reflection from the center of the detector, and L is the distance from the crystal to the detector. See text for details.

fere, or add up, are those that are reflected at the same angle as the x-ray beam (Figure 4-27a). The spacing of reflections is related to atomic distance in the unit cell, by a formula referred to as **Bragg's law**: $\lambda = 2d \sin \theta$. The distance (d) between reflection (lattice) planes—that is, between atoms—in the unit cell depends on the wavelength (λ) of the x-ray beam and the angle (θ) at which the beam strikes a reflecting plane in the crystal (Figure 4-27b). The distance from the crystal to the film (L) and the distance from the center of the film to the diffraction point (R) yield the angle (θ) of the x-ray beam to a reflecting plane in the crystal. The angle can be substituted into the Bragg's law formula to solve for d , the distance between lattice planes, or atoms. To obtain a sufficient number of reflections for determining the protein structure, the crystal is rotated during x-ray irradiation. Tens of thousands of reflections are usually collected to solve one protein structure.

Reconstructing the Protein Image An object illuminated in a light microscope also produces a diffraction pattern, but it is not visible, because the diffracted light is recombined into an image with a converging lens. Electron microscopy works in a similar fashion, using magnets to refocus the diffracted electrons into an image. However, no lens can recombine diffracted x rays. Instead, the diffraction pattern is recombined into an image by a mathematical converging series, called a Fourier series, that acts like a converging lens. X rays are photons and therefore behave as sine waves, each of which has an amplitude, a wavelength, and a phase. These parameters are required for the Fourier series. The wavelength (λ) is the same as that of the x-ray beam used to illuminate the crystal, and the amplitude (A) is calculated from the spot intensity ($I = A^2$). The problem lies with the phase. When a diffracted x-ray wave hits the detector, the wave collapses and the phase is lost. However, there are several methods for determining the phase of each reflection (discussed below).

The reconstructed image is displayed on a molecular graphics console as a volume encased in a meshwork referred to as an **electron density map** (Figure 4-28). The higher the resolution, the greater the detail that the electron density map will contain. Paradoxically, the reflections in the diffraction pattern that carry the highest resolution are those that are farthest from the center. This is easily explained by Bragg's law,

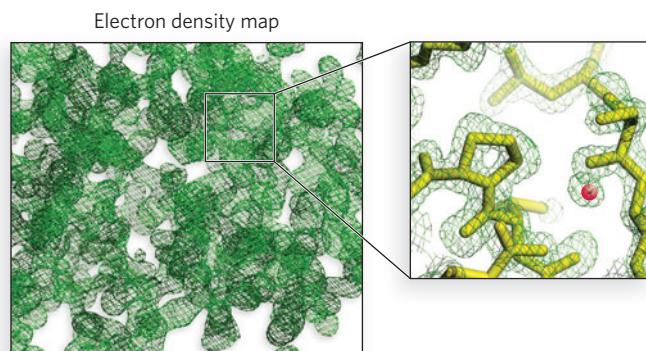


FIGURE 4-28 An electron density map. An electron density map (left) contains too much information to analyze when viewed all at once. The experimenter focuses instead on one small region at a time (right) and fits the polypeptide backbone into the density. Shown here are molecules that lie in the outlined portion of the electron density map. The small red sphere is an ordered water molecule. [Source: Based on images courtesy of Roxana Georgescu, laboratory of Mike O'Donnell, Rockefeller University.]

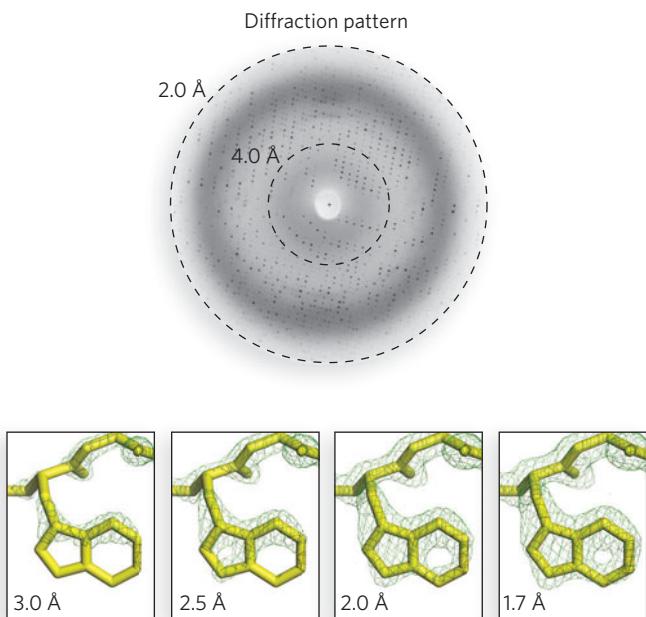


FIGURE 4-29 The relationship between resolution and diffraction pattern. The dotted circles in the diffraction pattern represent the locations of reflections responsible for two different resolutions of an electron density map. The maps below show four different resolutions. The area of electron density corresponds to Trp¹²³ of the *E. coli* DNA polymerase β subunit. Positions of the atoms of the Trp residue in the final model are shown in the four panels. [Source: Based on data and images courtesy of Roxana Georgescu, laboratory of Mike O'Donnell, Rockefeller University.]

$\lambda = 2d \sin \theta$: the larger the angle (θ), the smaller the distance (d) between the two points.

Examples of electron density maps of a tryptophan side chain in a protein, obtained by using reflections at increasing distances from the center of a diffraction pattern, are shown in Figure 4-29. Using reflections in the diffraction pattern that correspond to a resolution of 4 Å between atoms in an electron density map, the resolution is insufficient to trace the path of the peptide backbone or to place most of the side chains. Using reflections that correspond to a resolution of 3 Å, the peptide backbone is discernible as a continuous ribbon. Secondary structural elements are visible, and the general shape of side chains is often apparent, but there are usually some disordered regions in loops that cause breaks in the density and prevent a continuous chain trace. A range of 2.2 to 3.0 Å resolution is required to get the most complete information on a protein's structure.

The Initial Model The three-dimensional protein structure inferred from the electron density map is known as

the **initial model**. In the early stages of analysis, the initial model is hypothetical. To build the model, the known amino acid sequence of the protein must be fitted into the electron density mesh. Model building is aided by graphics on a computer screen, but it is mostly performed manually and requires the skill and patience of the experimenter. Because the peptide bond is planar, a peptide bond “ruler” helps identify the C_{α} atoms. To position the primary sequence in the electron density map, the experimenter looks for unusual arrangements of large, characteristic side chains. The remaining side chains are then filled in, and each is adjusted into the electron density. The resulting initial model is far from perfect, mainly owing to errors in determining the phases. These errors are minimized in the refinement process.

Refinement Improvements in the electron density map are generated by **refinement**, a process that increases the accuracy of the phases. The phases calculated during refinement eventually replace the less accurate phases determined initially. Refinement is an iterative process (Figure 4-30). It starts by taking the model and building a model crystal from it computationally (*in silico*). Then the Fourier series is used to compute a diffraction pattern for the model crystal, and the position and intensity of each calculated reflection are compared

with the observed diffraction pattern. The difference between the calculated and observed values yields a measurement of the error in the model, referred to as an **R factor** (R for residual error). At the first iteration, the R factor value is usually 0.4 to 0.5. Although refinement theoretically has no ending, in practice, structures are refined to an R factor value of 0.15 to 0.25.

The physical environment within a crystal is not identical to that in a solution or in a living cell. Therefore, the conformation of a protein in a crystal could, in principle, be affected by nonphysiological factors, such as incidental protein-protein contacts. However, when structures derived from crystal analysis are compared with structural information obtained by NMR (described on the next page), the crystal-derived structure almost always represents a functional conformation of the protein.

Confronting the Phase Problem The challenge of determining the phase of the reflections, referred to as the **phase problem**, is still a bottleneck in solving protein structure. The first method to solve the phase problem, known as isomorphous replacement, was developed by Max Perutz and John Kendrew to determine the structures of hemoglobin and myoglobin. Two other methods have since been developed, and their use is also widespread.

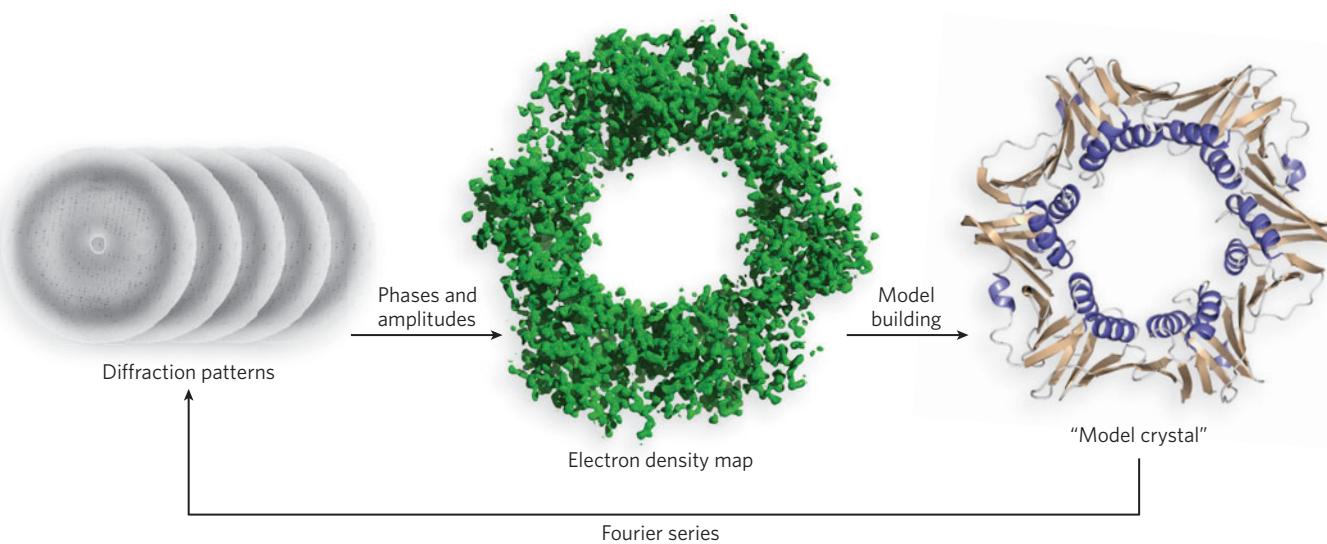


FIGURE 4-30 **Refinement.** During refinement in x-ray diffraction analysis, the initial model of the protein structure is used to calculate the theoretical diffraction pattern it would produce using a Fourier series. The phases are then adjusted to obtain a pattern close to the observed diffraction pattern. The adjusted phases generate a more detailed

electron density map, thus allowing more precise positioning of amino acid residues in the model. The process is repeated several times until the residual error (R factor) between observed and calculated diffraction patterns is reduced to an acceptable value. [Source: Based on images courtesy of Roxana Georgescu, laboratory of Mike O'Donnell, Rockefeller University.]



Max Perutz, 1914–2002 (left); John Kendrew, 1917–1997 (right) [Source: Corbis/Hulton Deutsch Collection.]

In **isomorphous replacement**, crystals are soaked with heavy metals, such as mercury, platinum, uranium, lead, or gold. The method requires the use of two heavy metal derivatives. Typically, only a few (one to three) heavy metal atoms bind specific sites in the protein without altering the structure (i.e., the structures with and without heavy atoms are isomorphous). These heavy atoms have dense electron clouds and interact strongly with x rays, thus altering the intensity of every reflection in the diffraction pattern. The difference in the reflection intensity of the crystal with and without heavy atoms mimics the diffraction pattern of a simple structure consisting only of the few heavy atoms. The coordinates for this simple structure can be solved, and the phases for the few heavy metals obtained. This information can be applied to determine the phases at each reflection of the unit cell.

Phases can also be determined by **multiwave-length anomalous dispersion (MAD)**, a method pioneered by Wayne Hendrickson at Columbia University. In MAD, only one heavy atom derivative is required. The most widespread use of this method employs selenium atoms to replace the sulfur in methionine. The selenomethionine protein is produced by growing cells in a medium containing selenomethionine as the only source of methionine. The beauty of this method is that it circumvents the arduous task of searching for heavy metal derivatives.

Molecular replacement is a third method of solving the phase problem. It starts with a protein of known structure as the initial model. The method works only if the initial model protein has a similar structure to the protein under study—for instance, two proteins with homologous (very similar) primary sequences, sharing a common ancestor in evolution and usually having similar three-dimensional structures. Continued iterative refinement improves the accuracy of the phases obtained from the initial model.

Smaller Protein Structures Can Be Determined by NMR

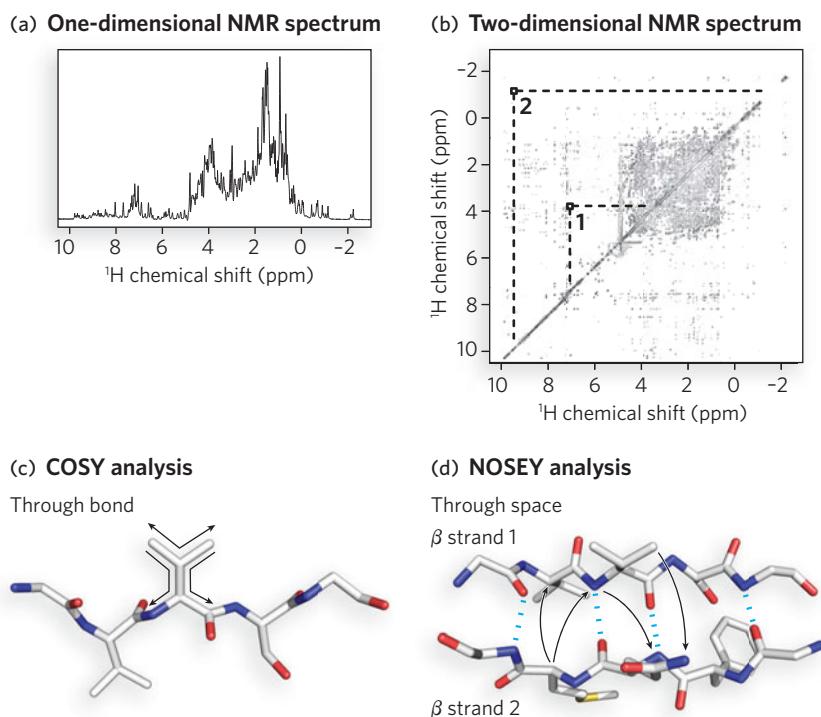
An important complementary method for determining protein structure is **nuclear magnetic resonance (NMR)**. NMR is performed in solution, which is an advantage over x-ray crystallography because protein

crystals can be difficult to obtain. However, only relatively small protein structures can be solved by NMR ($M_r < 25,000$).

Obtaining Primary Data The primary data from NMR are radio-frequency lightwave emissions from atomic nuclei. The technique involves placing the protein sample in a strong magnetic field, to align the spin of nuclei that contain a magnetic moment. Then the sample is pulsed with radio-frequency radiation to excite the nuclei. As the nuclei relax, they emit radiowaves. After many repetitions in rapid succession, the data are averaged. Repeated pulses of irradiation and collected emitted radiowaves are summed, thus increasing the signal-to-noise ratio to produce an NMR spectrum. The emissions are plotted as a spectrum of **chemical shifts**, expressed as parts per million (ppm). The chemical shift of a nucleus is sensitive to its environment and therefore carries environmental signatures that can be used to obtain structural information.

Only certain atoms, including ^1H , ^{13}C , ^{15}N , ^{19}F , and ^{31}P , possess the kind of nuclear spin that gives rise to an NMR signal. ^1H is particularly important in NMR experiments because of its high sensitivity and natural abundance. However, even a small protein has hundreds of ^1H atoms, typically resulting in a one-dimensional NMR spectrum too complex for analysis (Figure 4-31a). Structural analysis of proteins became possible with the advent of **two-dimensional NMR** techniques.

Many variations of two-dimensional NMR are performed by using different combinations of radio-frequency pulses and delays to separate the signals. In two-dimensional NMR, the data derived from the different pulses and delays are plotted along x and y axes, yielding a two-dimensional spectrum (Figure 4-31b). Instead of the plotting of peak height along the y axis, each spot carries a unique intensity that correlates with the peak height in the one-dimensional spectrum. The signals along the diagonal line through the two-dimensional spectrum are the same signals as in the one-dimensional spectrum, and the variation in intensity along the diagonal correlates with their peak heights. The signals that lie off the diagonal, called nonsequential signals, are derived by magnetization transfer between two protons that are close in space. In one type of two-dimensional NMR, called **correlation spectroscopy (COSY)**, the signals allow the identification of protons connected by covalent bonds (Figure 4-31c). In two-dimensional **nuclear Overhauser effect spectroscopy (NOESY)**, these nonsequential signals allow the measurement of distances through space between nearby atoms (Figure 4-31d).

**FIGURE 4-31** NMR spectra and protein-protein interactions.

(a) A one-dimensional NMR spectrum of a globin from a marine bloodworm. The spectrum represents the amount of chemical shift for each proton in a peptide segment. For a protein, the proton signals do not resolve in a one-dimensional spectrum, as indicated by the many overlapping peaks. (b) A two-dimensional NMR spectrum of the same globin molecule. The spots and their intensities that lie along the diagonal line are equivalent to the data contained in the peaks of the one-dimensional spectrum. The off-diagonal peaks (e.g., peaks 1 and 2) are nuclear Overhauser effect (NOE) signals generated

by close-range interactions of ^1H atoms that generate signals quite distant in the one-dimensional spectrum. (c) The two-dimensional COSY analysis identifies proton-proton signals through one or two covalent bonds (“through bond”) and thus is limited to individual amino acid units. (d) The NOESY analysis yields NOE signals resulting from proton-proton interactions occurring through empty space (“through space”) and thus identifies protons close in space but not necessarily close in the primary sequence. [Source: (c), (d) Based on images courtesy of Roxana Georgescu, laboratory of Mike O’Donnell, Rockefeller University.]

The COSY Spectrum The COSY spectrum identifies amino acid groups. Signals that lie off the diagonal in the COSY spectrum are called **through-bond correlation signals**, meaning that the signal is due to the interaction of two protons that are covalently connected through one or two intermediary atoms. Because the carbonyl group of the peptide bond lacks a proton, COSY signals are limited to proton couplings within single amino acid residues (see Figure 4-31c). The distinct coupling patterns between protons within amino acid residues act as fingerprints that identify individual residues in the COSY spectrum.

The NOESY Spectrum The NOESY spectrum identifies atoms that are not connected but are close in space. Signals that lie off the diagonal in the NOESY spectrum are NOE (nuclear Overhauser effect)

signals caused by magnetization transfer over distances of up to 5 Å through empty space. These **through-space NOE signals** can be used to identify adjacent amino acid residues, because the N–H proton of one residue is less than 3 Å from the proton on either the α -amino nitrogen, the C_{α} , or the first carbon on the R group of an adjacent residue. NOE signals also arise from residues that are far apart in the primary structure but close together in the tertiary structure (see Figure 4-31d).

Resolution in a ^1H NMR spectrum limits structure determination to small proteins of $M_r < 10,000$. For larger protein structures, either ^{13}C or ^{15}N spectra are required. Recombinant DNA technology can be used to prepare proteins that contain one of these rare isotopes. When either ^{13}C or ^{15}N is used, the analysis is referred to as three-dimensional NMR. When

both are used, it is called four-dimensional NMR. The NMR signals produced by these atoms, and the coupling with ^1H signals resulting from these substitutions, help in the assignment of individual ^1H signals and may be used to solve the structure of larger proteins, up to $M_r \sim 25,000$.

Tertiary Structure Determination Once the chemical shifts that derive from the primary sequence have been assigned, other data, such as the through-space NOE signals, provide information that restrains the possible tertiary structure solutions—information that is referred to as a “restraint.” Restraints are absolutely essential to the prediction of tertiary structure. More than 1,000 restraints are required to predict a structure containing 100 residues or more. Identifying these restraints is labor-intensive, although computational methods are improving. Many restraints are NOE signals that represent protons close in space but distant in the primary sequence. Another type of restraint is the torsion angles between residues, as obtained from the COSY spectrum. A third type is the known geometric restraints of all amino acids, such as chirality, van der Waals radii, and bond lengths.

With sufficient restraints, a structure can be predicted. First, a randomized configuration of the primary sequence is produced using the known geometry of the peptide bond and side-chain atoms. This still leaves a huge number of possible configurations, however, because the backbone $\text{N}-\text{C}_\alpha$ and $\text{C}_\alpha-\text{C}$ bonds are, to some degree, free to rotate. The computer program then tries to fold the chain in a way that best satisfies all the restraints, starting from those nearby in the sequence and proceeding to those that are farther apart. This procedure is repeated several times, but starting with a different randomized configuration of the primary sequence each time. If the structure is substantially the same after each trial, then the number of restraints was sufficient to arrive at a unique solution.

Structures determined by NMR are usually shown as a group of closely related structures (Figure 4-32). The individual structures are arrived at by independent trials and represent the range of conformations consistent with the list of restraints. Although the uncertainty in structures generated by NMR is in part a reflection of the molecular vibrations (commonly called breathing) within a protein structure in solution, the observed variation is also due to errors or insufficiencies in the list of restraints. For example, the areas of greatest variation between different structures of a group usually signify areas where there are fewer restraints. For this reason, in NMR analyses, the total

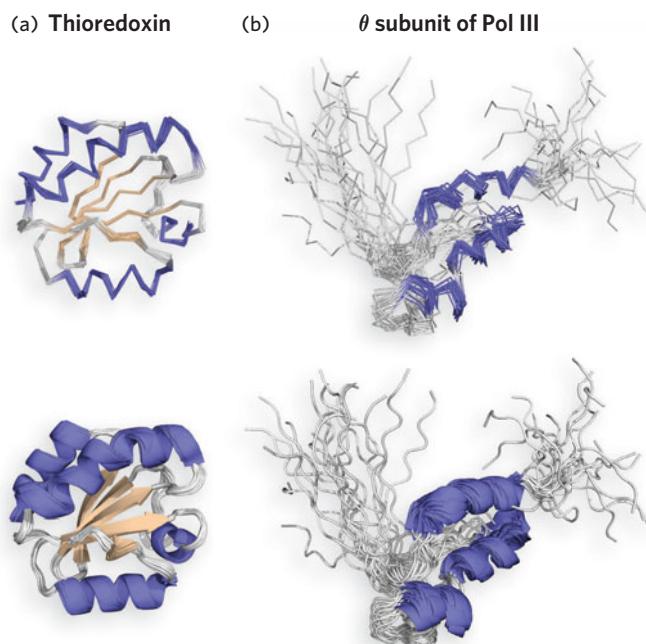


FIGURE 4-32 The structure of two proteins as determined by NMR. (a) Human thioredoxin (M_r , 12,000). Multiple lines represent structures consistent with the restraints from the NMR data. (b) The θ subunit (M_r , 8,600) of DNA polymerase III (Pol III). The divergent models reflect the lack of restraints in disordered areas. The protein contains a region that lacked sufficient restraints to arrive at a unique solution, and probably signifies a region of disordered residues. [Sources: (a) PDB ID 4TRX. (b) PDB ID 1DU2.]

number of restraints is far more important than the accuracy of individual restraints.

Whenever a protein structure is determined by both x-ray crystallography and NMR, the structures generally agree well. In some cases, the precise locations of particular amino acid side chains on the protein exterior are different, often because of effects related to the packing of adjacent protein molecules in a crystal. The two techniques together are at the heart of the rapid increase in the availability of structural information on the macromolecules of living cells.

SECTION 4.5 SUMMARY

- The two methods that reveal protein structure at atomic resolution are x-ray crystallography and nuclear magnetic resonance.
- X-ray crystallography can be applied to a protein of any size, but it requires a protein crystal.
- The diffraction pattern of x rays that have passed through a protein crystal must be recombined into an image mathematically, using the Fourier series.

- The major challenge in analyzing an x-ray diffraction pattern is determining the phase of the diffracted x rays, and three methods are commonly used: isomorphous replacement, multiwavelength anomalous dispersion (MAD), and molecular replacement.
- NMR is performed on proteins in solution and can be applied only to small proteins ($M_r < 25,000$).
- In NMR, the atomic nuclei are excited in a magnetic field and emitted radiation is collected; some of the signals are sensitive to environment and contain structural information.
- In NOESY and COSY two-dimensional NMR, atoms that are covalently bonded or otherwise in close proximity to each other are identified, and the distances between them are used to create a list of restraints from which a structure can be generated.

Unanswered Questions

Numerous protein structures have been determined, and one might think that, by now, researchers had deciphered most of the rules about how proteins fold into their unique shapes. Yet, the information that directs how proteins fold and how they associate with their proper partners in a cell remains largely unknown and continues to be a highly active area of research. Here are some of the many questions being actively pursued.

- What is the “code” in the primary sequence that determines how a protein folds?** We know that the instructions for folding lie in the primary sequence. However, despite the large database of protein

structures, we still do not know how these instructions are read. The problem lies in the relatively small difference in energy between the folded and unfolded states. Researchers remain hopeful that the “rules” of protein folding will someday be understood. Perhaps the accurate prediction of the structure adopted by a given sequence will be obtained by computations that draw on the vast empirical knowledge of structural folding patterns in proteins, combined with theoretical energy computations.

- How do proteins “know” they are to form multiprotein complexes?** Many of the important functions in a cell are performed by multiprotein complexes that act as machines to carry out complicated tasks. These tasks include central jobs such as transcription, replication, and translation. Given the thousands of different proteins in a cell, it is perplexing that particular subunits “know” how to join up, to the exclusion of others, to form these large complexes.
- How do chaperones and chaperonins “know” when to bind a protein?** Proteins that denature, or newly synthesized proteins that require assistance with folding, are targeted by chaperones and chaperonin complexes. However, most proteins contain disordered regions even when they are properly folded. We know little about how chaperones and chaperonins specifically target unfolded proteins. We know even less about how these protein-folding assistants recognize when their job is done, or when to keep working.

How We Know

Sequence Comparisons Yield an Evolutionary Roadmap from Bird Influenza to a Deadly Human Pandemic

Taubenberger, J.K., A.H. Reid, R.M. Lourens, R. Wang, G. Jin, and T.G. Fanning. 2005.

Characterization of the 1918 influenza virus polymerase genes. *Nature* 437:889–893.

Worldwide pandemic outbreaks of flu can lead to millions of deaths. A virus has little to gain by killing its host, and, usually, the more deadly a virus is, the more recently it evolved. The evolution of a deadly virus has been intensively studied for influenza strains that cause pandemics.

A given influenza virus is typically confined to a certain host species, such as birds, horses, pigs, or humans—partly because cell surface receptors that allow entry of a virus are different in each species. However, some influenza strains evolve and jump the species barrier. Trouble for humans starts when the viruses also acquire efficient human-to-human transmissibility. This rare event can result in a worldwide influenza pandemic. About one to three such pandemics occur every century. How do these deadly viruses evolve? This can be determined from their genome sequences and comparisons with other influenza viruses.

The influenza virus genome consists of eight segments of RNA that encode 10 different proteins. Evolution is facilitated in a couple of ways: through errors introduced by the viral replicase, an RNA-dependent RNA polymerase that copies the RNA genome, and through genetic reassortment of RNA segments between two different viruses to form a novel virus. Genetic reassortment of RNA segments occurs when one host animal becomes infected by two different viruses at the same time. The pig, for example, has cell surface receptors that allow infection by both avian and human influenza viruses and thus may act as a “mixing vessel” to produce recombinant influenza viruses.

Comparative sequence analysis of the avian and human viruses responsible for the influenza pandemics of 1957 and 1968 reveals that the viruses evolved by genetic reassortment of two or three genes between an avian and a human virus. Both pandemic viral strains contained an avian *PB1* gene, which encodes part of the viral replicase, a 1:1:1 protein complex composed of products of the *PB1*, *PB2*, and *PA* genes.

A comparison of *PB2* protein sequences of several human and avian influenza viruses is shown in **Figure 1**. Of particular note is the 1918 Brevig Mission strain, which resulted in the worst pandemic so far. Hundreds of millions of people were infected during the 1918–1919 “Spanish flu” pandemic, resulting in the deaths of about 50 million people worldwide. Comparative analysis of the *PB2* gene product of the 1918 Brevig Mission virus and the human and avian viruses shows only five residue

changes from the avian influenza genome. Sequence analysis of all the genes of the 1918 Brevig Mission virus indicates that it did not evolve by genetic reassortment with a second virus but, instead, adapted to humans directly from an avian source.

An avian influenza virus that could infect humans developed in the Far East in recent years, and such outbreaks present a constant threat of another pandemic.

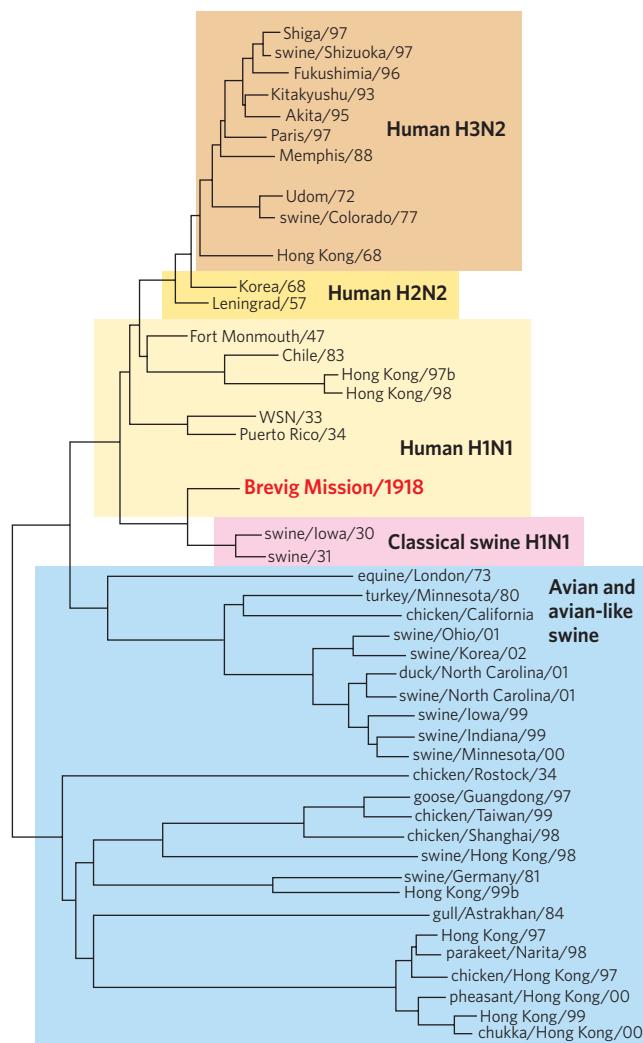


FIGURE 1 This phylogenetic tree of avian and human influenza viruses is based on sequence comparisons of the *PB2* gene. Branch points indicate the place where two *PB2* sequences diverged from a common ancestor. [Source: J. K. Taubenberger et al., *Nature* 437:889–893, 2005.]

We Can Tell That a Protein Binds ATP by Looking at Its Sequence

Koonin, E.V. 1993. A superfamily of ATPases with diverse functions containing either classical or deviant ATP-binding motif. *J. Mol. Biol.* 229:1165–1174.

Saraste, M., P.R. Sibbald, and A. Wittinghofer. 1990. The P-loop: A common motif in ATP- and GTP-binding proteins. *Trends Biochem. Sci.* 15:430–434.

Wouldn't it be handy to be able to figure out what role a protein plays without having to perform complicated experiments? Sequence comparisons can do just that. A classic example is the identification of an ATP/GTP-binding site, based on some characteristic features of ATP/GTP-binding proteins.

Many proteins that bind ATP or GTP have several structural features in common. The P-loop is a conserved, glycine-rich sequence that forms a loop and connects a β strand to an α helix (Figure 2). The P-loop (phosphate-binding loop) interacts with the phosphates of ATP or GTP (Figure 3), and the presence of a P-loop sequence usually indicates that the protein's function involves the use of ATP or GTP. The P-loop is found in association with the nucleotide-binding Rossmann fold motif.

The P-loop is also referred to as a Walker A sequence. It is often followed in the protein's primary sequence (after a variable number of residues) by a Walker B sequence (sometimes called a DEAD box, for the amino acid sequence Asp-Glu-Ala-Asp), which contains acidic residues that bind magnesium ions and assist in ATP or GTP hydrolysis. These two sequences

...(G/A)XXGXGK(T/S)...

FIGURE 2 The P-loop consensus sequence (i.e., the sequence found in many ATP-binding proteins). X means any amino acid; a pair of residues in parentheses means that one can substitute for the other—for example, G/A means either G or A in that position.

are widespread in proteins that function with ATP or GTP. A few examples are eukaryotic Ras family GTP-binding proteins, and eukaryotic and bacterial recombinases and mismatch repair proteins. Despite the widespread use of the P-loop motif, however, some proteins bind ATP using sequences that are unrelated to the Walker motif.

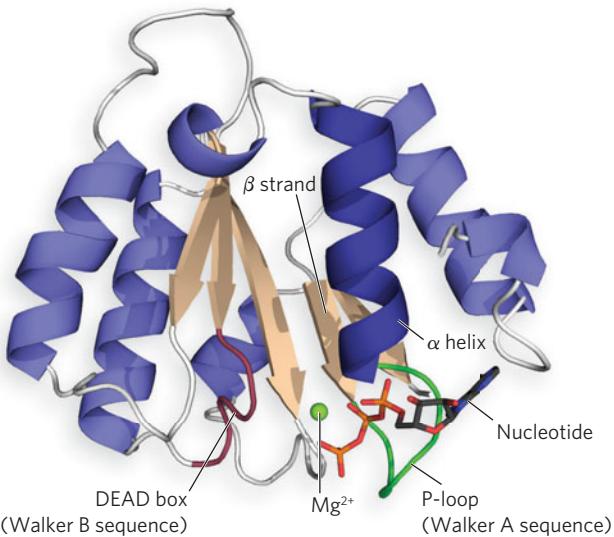


FIGURE 3 This ribbon representation of the ATP-binding RFC2 subunit of the yeast RFC clamp loader, which loads sliding clamps onto DNA for DNA polymerase (see Chapter 11), highlights the features of the Rossmann fold motif that participate in nucleotide binding. The P-loop (green) interacts with the nucleotide, and the DEAD box (red) binds ions that assist in nucleotide hydrolysis. [Source: PDB ID 1SXJ.]

Disulfide Bonds Act as Molecular Cross-Braces to Stabilize a Protein

Matsumura, M., G. Signor, and B.W. Matthews. 1989. Substantial increase of protein stability by multiple disulfide bonds. *Nature* 342:291–293.

In building construction, cross-braces make bridges stronger and walls stronger and sturdier. Proteins, too, utilize cross-braces. A disulfide bond connects two regions of one or more polypeptide chains within a protein and probably acts as a molecular cross-brace to enhance protein stability. But how would a researcher determine whether a disulfide bond really does work as a stabilizing cross-brace?

Brian Matthews's laboratory at the University of Oregon has examined disulfides for cross-brace function by engineering pairs of Cys residues into T4 lysozyme and then measuring their effect on protein stability. (A mutant protein with three disulfide bonds is shown in **Figure 4a**). The proteins were crystallized to detect structural alterations by x-ray diffraction, and protein stability was measured by circular dichroism (CD) spectroscopy at different temperatures. CD spectroscopy measures the amount of secondary structure in a protein and allows the researcher to follow the loss of α -helical content that accompanies protein denaturation.

Mutant proteins with a single disulfide bond had the same structure as wild-type lysozyme, with only small distortions at the replacement sites. Reduction of the disulfide bond (forming two unlinked Cys residues) resulted in a less-stable protein compared with wild-type lysozyme, indicating that Cys substitution had slightly destabilized each mutant protein. But in the oxidized form, the disulfide cross-link greatly increased the stability of several mutant proteins over the wild-type lysozyme. Addition of multiple disulfide cross-links to the wild-type lysozyme gave an additive effect in the stability of the mutant proteins (**Figure 4b**).

These elegant structural and biochemical studies demonstrated that disulfide bonds really do act as molecular cross-braces to enhance the stability of a protein. Antibodies, for example, contain many disulfide bonds, and we may assume that these bonds stabilize the proteins as they circulate in blood, outside the protective confines of the cell membrane.

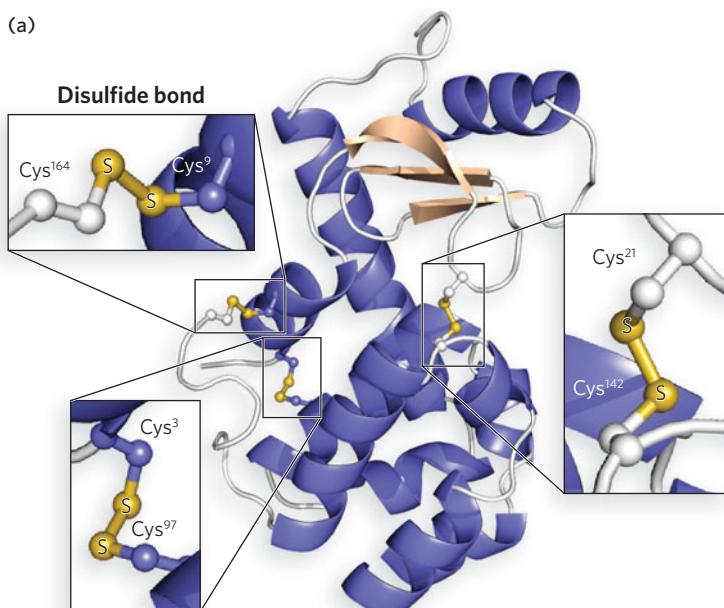
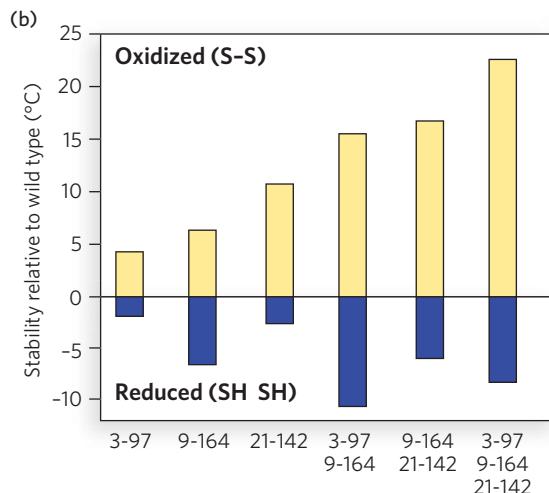


FIGURE 4 (a) Pairs of Cys residues (numbered for their position in the primary sequence) were engineered into T4 lysozyme. This mutant has three disulfide bonds. (b) Additional disulfide bonds stabilize protein structure relative to wild-type lysozyme. The numbers below the plot indicate the positions of



the Cys residues in the disulfide bonds. The rightmost column shows the results for the three-disulfide mutant protein in (a). The stability of these proteins is indicated by the temperature at which the protein loses its activity. [Source: M. Matsumura, G. Signor, and B. W. Matthews, *Nature* 342:291–293, 1989.]

Key Terms

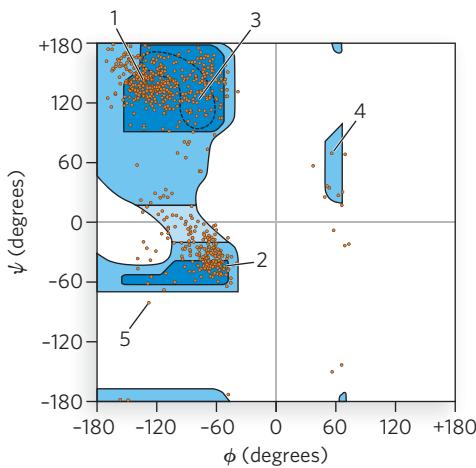
- primary structure, p. 97
 disulfide bond, p. 97
 Ramachandran plot, p. 102
 secondary structure, p. 103
 α helix, p. 104
 β sheet, p. 105
 reverse turn, p. 106
 tertiary structure, p. 107
 quaternary structure, p. 107
 domain, p. 108
 supersecondary structure, p. 109
- motif, p. 109
 Rossmann fold, p. 111
 helix-turn-helix motif, p. 111
 coiled-coil motif, p. 111
 amphipathic helix, p. 112
 oligomer, p. 112
 protomer, p. 112
 molten globule model, p. 117
 chaperone, p. 119
 x-ray crystallography, p. 121
 diffraction pattern, p. 122
- unit cell, p. 122
 Bragg's law, p. 122
 electron density map, p. 123
 phase problem, p. 124
 isomorphous replacement, p. 125
 nuclear magnetic resonance (NMR), p. 125
 nuclear Overhauser effect
 spectroscopy (NOESY), p. 125

Problems

- 1.** Each ionizable group of an amino acid can exist in one of two states, charged or neutral. The electric charge on the functional group is determined by the relationship between the group's pK_a and the pH of the solution. This relationship is described by the Henderson-Hasselbalch equation (see Chapter 3).
- (a) Histidine has three ionizable groups. Write the equilibrium equations for its three ionizations and assign the proper pK_a for each. Draw the structure of histidine in each ionization state. What is the net charge on the histidine molecule in each ionization state?
- (b) Draw the structures of the predominant ionization states of histidine at pH 1, 4, 8, and 12. Note that the ionization state can be approximated by treating each ionizable group independently.
- (c) What is the net charge of histidine at pH 1, 4, 8, and 12? For each pH, will histidine migrate toward the anode (+) or cathode (-) when placed in an electric field?
- 2.** A quantitative amino acid analysis reveals that bovine serum albumin (BSA) contains 0.58% tryptophan (M_r 204) by weight.
- (a) Calculate the *minimum* molecular weight of BSA (i.e., assume there is only one Trp residue per protein molecule).
- (b) The BSA protein is purified and its molecular weight is estimated to be 70,000. How many Trp residues are present in a molecule of serum albumin?
- 3.** A peptide has the following sequence:
- E-H-W-S-G-G-L-R-P-G
- (a) What is the net charge of the molecule at pH 3, 8, and 11? (Use pK_a values for side chains and terminal amino and carboxyl groups as given in Table 4-1.)
- (b) Purification of a peptide or protein is often easier if you understand its ionization properties. At a pH called the isoelectric point (pI), the net charge of the peptide or protein is zero. At lower or higher pH, it has a net positive or net negative charge, respectively. Estimate the pI for the above peptide.
- 4.** A biochemist isolates a peptide hormone with the following sequence:
- ADSERNCQLVILLAWLPGVKVQCALLDRET
- (a) Circle the residues that could contribute a positive charge.
- (b) Put an X above the residues that could contribute a negative charge.
- (c) Underline the residues that could be connected by a disulfide bond.
- 5.** The sequence shown below, with 86 amino acid residues, folds into a β -pleated sheet substructure within a protein. The residues that form the β strands are noted above the sequence, and the numbered residue positions outside the sheet are shown below. What type of β sheet structure is likely to form, parallel or antiparallel? Explain. What types of secondary structure are possible in the sequences between the β strands?



6. Given that the β strands in the β sheet structure of Problem 5 contain hydrophobic residues in most of the even-numbered positions, and that most of the odd-numbered residues have polar R groups, predict how the β sheet structure will fold in three dimensions.
7. Ramachandran plots can help increase the accuracy of models derived from x-ray crystallography data. The plot below was created by measuring ψ and ϕ for each amino acid residue in a 2.2 Å resolution crystal structure. In the plot, which excludes Gly and Pro residues, selected residues (dots) are numbered 1 through 5. What types of secondary structure are most consistent with the location of each of the numbered residues? Which residue(s) would you suspect of being incorrectly modeled into the electron density map? How would your suspicions change if the plot had included Gly residues?



8. Inspect the 20-residue sequence below and predict the most favorable region for an α helix that is 10 residues long. Explain your reasoning. Point out any stabilizing interactions that might occur.

AIPRKKREFICRGFAIRPNT

9. Predict which of the following sequences would bind ATP (or GTP), and explain your answer. (See How We Know.)
- YLFGGTRGVVGKTSIA
 - LLIQALPGMGGDARL
 - LLIFGPPGLPKTTKL
 - FINAGSQGIGKTACL
10. A polypeptide chain has 140 amino acid residues. How long will the polypeptide chain be if it is entirely α -helical? How long will it be if it is one continuous β strand?
11. Compare and contrast four aspects of the use of NMR and x-ray crystallography in protein structure determination.

12. Five proteins are listed below, each a monomer containing the number of amino acid residues indicated. How many domains would you expect each protein to have? Explain your reasoning.

- 70 amino acids
- 110 amino acids
- 150 amino acids
- 200 amino acids
- 250 amino acids

13. For the following 20-residue sequence in a protein, list five amino acid residues that are likely to be buried in the protein, inaccessible to water. Pick five that are good candidates for surface residues.

DLKFITISVGAPVLTREQLLE

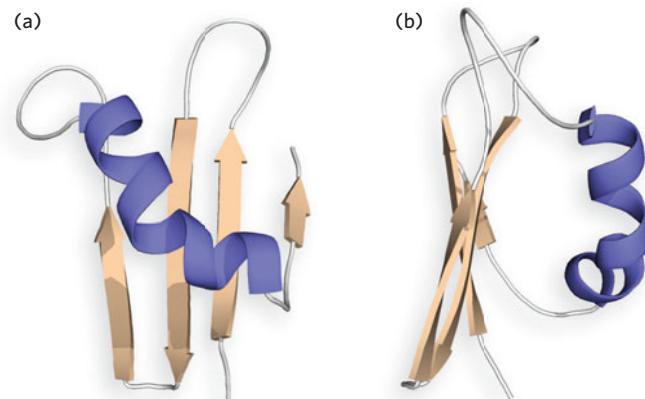
14. Consider an α helix in isolation, apart from the rest of the protein:

NRGAAEGAFCRAN

How would the following amino acid substitutions affect the stability of the helix?

- Change N1 to K.
- Change N1 to E.
- Change R2 to K and E6 to R.
- Change R2 to K and E6 to D
- Change both G3 and G7 to F.
- Change G7 to P.

15. A simple protein structure is shown below, from two different angles. In (a), label the N-terminus and C-terminus and the two β turns. In (b), indicate which side of the β sheet, left or right, is likely to be more hydrophobic.



16. Draw a topology diagram for a 10-stranded, up-and-down β barrel.

Additional Reading

General

Bolen, D.W., and G.W. Rose. 2008. Structure and energetics of the hydrogen-bonded backbone in protein folding. *Annu. Rev. Biochem.* 77:339–362.

Branden, C., and J. Tooze. 1999. *Introduction to Protein Structure*, 2nd ed. New York: Garland Publishing, Inc.

Primary Structure

Doolittle, J., J. Abelson, and M. Simon, eds. 2009. *Molecular Evolution: Computer Analysis and Nucleic Acid Sequences*. Methods in Enzymology, vol. 183 (Amsterdam: Elsevier). A collection of articles on the current state of affairs in computational analysis of DNA and protein sequences and the construction of phylogenetic trees.

Wolf, M.Y., Y.I. Wolf, and E.V. Koonin. 2008. Comparable contributions of structural-functional constraints and expression level to the rate of protein sequence evolution. *Biol. Direct* 3:40–55. A comprehensive resource for comparison of protein sequences and how they relate to evolution.

Zuckerkandl, E., and L. Pauling. 1965. Molecules as documents of evolutionary history. *J. Theor. Biol.* 8:357–366. This report is widely considered to be the founding paper in the field of molecular evolution.

Secondary Structure

Ramachandran, G.N., C. Ramakrishnan, and V. Sasisekharan. 1963. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* 7:95–99.

Rost, B. 2001. Review: Protein secondary structure prediction continues to rise. *J. Struct. Biol.* 134:204–218.

Tertiary and Quaternary Structures

Koonin, E.V., R.L. Tatusov, and M.Y. Galperin. 1998. Beyond complete genomes: From sequence to structure and function. *Curr. Opin. Struct. Biol.* 8:212–217. A review on accuracy in the correlation of sequences with function in genomics.

Ponting, C.P., and R.R. Russell. 2002. The natural history of protein domains. *Annu. Rev. Biophys. Biomol. Struct.* 31:45–71. A description of how structural databases can be used to study evolution.

Protein Folding

Clarke, A.R. 2006. Cytosolic chaperonins: A question of promiscuity. *Mol. Cell* 24:165–167.

Koloday, R., D. Petrev, and B. Honig. 2006. Protein structure comparison: Implications of the nature of “fold space,” and structure and function prediction. *Curr. Opin. Struct. Biol.* 16:393–398.

Determining the Atomic Structure of Proteins

Cavanagh, J., W. Fairbrother, A. Palmer, and A. Skelton. 2007. *Protein NMR Spectroscopy: Principles and Practice*, 2nd ed. San Diego: Academic Press.

Rhodes, G. 2006. *Crystallography Made Crystal Clear: A Guide to Users of Molecular Models*, 3rd ed. San Diego: Academic Press.

Protein Function



Tim Lohman [Source: Courtesy of Timothy Lohman]

DNA ranged from \sim 30 to 77 nucleotides. There was no consensus for a “correct” value and no explanation for why this range was so large.

Les Overman, a graduate student in my laboratory, and I noted that the measured site size varied with both salt (e.g., NaCl) concentration as well as salt type, but showed plateau values of \sim 33 and \sim 65 nucleotides per tetramer at low and high salt concentrations, respectively. Based on these experiments we surmised that the SSB tetramer can bind to single-stranded DNA in at least two distinct modes, a novel suggestion at the time, which was also suggested simultaneously by independent experiments from Jack Griffith’s laboratory. I remember the excitement in the lab when we made sense of these results, since our results also explained apparent discrepancies in a large amount of existing data from numerous laboratories.

—**Tim Lohman**, on his discovery of multiple DNA-binding modes for SSB

Moment of Discovery

One of my earliest “eureka!” moments came during my years as an assistant professor. My lab was studying the DNA-binding properties of the tetrameric *Escherichia coli* single-stranded DNA-binding protein (SSB), which is a central component of DNA replication, recombination, and repair processes. Previous estimates of its occluded site size (the length of DNA with which the protein directly interacts) when bound to single-stranded

5.1 Protein-Ligand Interactions 136

5.2 Enzymes: The Reaction Catalysts of Biological Systems 144

5.3 Motor Proteins 156

5.4 The Regulation of Protein Function 161

Biological information—in the form of the genome of every organism and virus—is the focus of molecular biology, and of this textbook. The packaging, function, and metabolism of this genomic information also involve a wide range of additional macromolecules, including proteins and RNA molecules. The macromolecules involved in DNA and RNA metabolism can be divided into three functional classes. First, some proteins or RNAs simply bind reversibly to nucleic acids; this binding often has a structural or regulatory function. Second, another large class of proteins (and some RNAs) act as biological catalysts, accelerating the reactions needed to sustain and propagate living systems. These are the **enzymes**, as critical to life as are the information-containing DNA and RNA genomes. And third, motor proteins do the work of moving cellular molecules from one location to another, of separating molecules, and of bringing molecules together.

The great majority of macromolecules that carry out these three functions are proteins, although several RNA enzymes are known and are increasingly well understood. The functions of proteins are particularly important to the topics of every chapter in this book, and an introduction to protein function now becomes our focus. The various functions of RNAs are described in Chapters 15 and 16, although the general principles described here apply to RNA molecules as well as proteins. In this chapter we explore each of the three major functions of proteins and conclude with a discussion of protein regulation.

5.1 Protein-Ligand Interactions

Sometimes, a simple reversible interaction of two macromolecules is all that is needed to elicit major changes in a cell or cellular process. A protein bound to another macromolecule can alter structure and/or function in many different ways. A few examples should suffice to illustrate the principle. A protein bound to a specific DNA sequence can regulate the expression of an adjacent or nearby gene. Proteins bound without sequence specificity can condense DNA in a chromosome or package a DNA molecule into a virus head. A protein subunit bound to an enzyme can increase or decrease that enzyme's activity. Polymeric structures built up of many noncovalently linked protein subunits help guide cell division. A protein bound reversibly to a small molecule can act as a transporter, facilitating the movement of that molecule within or between cells. Whether the binding association is prolonged or fleeting, it is often the basis of complex physiological processes, such as gene regulation, immune function, and

cellular signaling. Molecular biology deals with countless such interactions.

Many Proteins Bind to Other Molecules Reversibly

For the proteins that carry out these interactive processes, we can summarize several principles of protein function:

1. The bound molecule is called a **ligand**. A ligand can be any kind of molecule, including another protein. The transient nature of protein-ligand interactions is critical to life, allowing an organism to respond rapidly and reversibly to changing environmental and metabolic circumstances.
2. A ligand binds at a site on the protein called, appropriately, the **binding site**. The binding site is complementary to the ligand in size, shape, charge, and hydrophobic or hydrophilic character. The interaction is specific; the protein discriminates among the thousands of different molecules in its environment and selectively binds only one or a few. A given protein may have separate binding sites for several different ligands. These specific molecular interactions are crucial in maintaining the high degree of order in a living system. Our discussion here excludes the binding of water, which may interact weakly and nonspecifically with many parts of a protein.
3. Proteins exhibit conformational flexibility. Changes in conformation may be subtle, reflecting molecular vibrations and small movements of amino acid residues throughout the protein. A protein flexing in this way is sometimes said to “breathe.” Conformational changes may also be dramatic, with major segments of the protein structure moving as much as several nanometers. Specific conformational flexibility is frequently essential to a protein’s function.
4. Many protein-ligand interactions require a conformational change known as **induced fit**, in which a conformational change in the protein alters a binding site so that it becomes more complementary to the ligand, permitting tighter binding. The induced fit is the adaptation that occurs between protein and ligand.
5. The subunits in a multisubunit protein often exhibit **cooperativity**. A conformational change in one subunit can affect the conformation of other

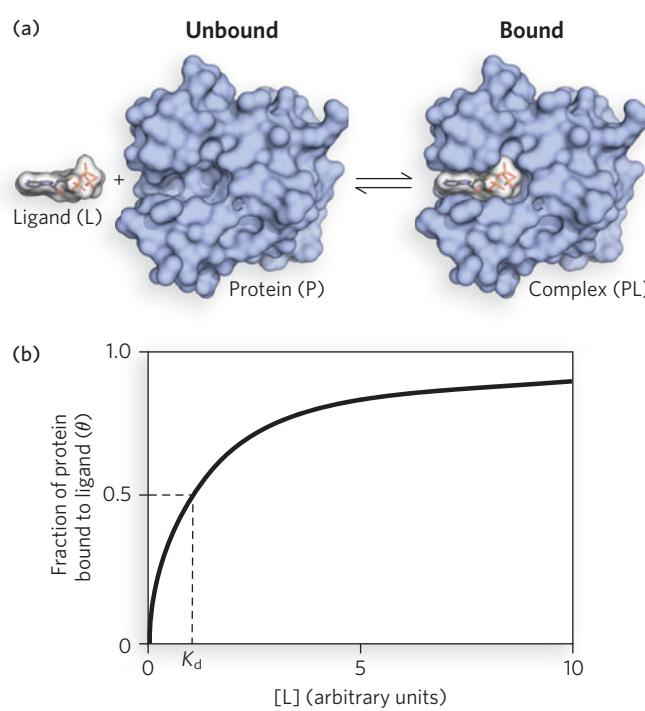
subunits. Thus, a conformational change triggered by the binding of a ligand to one subunit can increase or decrease the affinity of a neighboring subunit for the same ligand, giving rise to cooperative binding.

- The activities of many proteins are subject to regulation. Interactions of ligands and proteins may be regulated, usually through specific interactions with one or more additional ligands. These other ligands may cause conformational changes in the protein that affect the binding of the first ligand.

Protein-Ligand Interactions Can Be Quantified

The function of many proteins depends on their ability not only to bind to a ligand but also to release the ligand when and where it is needed. Function in molecular biology often revolves around a reversible protein-ligand interaction of this type. A quantitative description of this interaction is therefore a central part of many investigations.

In general, the reversible binding of a protein (P) to a ligand (L) can be described by a simple **equilibrium expression** (Figure 5-1a):



The reaction is characterized by an equilibrium constant, K_a , such that:

$$K_a = \frac{[PL]}{[P][L]} = \frac{k_a}{k_d} \quad (5-2)$$

where k_a and k_d are rate constants that describe, respectively, the rate of association and dissociation of the ligand with the protein. K_a is an **association constant** (not to be confused with the K_a that denotes an acid-base association constant; see Chapter 3). It describes the equilibrium between the complex and the separate, unbound components of the complex. The association constant provides a measure of the affinity of the ligand L for the protein P. K_a has units of M^{-1} ; a higher value of K_a corresponds to a higher affinity of the ligand for the protein.

It is more common (and intuitively simpler), however, to consider the **dissociation constant**, K_d , which is the reciprocal of K_a ($K_d = 1/K_a$) and is given in units of molar concentration (M). K_d is the equilibrium constant for the release of ligand. Note that a lower value of K_d corresponds to a higher affinity of ligand for the protein. The relevant expression changes to:

$$K_d = \frac{[P][L]}{[PL]} = \frac{k_d}{k_a} \quad (5-3)$$

We can now consider the binding equilibrium from the standpoint of the fraction, θ (theta), of ligand-binding sites on the protein that are occupied by ligand:

$$\theta = \frac{\text{binding sites occupied}}{\text{total binding sites}} = \frac{[PL]}{[PL]+[P]} \quad (5-4)$$

Substituting $K_a[P][L]$ for $[PL]$ (see Equation 5-2) and rearranging terms gives:

$$\theta = \frac{[L]}{[L] + K_d} \quad (5-5)$$

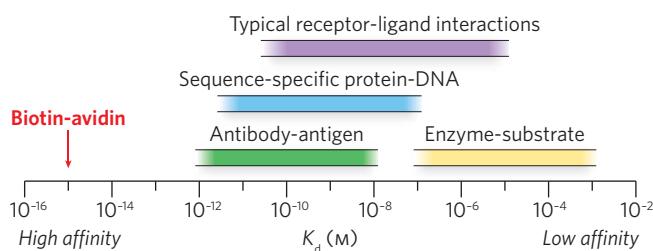
FIGURE 5-1 Ligand binding. (a) Reversible binding of a protein (P) to a ligand (L). The protein shown here is nucleoside diphosphate kinase and the ligand is ATP. (b) The fraction of ligand-binding sites occupied, θ , is plotted against the concentration of free ligand, $[L]$. A hypothetical binding curve is shown. The $[L]$ at which half the available ligand-binding sites are occupied is equivalent to $1/K_a$, or K_d . The curve has a horizontal asymptote at $\theta = 1$ and a vertical asymptote (not shown) at $[L] = -K_d$. [Source: (a) PDB ID 2BEF (ATP-binding domain only).]

Table 5-1 Protein Dissociation Constants

Protein	Ligand	K_d (M)*
Avidin (egg white)	Biotin	1×10^{-15}
Replication protein A (RPA; eukaryotes)	ssDNA	$1 \times 10^{-9} - 1 \times 10^{-11}$
SSB (as the SSB ₆₅ binding mode)	ssDNA	2×10^{-7}
Lactose repressor	dsDNA (nonspecific)	5×10^{-4}
	dsDNA (specific)	2×10^{-11}

Note: SSB, single-stranded DNA-binding protein; ssDNA, single-stranded DNA; dsDNA, double-stranded DNA.

*A reported dissociation constant is valid only for the particular solution conditions under which it was measured. K_d values for a protein-ligand interaction can be altered, sometimes by several orders of magnitude, by changes in the solution's salt concentration, pH, interactions with additional proteins, or many other variables.



The range of dissociation constants for typical interactions in biological systems, denoted by color for each class. A few interactions, such as that between the protein avidin and the enzyme cofactor biotin, fall outside the typical range. The avidin-biotin interaction is so tight that it may be considered irreversible.

Any equation of the form $x = y/(y + z)$ describes a hyperbola, and θ is thus found to be a hyperbolic function of $[L]$ (Figure 5-1b). The fraction of ligand-binding sites occupied approaches saturation asymptotically as $[L]$ increases. The $[L]$ at which half of the available ligand-binding sites are occupied (i.e., $\theta = 0.5$) corresponds to the K_d . When $[L] = K_d$, half of the ligand-binding sites are occupied. As $[L]$ falls below K_d , progressively less of the protein has ligand bound to it. For 90% of the available ligand-binding sites to be occupied, $[L]$ must be nine times K_d .

The mathematics can be reduced to simple statements: K_d equals the molar concentration of ligand at which half the available ligand-binding sites are occupied. At this point, the protein is said to have reached half-saturation with respect to ligand binding. The more tightly a protein binds a ligand, the lower the

concentration of ligand required for half the binding sites to be occupied, and thus the lower the value of K_d . Some representative dissociation constants are given in Table 5-1.

DNA-Binding Proteins Guide Genome Structure and Function

DNA-binding proteins are a key example of proteins that simply bind to a ligand (in this case DNA) reversibly without altering its covalent structure. DNA-binding proteins protect DNA, organize DNA, regulate genes or groups of genes, alter the conformation of DNA, facilitate all aspects of the metabolism of DNA, and ensure the proper segregation of chromosomes during cell division. DNA-binding proteins fall into two principal categories. Some bind to DNA nonspecifically, independent of DNA sequence; others recognize particular DNA sequences and bind tightly at the genomic locations where those sequences occur. The distinction is not absolute. “Nonspecific” DNA-binding proteins often display a measurable bias for binding of DNA sequences with particular features. “Specific” DNA-binding proteins generally exhibit some measurable (albeit much weaker) binding to nonspecific sequences.

The binding of proteins to DNA, and thus the measured K_d of these interactions, is almost always sensitive to parameters such as pH and salt concentration. DNA is a polyelectrolyte (an electrolyte of high molecular weight). In a cell, the negative charges of the phosphates in the DNA backbone are neutralized by interaction with counterions, and there is generally a high concentration of ions such as Mg^{2+} and K^+ surrounding the DNA. As a protein binds to

DNA, some of these ions are released, as are some bound water molecules from both the protein and the DNA. The release of ions and water has both positive and negative effects on the association of a protein with DNA. The positive effects come from a general gain in entropy ($\Delta S >> 0$) as the water and ions are released. The negative effects reflect the interaction energy between the water and ions and the macromolecules, interactions that must be eliminated to make the protein-DNA complex. Protein-DNA interactions are thus rarely as simple as the coming together of two complementary macromolecules. The interactions are affected in important ways by additional interactions of each macromolecule with water and with ions.

Two additional parameters are notable. First, the number of nucleotides (in single-stranded DNA) or base pairs (in double-stranded DNA) that are occluded by the bound protein defines the binding site size, n . This parameter helps determine the number of binding sites on the DNA that might be available to a protein, and a knowledge of n for a particular protein is necessary for any complete description of its binding equilibrium. Second, some DNA-binding proteins, particularly certain proteins that bind to DNA nonspecifically, exhibit cooperativity in binding; that is, when one protein molecule binds, it facilitates the binding of another. Cooperativity can also have important effects on binding equilibria.

The examples that follow are of proteins whose physical and structural properties are particularly well studied.

Nonspecific DNA-Binding Proteins As we'll discuss in Chapter 9, chromosomes are the largest macromolecules found in cells. If chromosomal DNA molecules were laid out linearly, they would typically be hundreds or even thousands of times longer than the cells in which they are housed. The protection and compaction of chromosomal DNA is largely the job of myriad nonspecific DNA-binding proteins found in every cell. These proteins also organize some key chromosomal functions, facilitating DNA replication and repair or guiding chromosomal segregation at cell division.

In most cases, nonspecific DNA-binding proteins exhibit only limited hydrogen-bonding interactions with bases in the DNA. Instead, electrostatic interactions with the negatively charged phosphate groups, hydrogen bonds to the backbone deoxyribose, and nonspecific hydrophobic interactions with the bases predominate, to varying degrees (Figure 5-2). The hydrophobic interactions often take the form of an aromatic amino acid side chain (Tyr, Trp, or Phe) intercalating between two adjacent bases.

An example of a nonspecific DNA-binding protein is the bacterial single-stranded DNA-binding

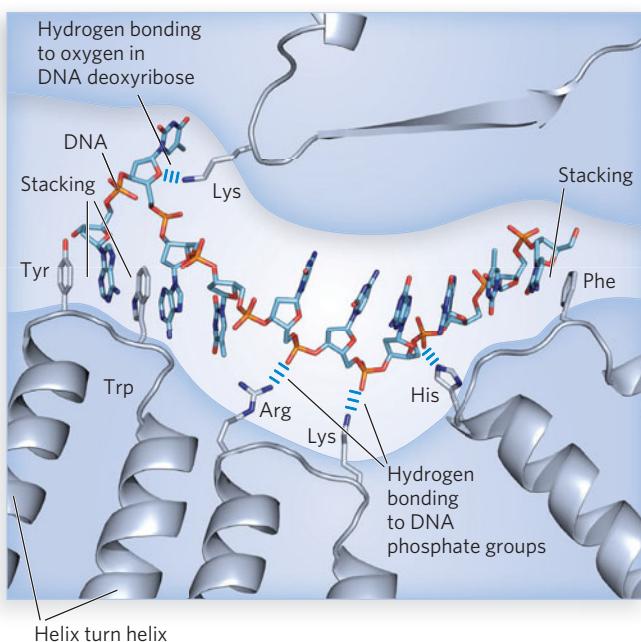


FIGURE 5-2 Nonspecific interactions of proteins with nucleic acids. Electrostatic interactions occur between proteins and the DNA backbone. The charged phosphate groups are exposed at the exterior surface of a single-stranded DNA molecule, where positively charged amino acid side chains (Arg, Lys, His) can interact. Hydrogen bonds occur between the protein and the deoxyribose groups in the DNA backbone. Hydrophobic interactions involving the intercalation of Tyr, Trp, or Phe side chains between two stacked bases are also prominent in many cases of nonspecific DNA-protein binding. Similar interactions occur with RNA.

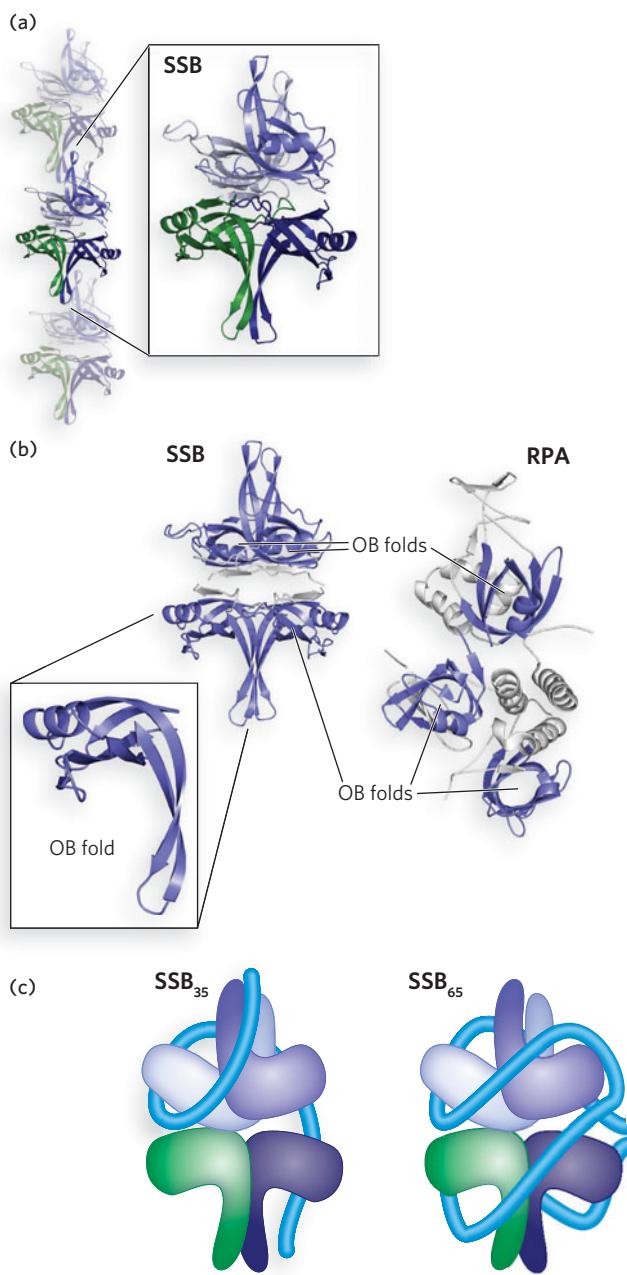


FIGURE 5-3 Binding of single-stranded DNA to single-stranded DNA-binding proteins. (a) Single-stranded DNA-binding protein (SSB) can bind in a filamentous form on single-stranded DNA. A single SSB tetramer is highlighted. (b) The key structural element that functions in single-stranded DNA binding is an OB fold. The basic fold is shown on the left, and the four OB folds in an SSB tetramer are highlighted in purple. The eukaryotic replication protein A (RPA) binds to single-stranded DNA in a similar fashion, using the OB folds in each subunit of the heterotrimer. (c) SSB can bind to single-stranded DNA in multiple binding modes, with the two most prominent modes shown in this schematic. The blue tube represents bound single-stranded DNA. [Sources: (a), PDB ID 1EYG. (b) PDB ID 1L1O.]

protein (SSB). SSB binds and protects single-stranded DNA as it is transiently created during DNA replication and repair. The importance of this protein is illustrated by a simple observation: if SSB function is lost, the cell dies. The structure of the SSB from *Escherichia coli* is typical for bacteria, consisting of four identical subunits around which the single-stranded DNA wraps (Figure 5-3a). Each subunit contains one oligonucleotide/oligosaccharide-binding fold (OB fold), a structural unit that often binds single-stranded DNA. The OB fold is common to all proteins in the SSB class (Figure 5-3b), as well as many others that associate with single-stranded DNA as part of their function.

In addition to binding to single-stranded DNA, bacterial SSBs interact directly with a range of other proteins in DNA metabolism. The *E. coli* SSB interacts with at least 15 different proteins, thereby helping organize the functions of DNA replication and repair (Table 5-2). Most, if not all, of these interactions occur through a conserved C-terminus with multiple negatively charged amino acid residues among the final 8 to 9 residues of the polypeptide.

The SSBs are found in every class of organism, and they always play essential roles in DNA metabolism. The eukaryotic SSB is called **replication protein A (RPA)**. It consists of three different subunits (i.e., is a heterotrimer) containing a total of six OB folds among them. Its function is quite similar to that of the bacterial SSBs, and it also interacts with a range of other proteins as part of its function in eukaryotic DNA metabolism.

In the test tube, the *E. coli* SSB binds to single-stranded DNA according to several different binding modes, depending on the concentrations of salt and protein. Two of these binding modes are notable. At relatively low concentrations of salt, SSB binds with a binding site size (n) of ~ 35 nucleotides and with a very high degree of cooperativity between tetramers. In this mode, called SSB₃₅, the single-stranded DNA is bound to two of the four subunits in each SSB tetramer (Figure 5-3c), and the tetramers are arranged on the DNA as a fairly regular filamentlike structure (see Figure 5-3a). When the salt concentration is higher, the SSB₆₅ ($n = 65$ nucleotides) binding mode predominates. Here, the single-stranded DNA is wrapped around all four SSB subunits (see Figure 5-3c); the cooperativity between SSB tetramers is reduced, and filaments form less readily. The SSB binding modes affect SSB function and interactions with other proteins *in vitro* and presumably *in vivo* (see Moment of Discovery). In both cases, SSB binds to single-stranded

Table 5-2 Proteins That Interact with Bacterial Single-Stranded DNA-Binding Protein

Protein	Function
χ subunit of DNA polymerase III	DNA replication
DnaG primase	DNA replication
RecQ helicase	Recombinational DNA repair
RecJ nuclease	Recombinational DNA repair
RecG helicase	Recombinational DNA repair
RecO recombination mediator	Recombinational DNA repair
PriA replication restart protein	Replication restart after repair
PriB replication restart protein	Replication restart after repair
Exonuclease I	DNA replication and repair
Uracil DNA glycosylase	DNA repair
DNA polymerase II	Mutagenic replication under stress
DNA polymerase V	Mutagenic replication under stress
Exonuclease IX	DNA repair
Bacteriophage N4 virion RNA polymerase	Viral nucleic acid metabolism

DNA with a combination of electrostatic interactions with the phosphoribose backbone and intercalation of particular Trp and Phe side chains between adjacent DNA bases. The *E. coli* SSB binds tightly to single-stranded DNA, with measured K_d values generally in the range of 8 to 700 nm, depending on the solution conditions.

Many other proteins that bind to single-stranded or double-stranded DNA with little specificity also bind such that the DNA is wrapped or bent around the protein. For example, duplex DNA wraps tightly around the nucleosomes of eukaryotic chromosomes (see Chapter 10). Although the histone proteins that make up a nucleosome are considered nonspecific DNA-binding proteins, the positioning of nucleosomes on double-stranded DNA is not entirely random. In particular, DNA sequence elements that facilitate the bending or wrapping of double-stranded DNA around a protein, such as regions with several contiguous A=T base pairs, can have strong effects on the locations of bound nucleosomes along the DNA.

Specific DNA-Binding Proteins Proteins that bind with an enhanced affinity to particular DNA sequences are critical to the regulation of many processes in DNA metabolism. Many of these proteins regulate the expression of genes. Their affinity for specific target sequences is roughly 10^4 to 10^6 times their affinity for any other DNA sequence. Most regulatory proteins have discrete DNA-binding domains

containing substructures that interact closely and specifically with the DNA. These binding domains usually include one or more of a relatively small group of recognizable and characteristic structural motifs (see Chapter 4).

To bind to specific DNA sequences, regulatory proteins must recognize and distinguish surface features on the DNA. Most of the chemical groups that differ among the four bases and thus permit discrimination between base pairs are hydrogen-bond donor and acceptor groups exposed in the major groove of DNA. These interactions are illustrated by the binding of a homeodomain of a eukaryotic regulatory protein with its DNA binding site (Figure 5-4; see Figure 3-3 for DNA structural features). Most of the protein-DNA contacts that impart specificity are hydrogen bonds. A notable exception is the nonpolar surface near C-5 of pyrimidines, where thymine is readily distinguishable from cytosine by its protruding methyl group. Protein-DNA contacts are also possible in the minor groove of DNA, but the hydrogen-bonding patterns here generally do not allow ready discrimination between base pairs.

In specific DNA-binding proteins, the amino acid side chains that most often hydrogen-bond to bases in the DNA are those of Asn, Gln, Glu, Lys, and Arg residues. Is there a simple recognition code in which a particular amino acid always pairs with a particular base? Two hydrogen bonds can form between Gln or Asn and the N⁶ and N-7 positions of adenine but not any other base. An Arg residue can form two hydrogen bonds

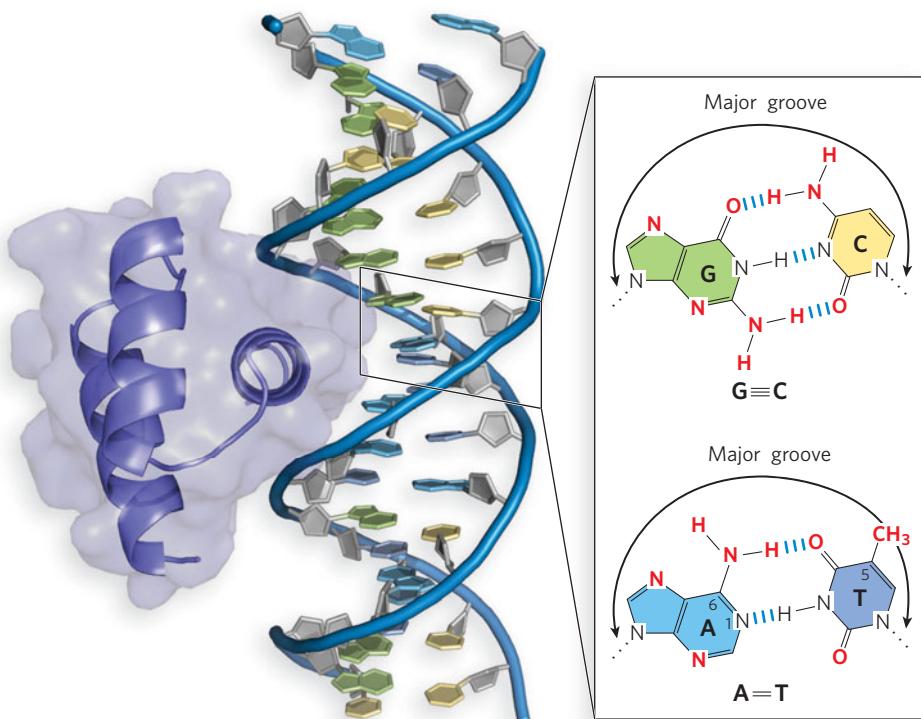


FIGURE 5-4 Groups in DNA that can guide specific protein binding. The eukaryotic DNA-binding protein Engrailed interacts with the major groove of DNA. Shown at right are functional groups on base pairs displayed in the

major and minor grooves. Red indicates groups that can be used for base-pair recognition by proteins. Most specific binding is through interactions with the major groove.
[Source: PDB ID 2HDD.]

with N-7 and O⁶ of guanine (Figure 5-5). Examination of the structures of many DNA-binding proteins, however, has shown that proteins can recognize each base pair in more than one way, leading to the conclusion that there is no simple amino acid–base code. For some proteins, the Gln–adenine interaction can specify A=T base pairs; for others, a van der Waals pocket for the methyl group of thymine can recognize A=T base pairs. As yet, researchers cannot examine the structure of a DNA-binding protein and predict the DNA sequence to which it binds.

An example of a specific DNA-binding protein is the well-studied lactose (Lac) repressor of *E. coli* (see How We Know). This protein is part of a regulatory network that controls the expression of three consecutive genes in the *E. coli* chromosome, all of them involved in some aspect of lactose metabolism. The three genes are transcribed together in a unit described as an operon (Figure 5-6), and regulation of transcription occurs in and around a specific sequence, called the Lac operator, where the Lac repressor binds to the DNA. When the Lac repressor is bound to the Lac operator, transcription of the operon genes is blocked. The lactose operon

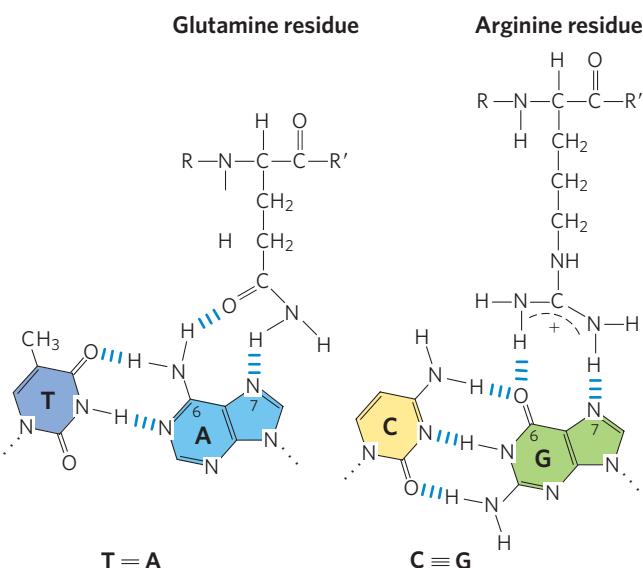


FIGURE 5-5 DNA-protein binding. Two examples of amino acid–base pair interactions that have been observed in DNA-protein binding. An Asn residue can participate in the same type of interaction as the Gln residue.

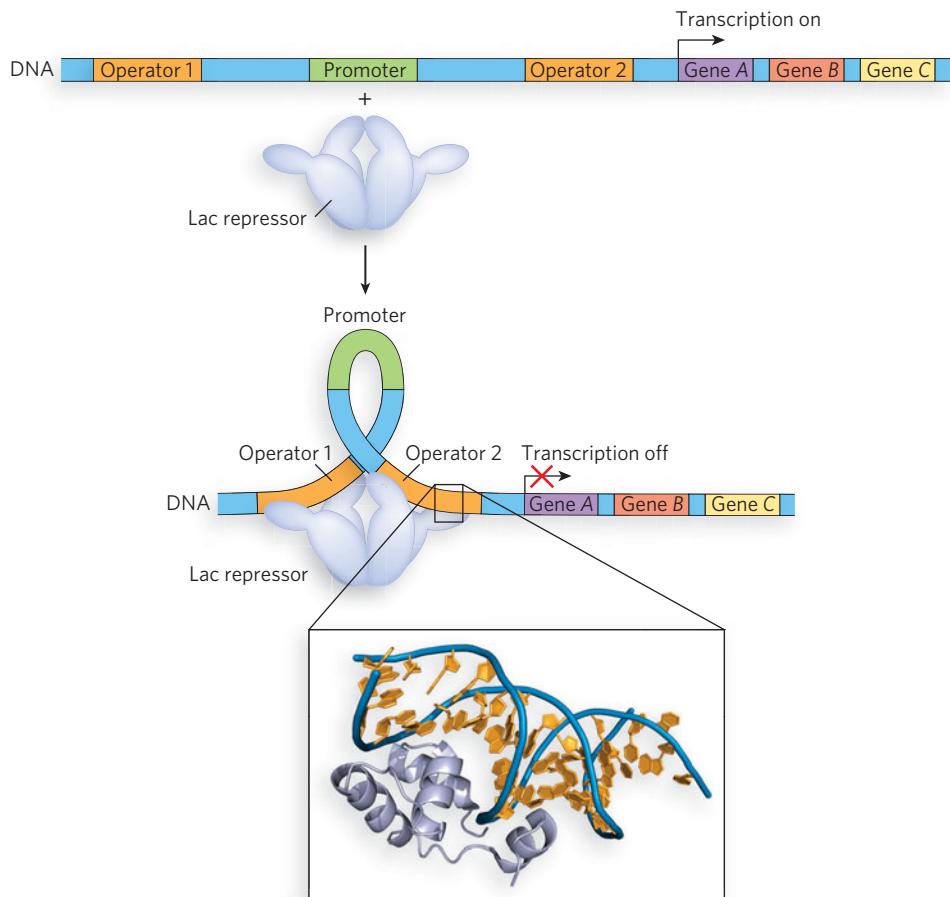


FIGURE 5-6 Simplified representation of the *lac* operon.

When unbound by the Lac repressor, RNA polymerase can bind to the promoter and transcribe several linked genes, here labeled A, B, and C (see Chapter 20). The Lac repressor binds to two operators and shuts down transcription,

apparently forming a loop in the DNA that prevents RNA polymerase from binding to the promoter. The helix-turn-helix motif of the repressor protein binds specifically in the major groove of the operator recognition sequences.

is described in detail in Chapter 20, and we use the Lac repressor here simply to illustrate a common property of sequence-specific DNA-binding proteins. The DNA binding sites for regulatory proteins are often inverted repeats of a short DNA sequence where multiple (usually at least two) subunits of a regulatory protein bind cooperatively. The Lac repressor functions as a tetramer, with two dimers tethered together at the end distant from the DNA-binding sites.

An *E. coli* cell usually contains about 20 tetramers of the Lac repressor. Each of the tethered dimers separately binds to an operator sequence, in contact with 17 base pairs of a 22 base pair region in the *lac* operon (see Figure 5-6). Each of the tethered dimers can independently bind to an operator sequence. The tetrameric Lac repressor binds to two proximal operator sequences in vivo with an estimated K_d of about 10^{-10} M.

The repressor discriminates between the operators and other sequences by a factor of about 10^6 , so binding to these few dozen base pairs among the 4.6 million or so of the *E. coli* chromosome is highly specific.

Specific DNA-binding proteins interact with their specific DNA sequences through a particular part of the protein structure, referred to as a DNA-binding motif. Several common DNA-binding motifs are now known (see Chapter 4). The Lac repressor interacts with DNA via a **helix-turn-helix motif**, a DNA-binding motif that is crucial to the interaction of many bacterial regulatory proteins with DNA; similar motifs occur in some eukaryotic regulatory proteins (see Figure 5-2). The helix-turn-helix motif comprises about 20 amino acid residues in two short α -helical segments, each 7 to 9 residues long, separated by a β turn. This structure generally is not stable by itself; it is simply the reactive portion of a

somewhat larger DNA-binding domain. One of the two α -helical segments is called the recognition helix, because it usually contains many of the amino acid residues that interact with the DNA in a sequence-specific way. This α helix is stacked on other segments of the protein structure so that it protrudes from the protein surface.

One set of amino acid residues in the recognition helix of the Lac repressor's helix-turn-helix domain participates in both nonspecific and specific interactions with DNA. The nonspecific interactions, even though they are weaker, usually occur first and play an important role in accelerating the search for the specific DNA binding site. In the nonspecific binding mode, these residues interact electrostatically with the DNA's phosphoribose backbone (Figure 5-7). When bound to the specific recognition sequences in the Lac operator, the recognition helix is positioned in, or nearly in, the major groove. A network of specific hydrogen bonds and hydrophobic interactions governs the stability of this complex.

There are many proteins that bind to specific sequences in a nucleic acid. We'll encounter several of these in later chapters.

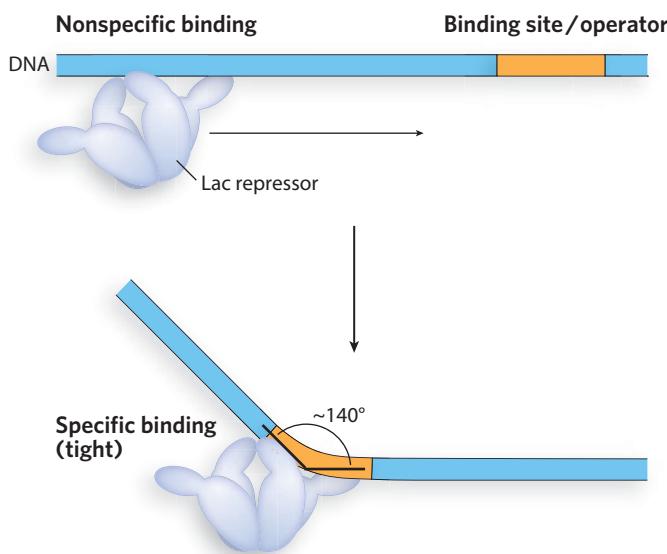


FIGURE 5-7 Nonspecific versus specific DNA binding. The Lac repressor interacts transiently and nonspecifically with the DNA phosphoribose backbone during the search for its specific DNA binding site. When the sequence of its normal binding site is located, the repressor interacts specifically with the nucleotide bases in that site. Binding results in DNA bending.

SECTION 5.1 SUMMARY

- Many proteins bind reversibly to other molecules, known as ligands. Ligand binding often involves protein conformational changes, in the process of induced fit.
- Ligand binding can be quantified, and the key parameter is the dissociation constant, K_d , which is the concentration of ligand at which half of the protein's binding sites are occupied by ligand. A lower K_d corresponds to higher affinity for (tighter binding to) the ligand.
- A DNA- or RNA-binding protein may bind to a nucleic acid either nonspecifically or in a sequence-dependent manner. Nonspecific binding usually involves interactions with the phosphoribose backbone of the nucleic acid. Specific DNA or RNA binding requires an interaction of amino acid side chains in the protein with functional groups in the nucleic acid bases, especially those exposed in the major groove of DNA.

5.2 Enzymes: The Reaction Catalysts of Biological Systems

Rare indeed is the organic chemical reaction that proceeds unaided at a rate sufficient to support living systems. Enzymes have extraordinary catalytic power, often far greater than that of synthetic or inorganic catalysts. They have a high degree of specificity for their substrates, they substantially accelerate chemical reactions, and they function in aqueous solutions under very mild conditions of temperature and pH. Few nonbiological catalysts have all these properties.

Enzymes are central to every cellular process. Acting in organized sequences, they catalyze the hundreds of stepwise reactions that degrade nutrient molecules, conserve and transform chemical energy, make biological macromolecules from simple precursors, and carry out the various processes of DNA and RNA metabolism.

In molecular biology, the study of enzymes has immense practical importance. In some diseases, especially hereditary genetic disorders related to DNA or RNA metabolism, there may be a deficiency or even a total absence of one or more enzymes. Other disease conditions may be caused by excessive activity of an enzyme. Many medicines act through interactions with enzymes. Furthermore, researchers can isolate and harness enzyme functions to suit their purposes in the

Table 5-3 Inorganic Elements as Cofactors for Enzymes and Regulatory Proteins

Cofactor	Enzyme	Function
Cu ²⁺	Superoxide dismutase	Cellular protection from reactive oxygen species
Fe ²⁺ or Fe ³⁺	AlkB	DNA repair
Mg ²⁺	RecA protein	Recombinational DNA repair
	ATPases (all)	Many functions
	Nucleases	DNA cleavage
	DNA and RNA polymerases	Nucleic acid synthesis
Mn ²⁺	Ribonucleotide reductase	Biosynthesis of deoxynucleotides
Mo ⁶⁺	Molybdate sensor protein	Gene regulation
	Certain bacterial riboswitches	Gene regulation
Zn ²⁺	Many DNA-binding proteins	Gene regulation

laboratory. The set of methods collectively described as “biotechnology” is made possible by our understanding of the enzymes of DNA and RNA metabolism (see Chapter 7).

Enzymes Catalyze Specific Biological Reactions

With the exception of a small group of catalytic RNA molecules (see Chapter 16), all enzymes are proteins. Their catalytic activity depends on the integrity of their native protein conformation. Enzymes, like other proteins, have molecular weights ranging from about 12,000 to more than 1 million. Some enzymes require for their activity no chemical groups other than their amino acid residues. Others require an

additional chemical component called a **cofactor**, either one or more inorganic metal ions (Table 5-3) or a complex organic or metallo-organic molecule called a **coenzyme**, which acts as a transient carrier of specific functional groups (Table 5-4). Most coenzymes are derived from vitamins, organic nutrients required in small amounts in the human diet. Some enzymes require *both* a coenzyme and one or more metal ions for activity. A coenzyme or inorganic cofactor that is very tightly or even covalently bound to the enzyme protein is known as a **prosthetic group**. A complete, catalytically active enzyme, together with its bound coenzyme and/or inorganic cofactor, is referred to as a **holoenzyme**. The protein part of such an enzyme is called the **apoenzyme** or **apoprotein**.

Table 5-4 Coenzymes: Transient Carriers of Specific Atoms or Functional Groups

Coenzyme*	Examples of Chemical Group(s) Transferred	Dietary Precursor in Mammals
Biocytin	CO ₂	Biotin
Coenzyme A	Acyl groups	Pantothenic acid and other compounds
5'-Deoxyadenosylcobalamin (coenzyme B ₁₂)	H atoms and alkyl groups	Vitamin B ₁₂
Flavin adenine dinucleotide (FAD)	Electrons	Riboflavin (vitamin B ₂)
Lipoate	Electrons and acyl groups	Not required in diet
Nicotinamide adenine dinucleotide (NAD)	Hydride ion (:H ⁻)	Nicotinic acid (niacin)
Pyridoxal phosphate	Amino groups	Pyridoxine (vitamin B ₆)
Tetrahydrofolate	One-carbon groups	Folate
Thiamine pyrophosphate	Aldehydes	Thiamine (vitamin B ₁)
S-Adenosylmethionine (adoMet)	Methyl groups	Not required in diet

*The structures and modes of action of these coenzymes are described in most biochemistry textbooks.

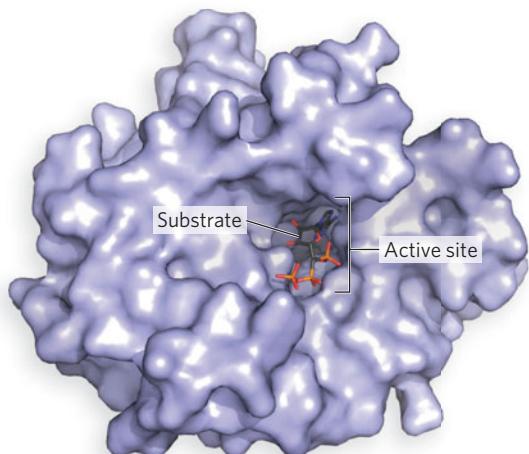


FIGURE 5-8 Binding of a substrate to an enzyme at the active site. This is the enzyme T4 RNA ligase, with a bound substrate, ATP (red). Active sites are typically pockets in the surface of enzymes. [Source: PDB ID 2C5U.]

An enzyme catalyzes a reaction by providing a specific environment in which the reaction can occur more rapidly. Here are some key principles of enzyme-catalyzed reactions:

1. A molecule that undergoes an enzyme-catalyzed reaction is referred to as a **substrate**. A substrate differs from a ligand in that it can undergo a chemical transformation while bound to the enzyme.
2. The substrate interacts with the enzyme in a pocket known as the **active site** (Figure 5-8). The active site is typically lined with multiple chemical groups—amino acid side chains, metal ion cofactors, and/or coenzymes—all oriented to facilitate the reaction.
3. An enzyme-catalyzed reaction is highly specific for that particular reaction. An active site that is set up to catalyze one reaction with one substrate will not interact well with other substrates. Specificity is an important property of every enzyme. The catalysis of a different reaction requires a different enzyme.
4. Catalysis often requires conformational flexibility. As we have seen with proteins that reversibly bind to ligands, conformational changes can have an essential role in enzyme function. Induced fit and cooperativity also play roles in enzyme catalysis.
5. Many enzymes are regulated. The panoply of enzymes available to a given cell confers the

opportunity not just to accelerate reactions but to control them. In this way, cellular metabolism can be modulated as resources and circumstances demand.

The enormous and highly selective rate enhancements achieved by enzymes can be explained by the many types of covalent and noncovalent interactions between enzyme and substrate. Chemical reactions of many types may take place between substrates and the functional groups (specific amino acid side chains, metal ions, and coenzymes) on enzymes. The particular reactions that occur depend on the requirements of the overall reaction to be catalyzed. An enzyme's catalytic functional groups may form a transient covalent bond with the substrate and activate it for reaction. Or a group may be transiently transferred from the substrate to the enzyme. The most common type of group transfer involves the transfer of protons between ionizable amino acid side chains in the active site and groups on the substrate molecule, a process called general acid and base catalysis (Figure 5-9). In the enzymes important to molecular biology, phosphoryl group transfers are also common. In many cases, these group transfer reactions occur only in the enzyme active site. The capacity to facilitate multiple interactions and transfers of this type, sometimes all at once, is one of the factors contributing to the rate enhancements provided by enzymes.

Covalent interactions are only part of the story, however. Much of the energy required to increase the reaction rate is derived from weak, noncovalent interactions between substrate and enzyme, including hydrogen bonds and hydrophobic and ionic interactions. The formation of each weak interaction is accompanied by the release of a small amount of free energy that stabilizes the interaction. The energy derived from enzyme-substrate interaction is called **binding energy**, ΔG_B . Its significance extends beyond a simple stabilization of the enzyme-substrate interaction. *Binding energy is a major source of the free energy used by enzymes to increase the rates of reactions.*

In the context of nucleic acid metabolism, one type of weak interaction merits special mention. Ionic interactions can include interactions of bound metals (such as Mg^{2+} , Mn^{2+} , and Fe^{2+} or Fe^{3+} ions) and substrates. About one-third of all enzymes utilize metals in their catalytic mechanisms, and that proportion is much higher for the enzymes that act on DNA and RNA. For example, the active sites of DNA and RNA polymerases universally feature two metal ions, usually two Mg^{2+} ions, that help orient substrates and

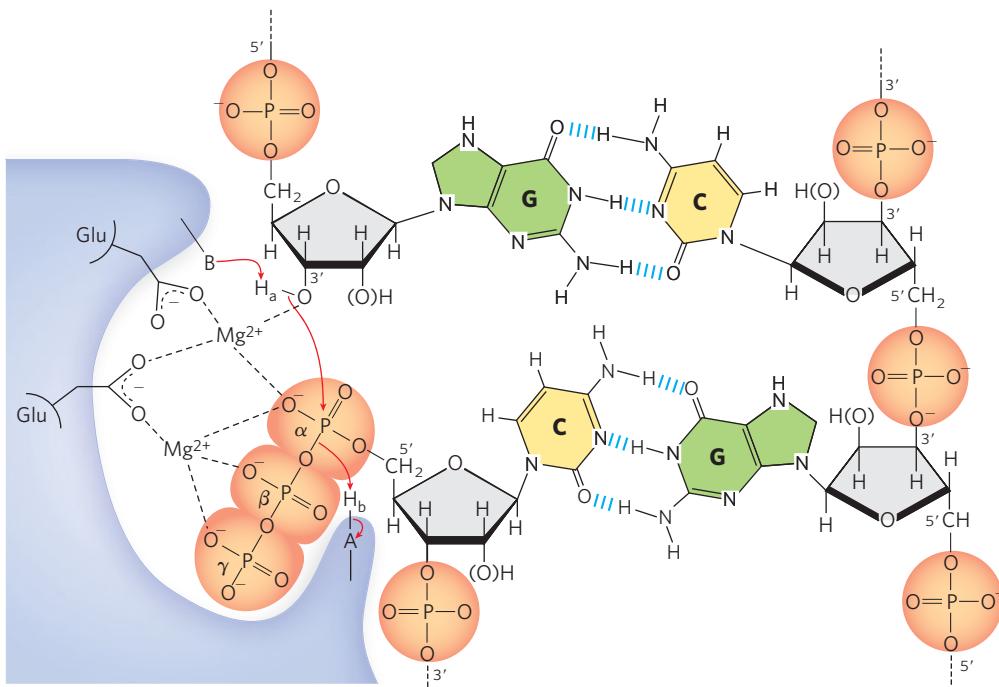


FIGURE 5-9 An enzyme-catalyzed reaction. Shown here is the key step in the formation of a new phosphodiester bond in the active site of a DNA or RNA polymerase. The end of a growing chain of nucleic acid (the primer chain) is at the top left, and an incoming (deoxy)nucleoside triphosphate is at the lower left. The reaction begins with general base catalysis by an active-site residue (B) that abstracts a proton H_a from the attacking 3' hydroxyl at the end of the primer chain. The oxygen of the hydroxyl group concurrently attacks the phosphorus of the α -phosphoryl

group of the nucleoside triphosphate, displacing pyrophosphate (PP_i). The pyrophosphate is protonated by another active-site residue (usually a Lys, shown here as A), an example of general acid catalysis, which facilitates ejection of the PP_i. Two metal ions, usually two Mg²⁺, are in the active site. One metal ion lowers the pK_a of the primer 3' hydroxyl to facilitate the general base catalysis. The other metal ion coordinates with and orients oxygens of the triphosphate and also aids catalysis by stabilizing the transition state of the reaction.

facilitate the overall reaction in multiple ways (see Figure 5-9). Mg²⁺ ions play key roles at the active sites of a wide range of enzymes discussed in later chapters.

Enzymes Increase the Rate of a Reaction by Lowering the Activation Energy

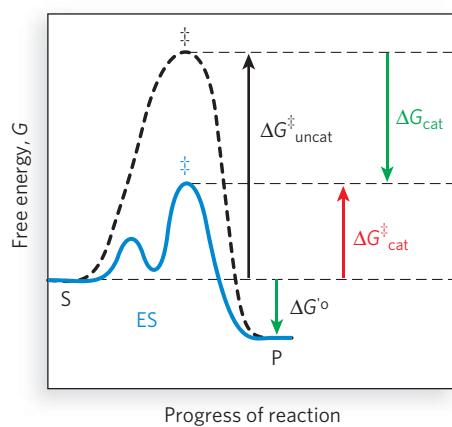
A simple enzyme reaction might be written like this:



where E, S, and P represent the enzyme, substrate, and product, and ES and EP are transient complexes of the enzyme with the substrate and with the product.

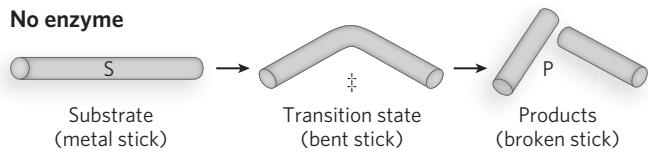
To understand catalysis, we must recall the important distinction between reaction equilibria and reaction rates. The equilibrium between S and P reflects the difference in the free energies of their ground states. Any reaction, such as S ⇌ P, can be described by a reaction coordinate diagram, a picture of the energy changes during the reaction (Figure 5-10a). Energy in biological systems is described in terms of free energy, G (see Chapter 3). In the diagram, the free energy of the system is plotted against the progress of the reaction (the reaction coordinate). In this example, the free energy of the ground state of P is lower than that of S, so the **biochemical standard free-energy change**, or ΔG° , for the reaction is negative and the equilibrium favors P.

(a) Reaction coordinate diagram

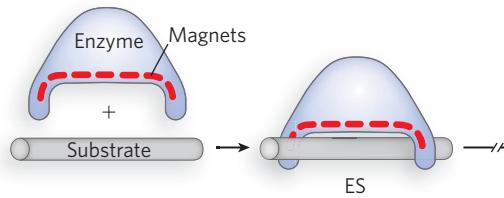


(b)

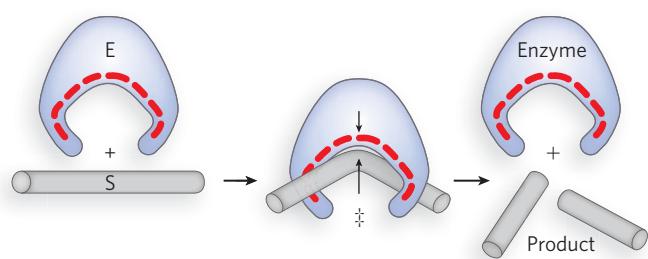
No enzyme



Enzyme complementary to substrate



Enzyme complementary to transition state



KEY CONVENTION

To describe the free-energy changes for reactions, chemists define a standard set of conditions (temperature 298 K; partial pressure of each gas 1 atm, or 101.3 kPa; concentration of each solute 1 M) and express the free-energy change for a reacting system under these conditions as ΔG° , the standard free-energy change. Because living systems commonly involve H^+ concentrations far below 1 M, biochemists and molecular biologists define a biochemical standard free-energy change, $\Delta G'^\circ$, the standard free-energy change at pH 7.0; we use this definition throughout the book.

FIGURE 5-10 The use of noncovalent binding energy to accelerate an enzyme-catalyzed reaction. (a) A reaction coordinate diagram. The free energy of a system is plotted against the progress of the reaction $S \rightarrow P$. This kind of diagram describes the energy changes during the reaction; the horizontal axis (reaction coordinate) reflects the progressive chemical changes (e.g., bond breakage or formation) as S is converted to P . The activation energies, ΔG^\ddagger , for the $S \rightarrow P$ and $P \rightarrow S$ reactions are indicated. ΔG° is the overall biochemical standard free-energy change in the direction $S \rightarrow P$. The ES intermediate occupies a minimum in the energy progress curve of the enzyme-catalyzed reaction. The terms $\Delta G^\ddagger_{\text{uncat}}$ and $\Delta G^\ddagger_{\text{cat}}$ correspond to the activation energy for the uncatalyzed reaction (black, dashed curve) and the overall activation energy for the catalyzed reaction (blue, solid curve), respectively. The activation energy is lower by the amount ΔG_{cat} when the enzyme catalyzes the reaction. (b) An imaginary enzyme (stickase) designed to catalyze the breaking of a metal stick. Before the stick is broken, it must be bent (transition state). Magnetic interactions replace weak enzyme-substrate bonding interactions. A stickase with a magnet-lined pocket structurally complementary to the stick (substrate) stabilizes the substrate (middle). Bending is impeded by the magnetic attraction between stick and stickase. An enzyme with a pocket complementary to the reaction transition state helps destabilize the stick (bottom), contributing to catalysis. The binding energy of the magnetic interactions compensates for the increase in free energy needed to bend the stick. In enzyme active sites, weak interactions that occur only in the transition state aid in catalysis.

A favorable equilibrium does not mean that the $S \rightarrow P$ conversion will occur at a fast or even detectable rate. An unfavorable equilibrium does not mean that the reaction will be slow. Instead, the *rate* of a reaction depends on the height of the energy hill that separates the product from the substrate. At the top of this hill lies the **transition state** (denoted by \ddagger in Figure 5-10a). The transition state is not a stable species, but is a transient moment when the alteration in the substrate has reached a point corresponding to the highest energy in the reaction coordinate diagram. The difference between the energy levels of the ground state and the transition state is the **activation energy**, ΔG^\ddagger . A higher activation energy corresponds to a slower reaction.

The function of a catalyst is to increase the *rate* of a reaction. Catalysts do not affect reaction equilibria. Enzymes are no exception. The bidirectional arrows in Equation 5-6 make this point: any enzyme that catalyzes

the reaction $S \rightarrow P$ also catalyzes the reaction $P \rightarrow S$. The role of enzymes is to *accelerate* the interconversion of S and P . The enzyme is not used up in the process, and the equilibrium point is unaffected. However, the reaction reaches equilibrium much faster when the appropriate enzyme is present, because the rate of the reaction is increased. Enzymes increase reaction rates by lowering the activation energy of the reaction. To achieve this, enzymes utilize noncovalent and covalent interactions in somewhat different ways.

Two fundamental and interrelated principles provide a general explanation for how enzymes use noncovalent binding energy to accelerate a reaction:

1. Much of the catalytic power of an enzyme is ultimately derived from the free energy released in forming many weak bonds and interactions between the enzyme and its substrate. This binding energy contributes to specificity as well as to catalysis.
2. Weak interactions are optimized in the reaction transition state; enzyme active sites are complementary not to substrates per se but to the transition states through which substrates pass as they are converted to products during the reaction (see **Figure 5-10b**).

When the enzyme active site is complementary to the reaction transition state, some of the noncovalent interactions between enzyme and substrate occur only in the transition state. The free energy (binding energy) released by the formation of these interactions partially offsets the energy required to reach the top of the energy hill. The summation of the unfavorable (positive) activation energy ΔG^\ddagger and the favorable (negative) binding energy ΔG_B results in a lower net activation energy (see **Figure 5-10a**). Even on the enzyme, the transition state is not a stable species but a brief point in time that the substrate spends atop an energy hill. The enzyme-catalyzed reaction is much faster than the uncatalyzed process, however, because the hill is much smaller. The groups on the substrate that are involved in the weak interactions between the enzyme and transition state can be at some distance from the substrate bonds that are broken or changed. The weak interactions formed only in the transition state are those that make the primary contribution to catalysis (see **Figure 5-10b**; see also **Figure 5-9**).

Covalent interactions can accelerate some enzyme-catalyzed reactions by creating a different, lower-energy reaction pathway. When the reaction occurs in solution in the absence of the enzyme, the reaction takes a particular (and usually very slow) path. In the enzyme-catalyzed reaction, if a group on the enzyme is

transferred to or from the substrate during the reaction, the reaction path is altered. The new pathway results in reaction acceleration only if its overall activation energy is lower than that of the uncatalyzed reaction.

The Rates of Enzyme-Catalyzed Reactions Can Be Quantified

The oldest approach to understanding enzyme mechanisms, and the one that remains most important, is to determine the rate of a reaction and how it changes in response to changes in experimental parameters, a discipline known as **enzyme kinetics**. We provide here a brief review of key concepts related to the kinetics of enzyme-catalyzed reactions (for more advanced treatments, see Additional Reading at the end of the chapter.)

Substrate concentration affects the rate of enzyme-catalyzed reactions. Studying the effects of substrate concentration $[S]$ *in vitro* is complicated by the fact that it changes during the course of the reaction, as substrate is converted to product. One simplifying approach in kinetics experiments is to measure the **initial velocity**, designated V_0 (**Figure 5-11a**). In a typical reaction, the enzyme may be present in nanomolar quantities, whereas $[S]$ may be five or six orders of magnitude higher. If just the beginning of the reaction is monitored (often no more than the first few seconds), changes in $[S]$ can be limited to a small percentage, and $[S]$ can be regarded as constant. V_0 can then be explored as a function of $[S]$, which is adjusted by the investigator. The effect on V_0 of varying $[S]$ when the enzyme concentration is held constant is shown in **Figure 5-11b**. At relatively low concentrations of substrate, V_0 increases almost linearly with an increase in $[S]$. At higher substrate concentrations, V_0 increases by smaller and smaller amounts in response to increases in $[S]$. Finally, a point is reached beyond which V_0 increases are vanishingly small as $[S]$ increases. In this plateau-like V_0 region, the reaction approaches its **maximum velocity**, V_{max} .

When the enzyme is first mixed with a large excess of substrate, there is an initial **pre-steady state**, a period when the concentration of ES (enzyme-substrate) builds up. This period is usually too short to be observed easily, lasting just microseconds, and is not evident in **Figure 5-11a**. The reaction quickly achieves a **steady state** in which $[ES]$ (and the concentration of any other intermediates) remains approximately constant over time. The concept of a steady state was introduced by G. E. Briggs and J. B. S. Haldane in 1925. The measured V_0 generally reflects the steady state, even though V_0 is limited to the early part of the reaction,

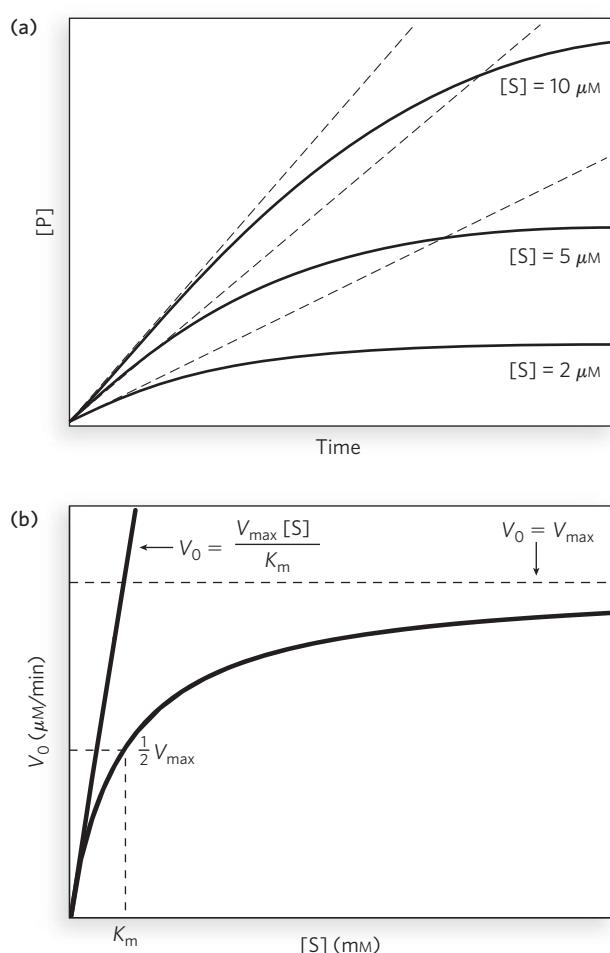
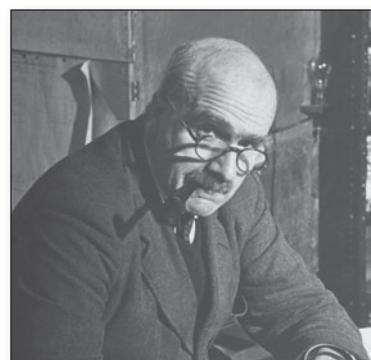


FIGURE 5-11 The initial velocity of an enzyme-catalyzed reaction. (a) A theoretical enzyme catalyzes the reaction $S \rightleftharpoons P$. Progress curves for the reaction (product concentration, $[P]$, vs. time) measured at three different initial substrate concentrations show that the rate of the reaction declines as substrate is converted to product. A tangent to each curve taken at time zero defines the initial velocity, V_0 , of the reaction. (b) The maximum velocity, V_{\max} , is indicated as a dashed line. V_0 approaches but never quite reaches V_{\max} . The substrate concentration at which V_0 is half maximal is K_m , the Michaelis constant. The concentration of enzyme in an experiment such as this is generally so low that $[S] \gg [E]$ even when $[S]$ is described as low or relatively low. At low $[S]$, the slope of the line is defined by $V_0 = V_{\max}[S]/K_m$, and V_0 exhibits a linear dependence on $[S]$. The units shown here are typical for enzyme-catalyzed reactions and help illustrate the meaning of V_0 and $[S]$. (Note that the curve describes part of a rectangular hyperbola, with one asymptote at V_{\max} . If the curve were continued below $[S] = 0$, it would approach a vertical asymptote at $[S] = -K_m$.)



J. B. S. Haldane, 1892–1964 [Source: Hans Wild/Time Life Pictures/Getty Images.]

and analysis of these initial rates is referred to as **steady-state kinetics**.

The curve expressing the relationship between $[S]$ and V_0 (see Figure 5-11b) has the same general shape for most enzymes (approaching a rectangular hyperbola). It can be expressed algebraically by an equation developed by Leonor Michaelis and Maud Menten, called the **Michaelis-Menten equation**:

$$V_0 = \frac{V_{\max} [S]}{K_m + [S]} \quad (5-7)$$

The important terms are $[S]$, V_0 , V_{\max} , and a constant called the **Michaelis constant**, K_m . All these terms are readily measured experimentally.

The Michaelis-Menten equation is the rate equation for a one-substrate enzyme-catalyzed reaction. It states the quantitative relationship between the initial velocity V_0 , the maximum velocity V_{\max} , and the initial substrate concentration $[S]$, all related through the Michaelis constant K_m . Note that K_m has units of concentration. Does the equation fit experimental observations? Yes; we can confirm this by considering the



Leonor Michaelis, 1875–1949 [Source: Rockefeller University Archive Center.]



Maud Menten, 1879–1960 [Source: Courtesy of Dorothy C. Craig.]

limiting situations where $[S]$ is very high or very low, as shown in Figure 5-11b. The K_m is functionally equivalent to the $[S]$ at which V_0 is one-half V_{max} . At low $[S]$, $K_m \gg [S]$ and the $[S]$ term in the denominator of the Michaelis-Menten equation (Equation 5-7) becomes insignificant. The equation simplifies to $V_0 = V_{max}[S]/K_m$, and V_0 exhibits a linear dependence on $[S]$. At high $[S]$, where $[S] \gg K_m$, the K_m term in the denominator of the Michaelis-Menten equation becomes insignificant and the equation simplifies to $V_0 = V_{max}$; this is consistent with the plateau observed at high $[S]$. The Michaelis-Menten equation is therefore consistent with the observed dependence of V_0 on $[S]$, and the shape of the curve is defined by the terms V_{max}/K_m at low $[S]$ and V_{max} at high $[S]$.

Kinetic parameters are used to compare enzyme activities. Many enzymes that follow Michaelis-Menten kinetics have different reaction mechanisms, and enzymes that catalyze reactions with six or eight identifiable intermediate steps within the enzyme active site often exhibit the same steady-state kinetic behavior. Even though Equation 5-7 holds true for many enzymes, both the magnitude and the real meaning of V_{max} and K_m can differ from one enzyme to the next. This is an important limitation of the steady-state approach to enzyme kinetics. The parameters V_{max} and K_m can be obtained experimentally for any given enzyme, but by themselves they provide little information about the number, rates, or chemical nature of discrete steps in the reaction. Nevertheless, steady-state kinetics is the standard language with which biochemists compare and characterize the catalytic efficiencies of enzymes.

Figure 5-11b shows a simple graphical method for obtaining an approximate value for K_m . More convenient procedures can be found in more advanced treatments of enzyme kinetics (see Additional Reading). The K_m , as noted, can vary greatly from enzyme to enzyme, and even for different substrates of the same enzyme. The term is sometimes used (often inappropriately) as an indicator of the affinity of an enzyme for its substrate.

The term V_{max} depends on both the concentration of enzyme and the rate of the rate-limiting step in the reaction pathway. Because the number of steps in a reaction and the identity of the rate-limiting step can vary, it is useful to define a **general rate constant**, k_{cat} , to describe the limiting rate of any enzyme-catalyzed reaction at saturation. If the reaction has several steps and one is clearly rate-limiting, k_{cat} is equivalent to the rate constant for that limiting step. When several steps are partially rate-limiting, k_{cat} can become a complex

function of several of the rate constants that define each individual reaction step. In the Michaelis-Menten equation, $V_{max} = k_{cat}[E_t]$, where $[E_t]$ is the total concentration of enzyme, and Equation 5-7 becomes:

$$V_0 = \frac{k_{cat} [E_t][S]}{K_m + [S]} \quad (5-8)$$

The constant k_{cat} is a first-order rate constant and hence has units of reciprocal time. It is equivalent to the number of substrate molecules converted to product in a given unit of time on a single enzyme molecule when the enzyme is saturated with substrate. Hence, this rate constant is also called the **turnover number**. A k_{cat} may be 0.01 s^{-1} for an enzyme with an intrinsically slow function (some enzymes with regulatory functions act very slowly) or as high as $10,000\text{ s}^{-1}$ for a fast enzyme catalyzing some aspect of intermediary metabolism.

In cells, there are many situations in which enzyme activity is inhibited by specific molecules, including other proteins. From a practical standpoint, the development of pharmaceutical and agricultural agents almost always involves the development of inhibitors for particular enzymes. Some aspects of enzyme inhibition are reviewed in Highlight 5-1.

DNA Ligase Activity Illustrates Some Principles of Catalysis

An understanding of the complete mechanism of action of a purified enzyme requires the identification of all substrates, cofactors, products, and regulators. Moreover, it requires a knowledge of (1) the temporal sequence in which enzyme-bound reaction intermediates form, (2) the structure of each intermediate and each transition state, (3) the rates of interconversion between intermediates, (4) the structural relationship of the enzyme to each intermediate, and (5) the energy contributed by all reacting and interacting groups to intermediate complexes and transition states. As yet, there is probably no enzyme for which we have an understanding that meets all these requirements.

It is impractical, of course, to cover all possible classes of enzyme chemistry, and we focus here on an enzyme reaction important to molecular biology: the reaction catalyzed by DNA ligases. The discussion concentrates on selected principles, along with some key experiments that have helped bring these principles into focus. We also use the DNA ligase example to review some of the conventions used to depict enzyme mechanisms. Many mechanistic details and pieces of experimental evidence are necessarily omitted; an

HIGHLIGHT 5-1 A CLOSER LOOK

Reversible and Irreversible Inhibition

Enzyme inhibitors are molecules that interfere with catalysis by slowing or halting enzyme reactions. The two general categories of enzyme inhibition are reversible and irreversible. There are three types of **reversible inhibition**: competitive, uncompetitive, and mixed.

A **competitive inhibitor** competes with the substrate for the active site of an enzyme (Figure 1a). While the inhibitor (I) occupies the active site, it prevents binding of the substrate to the enzyme. Many competitive inhibitors are structurally similar to the substrate and combine with the enzyme to form an EI complex, but without leading to catalysis. Even fleeting combinations of this type will reduce the efficiency of the enzyme. The two other types of reversible inhibition, though often defined in terms of one-substrate enzymes, are in practice observed only with enzymes having two or more substrates. An **uncompetitive inhibitor** binds at a site distinct from the substrate active site and, unlike a competitive inhibitor, binds only to the ES complex (Figure 1b). A **mixed inhibitor** also binds at a site distinct from the substrate active site, but it binds to either E or ES (Figure 1c).

All of these inhibition patterns can be analyzed with the aid of a single equation derived from the Michaelis-Menten equation:

$$V_0 = \frac{V_{\max} [S]}{\alpha K_m + \alpha' [S]}$$

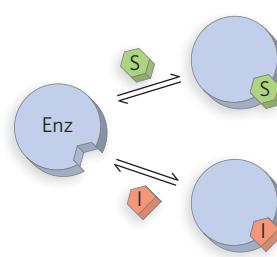
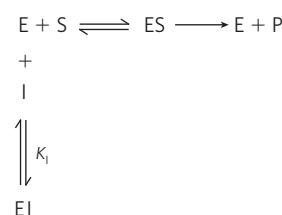
where α and α' reflect the interaction of an inhibitor with the free enzyme (through K_I) and the ES complex (through K'_I), respectively. These terms are defined as:

$$\alpha = 1 + \frac{[I]}{K_I}, \text{ and } K'_I = \frac{[E][I]}{[EI]}$$

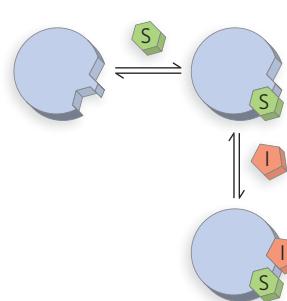
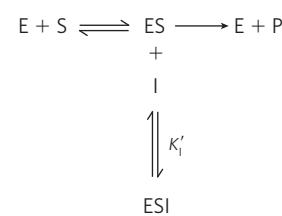
entire book would be needed to document the rich experimental history of enzyme research.

DNA ligases were discovered in 1967, and reports were published from four different research groups in that year. These enzymes catalyze the joining of DNA ends at strand breaks (also called nicks) where the 5' terminus is phosphorylated and the 3' terminus has a free hydroxyl group. In the intervening decades, many details of the reaction mechanism of these enzymes have been elucidated, and DNA ligases have become essential tools of biotechnology. RNA ligases have

(a) Competitive inhibition



(b) Uncompetitive inhibition



(c) Mixed inhibition

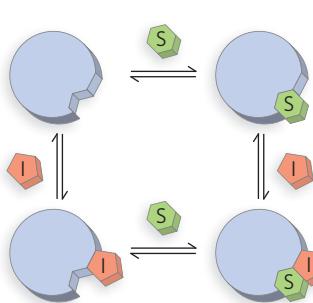
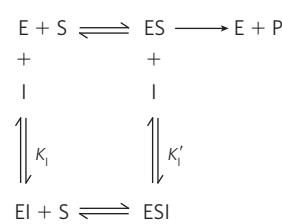


FIGURE 1 The three types of reversible inhibition.

- (a) Competitive inhibitors bind to the enzyme's active site.
- (b) Uncompetitive inhibitors bind at a separate site, but bind only to the ES complex.
- (c) Mixed inhibitors bind at a separate site, but may bind to either E or ES.

also been characterized; they use a similar reaction mechanism.

DNA ligases make use of two cofactors: Mg^{2+} ions and either ATP or nicotinamide adenine dinucleotide (NAD^+). ATP-dependent DNA ligases are found in eukaryotes, viruses, and some bacteria and archaea. NAD^+ -dependent ligases are found in most bacteria, as well as in some viruses and archaea. No eukaryotic ligases utilize NAD^+ . Commonly, NAD^+ is a cofactor participating in oxidation-reduction reactions. Its role in DNA ligase reactions is quite different, however,

Table 1 Effects of Reversible Inhibitors on Apparent V_{\max} and Apparent K_m

Inhibitor Type	Apparent V_{\max}	Apparent K_m
None	V_{\max}	K_m
Competitive	V_{\max}	αK_m
Uncompetitive	V_{\max}/α'	K_m/α'
Mixed	V_{\max}/α'	$\alpha K_m/\alpha'$

and

$$\alpha' = 1 + \frac{[I]}{K_I}, \text{ and } K_I^t = \frac{[E][I]}{[ESI]}$$

For a competitive inhibitor, there is no binding to the ES complex, and $\alpha' = 1$. For an uncompetitive inhibitor, there is no binding to the free enzyme (E), and $\alpha = 1$. For a mixed inhibitor, both α and α' are greater than 1. Each class of inhibitor has characteristic effects on the key kinetic parameters in the Michaelis-Menten equation, as summarized in Table 1. The altered K_m or V_{\max} measured in the presence of an inhibitor is often referred to as an *apparent K_m* or *V_{\max}* .

Many reversible enzyme inhibitors are used as pharmaceutical drugs; two examples are shown in Figure 2. The human immunodeficiency virus (HIV) encodes a DNA polymerase that can use either RNA or DNA as template; this enzyme is a reverse transcriptase (see Chapter 14). It uses deoxynucleoside triphosphates as substrates, and it is competitively inhibited by the drug AZT—the first drug to be employed in treating HIV infections. Similarly, quinolone antibiotics widely used to treat bacterial infections are uncompetitive inhibitors of enzymes called topoisomerases (described in Chapter 9).

An **irreversible inhibitor** can bind covalently with or destroy a functional group on an enzyme that

is essential for the enzyme's activity, or it can form a particularly stable noncovalent association. The formation of a covalent link between an irreversible inhibitor and an enzyme is common. Because the enzyme is effectively inactivated, irreversible inhibitors affect both V_{\max} and K_m . An inhibitor that does not form a covalent link but binds so tightly to the enzyme active site that it does not dissociate within hours or days is also effectively an irreversible inhibitor. Given that enzyme active sites bind most tightly to the transition state of the reactions they catalyze, a molecule that mimics the transition state can be a tight-binding inhibitor. Inhibitors designed in this way are called transition state analogs. Many drugs used to treat AIDS are designed in part as transition state analogs that bind tightly to the HIV protease.

Note that uncompetitive and mixed inhibitors should not be confused with allosteric modulators (see Section 5.4). Although the inhibitors bind at a second site on the enzyme, they do not necessarily mediate conformational changes between active and inactive forms, and the kinetic effects are distinct.

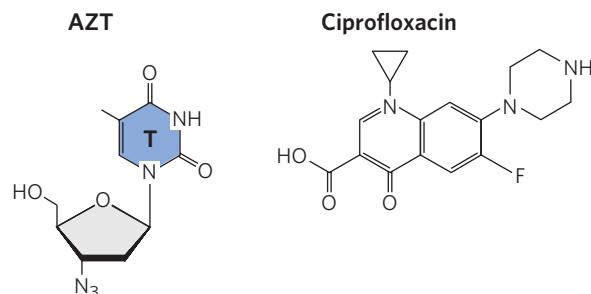


FIGURE 2 Examples of inhibitors with medical applications. AZT (3'-azido-3'-deoxythymidine) is used in the treatment of HIV/AIDS; ciprofloxacin is a quinolone antibiotic.

and it parallels the role of ATP in the ATP-dependent enzymes.

All DNA ligases promote a reaction that involves three chemical steps (Figure 5-12). In step 1, an adenylate group, adenosine 5'-monophosphate (AMP), is transferred from either ATP or NAD⁺ to a Lys residue in the enzyme active site. This process occurs readily in the absence of DNA. In step 2, the enzyme binds to DNA at the site of a strand break and transfers the AMP to the 5' phosphate of the DNA substrate. This activates the 5' phosphate for nucleophilic attack by the

3'-hydroxyl group of the DNA, in step 3, leading to displacement of the AMP and formation of a new phosphodiester bond in the DNA that seals the nick. Each of the three steps has a highly favorable reaction equilibrium that renders it effectively irreversible. In the absence of DNA, the adenylated enzyme formed in step 1 is quite stable, and it is likely that most DNA ligases in a cell are adenylated and ready to react with DNA. In addition to illustrating the reaction pathway, Figure 5-12 introduces the conventions commonly used to describe enzyme-catalyzed reactions.

How to Read Reaction Mechanisms—A Refresher

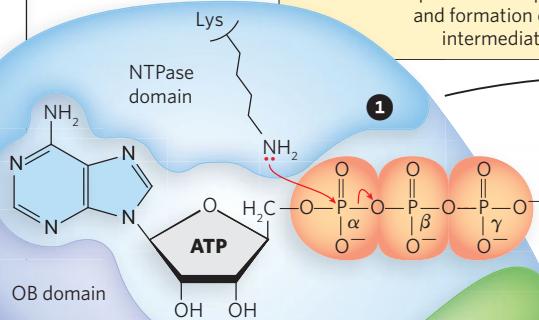
Chemical reaction mechanisms, which trace the formation and breakage of covalent bonds, are communicated with dots and curved arrows, a convention known informally as “electron pushing.” A covalent bond consists of a shared pair of electrons. Nonbonded electrons important to the reaction mechanism are designated by dots ($\text{---}\ddot{\text{O}}\text{H}$). Curved arrows (\curvearrowright) represent the movement of electron pairs. For movement of a single electron (as in a free radical reaction), a single-headed (fishhook-type) arrow is used (\curvearrowleft). Most reaction steps involve an unshared electron pair (as in the ligase mechanism).

Some atoms are more electronegative than others; that is, they more strongly attract electrons. The relative electronegativities of atoms encountered in this text are $\text{F} > \text{O} > \text{N} > \text{C} \approx \text{S} > \text{P} \approx \text{H}$. For example, the two electron pairs making up a C=O (carbonyl) bond are not shared equally; the carbon is relatively electron-deficient as the oxygen draws away the electrons. Many reactions involve an electron-rich atom (a nucleophile) reacting with an electron-deficient atom (an electrophile). Some common nucleophiles and electrophiles in biochemistry are shown at right.

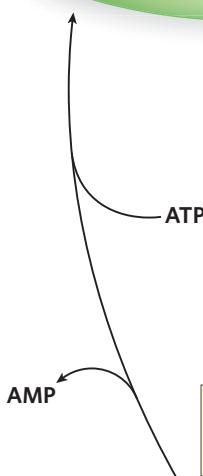
In general, a reaction mechanism is initiated at an unshared electron pair of a nucleophile. In mechanism diagrams, the base of the electron-pushing arrow originates near the electron-pair dots, and the head of the arrow points directly at the electrophilic center being attacked. Where the unshared electron pair confers a formal negative charge on the nucleophile, the negative charge symbol itself can represent the unshared electron pair, and serves as the base of the arrow. In some cases, the electron pair is the one that makes up a covalent bond, and the base of the arrow is then shown at the middle of the bond. In the ligase mechanism, the nucleophilic electron pair in the first chemical step is provided by the nitrogen of the ϵ -amino group of the Lys residue. This electron pair provides the base of the curved arrow. The electrophilic center under attack is the phosphorus atom of the α -phosphoryl group of ATP. The C, O, P, and N atoms have a maximum of 8 valence electrons, and H has a maximum of 2. These atoms are occasionally found in unstable states with less than their maximum allotment of electrons, but C, O, P, and N cannot have more than 8. Thus, when the electron pair from the ligase N attacks the substrate's phosphorus, an electron pair is displaced from the phosphorus valence shell. These electrons move toward the electronegative oxygen atoms. The oxygen shown as P=O has 8 valence electrons both before and after this chemical process, but the number shared with the phosphorus is reduced from 4 to 2, and the oxygen acquires a negative charge. To complete the process (not shown), the electron pair conferring the negative charge on the oxygen moves back to re-form a double bond with phosphorus and reestablish the P=O linkage. Again, an electron pair must be displaced from the phosphorus, and this time it is the electron pair shared with the oxygen that bridges the α and β phosphoryl groups so that pyrophosphate is released. The remaining steps follow a similar pattern.

DNA ligase

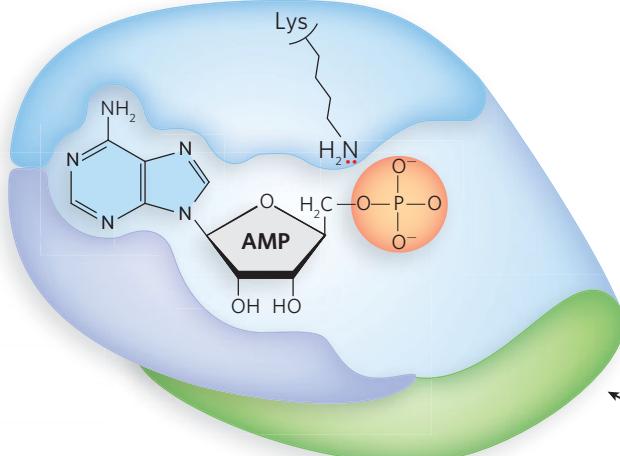
ATP (or NAD) is first bound to the enzyme. A Lys residue in the NTPase domain attacks the α -phosphate in the first chemical step, leading to displacement of pyrophosphate and formation of the E-AMP intermediate.

DNA ligase

DNA binding domain



The AMP product is released to regenerate free enzyme. It is quickly replaced by a new ATP molecule to re-start the catalytic cycle.



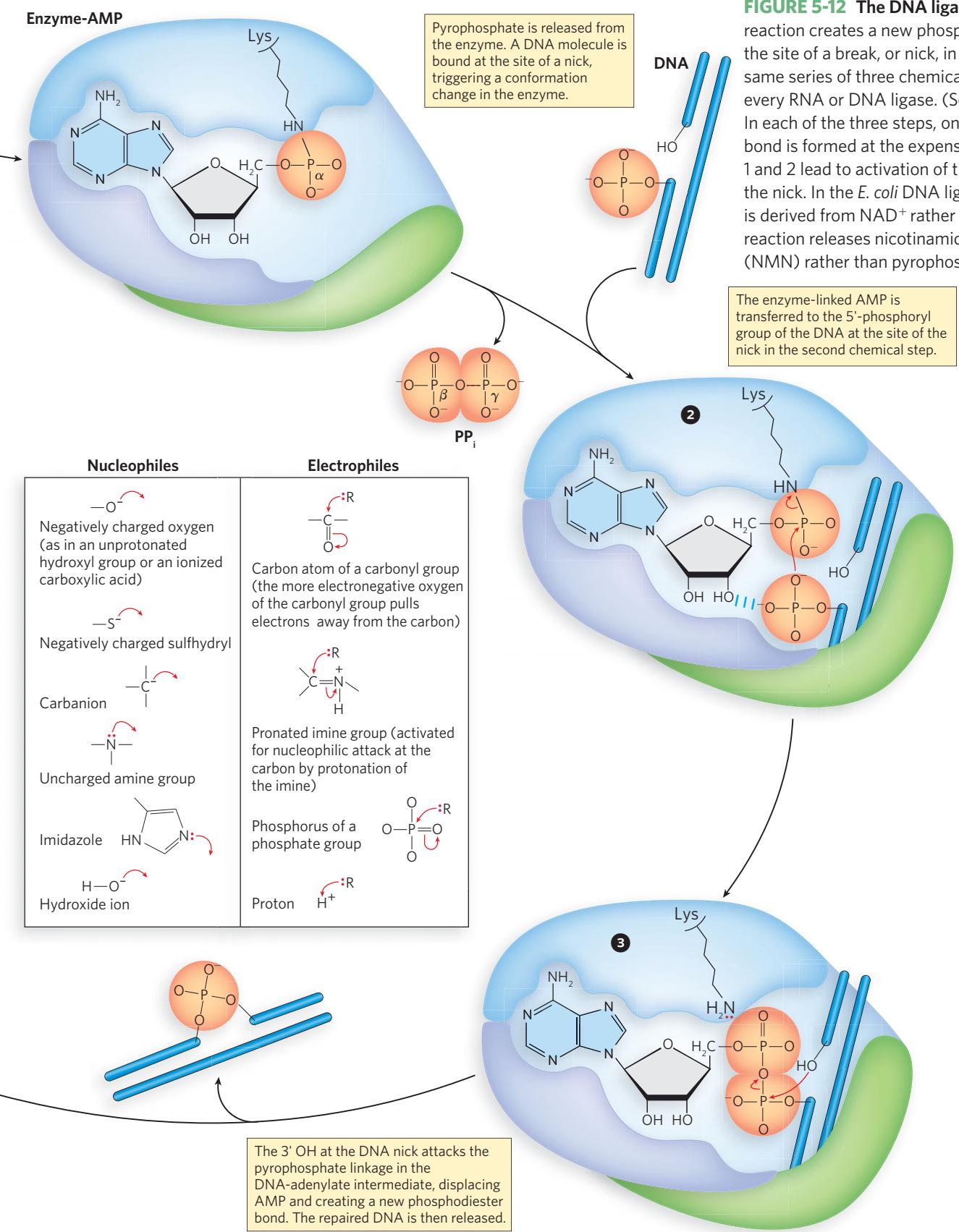


FIGURE 5-12 The DNA ligase reaction. The reaction creates a new phosphodiester bond at the site of a break, or nick, in the DNA. The same series of three chemical steps is used by every RNA or DNA ligase. (See text for details.) In each of the three steps, one phosphodiester bond is formed at the expense of another. Steps 1 and 2 lead to activation of the 5' phosphate in the nick. In the *E. coli* DNA ligase reaction, AMP is derived from NAD⁺ rather than ATP, and the reaction releases nicotinamide mononucleotide (NMN) rather than pyrophosphate.

The overall picture of the DNA ligase reaction mechanism is a composite derived from kinetic and structural studies of many closely related enzymes, including those isolated from bacteriophages T4 and T7, bacteria, eukaryotic viruses (viruses with eukaryotic host cells), and mammals. The ligases typically consist of a DNA-binding domain, a nucleotidyltransferase (NTase) domain, and an OB-fold domain. In DNA ligases, the OB fold, normally associated with binding to single-stranded DNA (see Section 5.1), interacts with the minor groove of double-stranded DNA. These domains provide a flexible structure that closes to completely encircle a nicked DNA molecule, with all three domains in contact with the DNA (see Figure 5-12). During step 1, some residues in the OB fold become part of the active site for transfer of AMP to the active-site Lys residue. As the covalent link between the enzyme and AMP is formed, the adenine base of AMP is fixed in a binding site in the NTase domain, where it stays throughout the remaining steps. In steps 2 and 3, a conformational change rearranges this part of the OB fold so that the same residues face the solvent, while other parts of the OB fold bind to the DNA and interact primarily with the strand adjacent to the 5'-phosphate end of the strand break. At the same time, the conformational change closes the enzyme around the nicked DNA, and the *N*-glycosyl bond between the adenine base and the ribose moiety of AMP is rotated, thereby realigning the phosphate of AMP for reaction, in step 2, with the 5' phosphate of DNA. Step 3 follows closely behind step 2 within the same complex.

This one example cannot provide a complete overview of the broad range of strategies that enzymes use, but it does serve as a good introduction to the complexity of enzyme reaction mechanisms. In addition, it illustrates the transfer of phosphoryl groups, a reaction catalyzed by protein enzymes and ribozymes involved in almost every aspect of molecular biology—from DNA polymerases and RNA polymerases to nucleases, topoisomerases, spliceosomes, and ligases.

SECTION 5.2 SUMMARY

- Most enzymes are proteins. They facilitate the reactions of substrate molecules. The catalyzed reaction occurs in an active site, a pocket on the enzyme that is lined with amino acid side chains and, in many cases, bound cofactors that participate in the reaction.
- Enzymes are catalysts. Catalysts do not affect reaction equilibria; they enhance reaction rates by lowering activation energies.

- Enzyme catalysis involves both covalent enzyme-substrate interactions and noncovalent interactions. Enzyme active sites bind most tightly to the transition states of the reactions they catalyze.
- The rates of most enzyme-catalyzed reactions are described by the Michaelis-Menten equation, which relates the key kinetic parameters V_{\max} , K_m , and k_{cat} .
- A DNA ligase catalyzes a series of phosphoryl transfer reactions to seal nicks in the DNA backbone; its reaction mechanism illustrates several general principles of enzyme-catalyzed reactions.

5.3 Motor Proteins

Organisms move. Cells move. Organelles and macromolecules within cells move. Most of these movements arise from the activity of **motor proteins**, a fascinating class of protein-based molecular motors. Fueled by chemical energy, usually derived from ATP, organized groups of motor proteins undergo cyclic conformational changes that create a unified, directional force—the tiny force that pulls apart chromosomes in a dividing cell and the immense force that levers a quarter-ton jungle cat into the air.

As in all proteins, interactions among different motor proteins, or between motor proteins and other types of proteins, include complementary arrangements of ionic bonds, hydrogen bonds, hydrophobic interactions, and van der Waals interactions at protein-binding sites. In motor proteins, however, these interactions achieve exceptionally high levels of spatial and temporal organization.

Motor proteins promote the contraction of muscles (actin and myosin), the migration of organelles along microtubules (kinesin and dynein), the rotation of bacterial flagella, and the movement of some proteins along DNA. In molecular biology, the motor proteins of most interest are those that function in the transactions involving nucleic acids. These include the helicases, involved in a wide range of processes; DNA and RNA polymerases (Chapters 11 and 15); DNA topoisomerases (Chapter 9); and other proteins that move along DNA as they carry out their functions in DNA metabolism (Chapters 11–14). Here, we focus on helicases, motor proteins that are highly relevant to the information pathways explored in this text.

Motor proteins combine the functions of ligand binding and catalysis. Each motor protein must interact transiently with another macromolecular ligand, binding and releasing it in a reversible process. To bring about productive motion, the sequence of binding and

release cannot be random; it must have a unidirectional component. This requires energy, usually derived from ATP hydrolysis, which is coupled to a directed bind-and-release process. For the motor proteins we are concerned with, the ligand is generally DNA or RNA.

Helicases Abound in DNA and RNA Metabolism

A **helicase** is a protein that separates the paired strands of a nucleic acid, converting a duplex into two single strands. Helicases are part of a larger family of motor proteins that promote reactions by translocating along the DNA (or RNA) substrate, resulting in the displacement of proteins from nucleic acids, the movement of branches

that occur in DNA during replication and recombination, conformational changes in nucleic acids, and the remodeling of chromatin. All of these processes are coupled to the hydrolysis of ATP. A great many human genetic diseases have been traced to defects in motor proteins of this class, attesting to their general importance.

Two structural classes of ATPase domain are found in helicases. The first is structurally related to the core domain of the bacterial RecA recombinase (see Chapter 13); ATP is often bound in a site near the intersection of two RecA-like domains (**Figure 5-13a**). The second is related to a class of enzymes called AAA⁺ (ATPases associated with various cellular activities); again, ATP is bound in sites located at the subunit-subunit interfaces (**Figure 5-13b**).

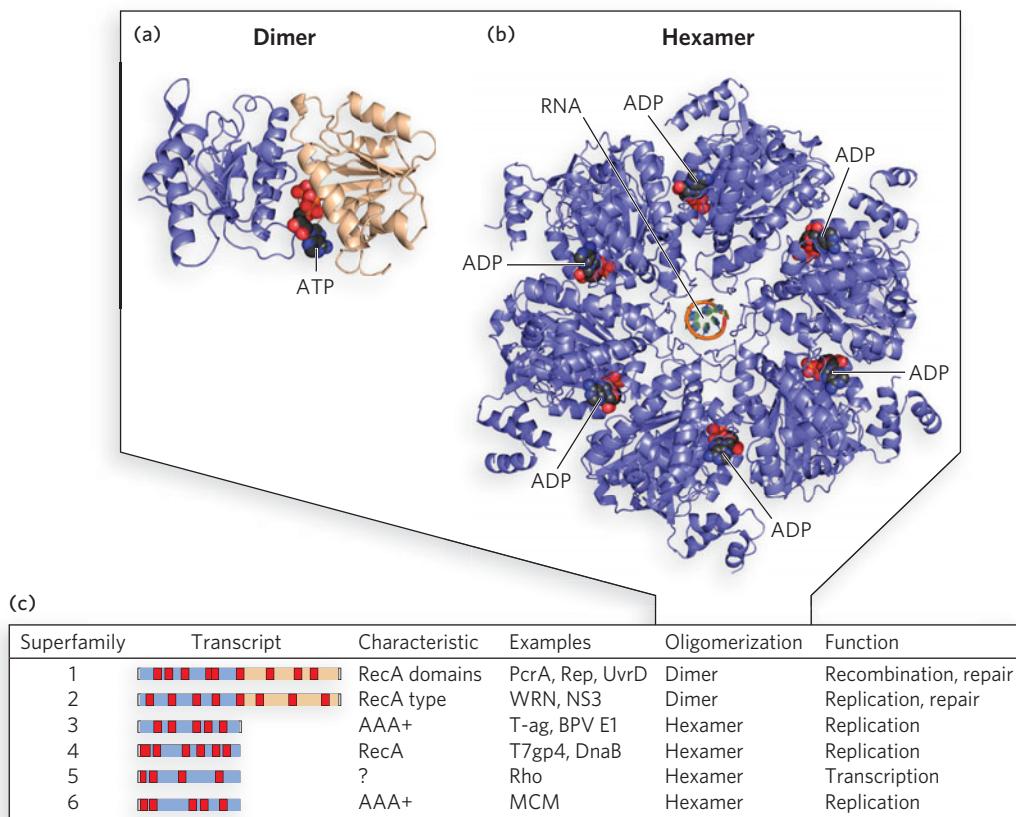


FIGURE 5-13 DNA helicases. (a) Enzymes of helicase superfamilies 1 and 2 (SF1 and SF2) have a conserved core structure, the RecA-like fold. Here, an ATP analog (black and red) that mimics ATP structure but is not hydrolyzed is bound at the interface of the two subunits in the helicase PcrA. (b) The core of helicases of SF3 through SF6 consists of six subunits with nucleotide-binding pockets at the subunit-subunit interfaces. Shown here is the Rho transcription termination factor (see Chapter 15). (c) Helicase superfamilies. Core domains and positions of the conserved motifs are shown for each family (“Transcript”). Red motifs are universal structural elements

in all helicases, including those involved in the binding and hydrolysis of a nucleoside triphosphate. In contrast to core domains, accessory domain positions and functions are specific to each protein. Their presence, function, and precise location within different members of the same superfamily vary widely. There are two different ATPase domain types in helicases, as shown in the “Characteristic” column. Well-studied examples of each family, typical oligomeric structure, and functions are listed in the remaining three columns. [Source: (a) PDB ID 3PJR (dimer). (b) PDB ID 3ICE (hexamer). (c) Adapted from M. R. Singleton et al., *Annu. Rev. Biochem.* 76:23–50, 2007.]

Based on extensive sequence comparisons, six superfamilies of helicases have been defined (Figure 5-13c). Most of them include proteins that do not function in DNA or RNA strand separation, but instead are involved in translocation along DNA or RNA. Helicases of superfamilies 1 and 2 (SF1 and SF2), the most common, usually have two RecA-like domains. These proteins often, but not always, function best as oligomers. Helicases of superfamilies 3 through 6 (SF3–6) generally function as circular hexamers (see Figure 5-13b). The individual domains are RecA-like in many cases, although the subunits in SF3 and SF6 helicases feature the AAA⁺ structural domain.

The superfamilies are further defined on the basis of a series of motifs (see Figure 5-13c). At least three motifs found in all helicase superfamilies are involved with ATP binding and hydrolysis. Other motifs are involved in DNA or RNA binding or oligomerization.

Helicase Mechanisms Have Characteristic Molecular Parameters

Any discussion of helicase mechanisms must take into account several key biochemical properties of these motor proteins: oligomeric state, rate of nucleic acid unwinding or translocation, directionality, processivity, step size, and ATP-coupling stoichiometry.

For some helicases, the **oligomeric state** in which they are active has been quite controversial. A few helicases exhibit some activity as monomers but are greatly stimulated by the addition of more subunits or auxiliary proteins. The observed rates of DNA unwinding or translocation can thus be a complicated function of reaction conditions and proteins added.

Helicases are associated with DNA **unwinding**, the separation of the two paired strands of a nucleic acid. Closely related proteins are often involved in **translocation**, which is movement along duplex DNA or RNA without separating the paired strands. These latter proteins are sometimes called **translocases**, but many other names are used for specific proteins that are more closely attuned to their function. In some cases, these proteins are bound to a structure such as a membrane or a viral coat and function to pump DNA or RNA through it. Others function to eject bound proteins from a nucleic acid.

Helicases move unidirectionally on a strand of nucleic acid, and **directionality**—the direction in which the enzyme moves along the strand—is an important distinguishing characteristic of a helicase mechanism. Some helicases move uniquely 3'→5' and some move 5'→3' (Figure 5-14). With double-stranded DNA, the direction of movement is defined by only one strand. The strand that guides the direction of movement is established during the process of loading the helicase

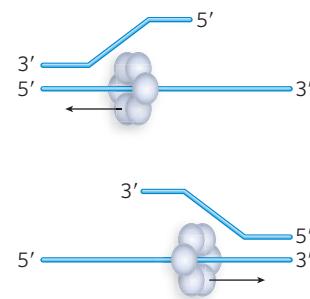


FIGURE 5-14 Helicases can be classified based on translocation directionality: 3'→5' (top) or 5'→3' (bottom).

onto the DNA. Helicases that translocate in the 3'→5' direction seem to be more common than those moving in the 5'→3' direction.

Processivity refers to the number of base pairs unwound or the number of nucleotides translocated, on average, each time an enzyme of this type binds to a DNA molecule. Some helicases may unwind only a few base pairs before dissociating, whereas a much greater processivity is the norm for hexameric helicases that encircle DNA and function in such processes as chromosomal DNA replication.

The **step size** is the average number of base pairs or nucleotides over which the helicase moves for each ATP molecule hydrolyzed. The **ATP-coupling stoichiometry** is the average number of ATPs consumed per base pair or nucleotide traversed. This may seem like another way of describing step size, and the two can be closely related. However, if coupling is not perfect and some ATP is hydrolyzed unproductively (e.g., in a futile cycle), the step size measured experimentally can be lower than the true step size that reflects the coupled ATP hydrolysis-movement cycle.

Most helicases seem to use a translocation mechanism that can be described as “stepping.” It requires the helicase to have at least two DNA-binding sites. Using a series of conformational changes facilitated by ATP binding and hydrolysis, the enzyme moves along the DNA. Different sites on the same or different helicase subunits bind alternately to the DNA strand defining directionality, in a closely orchestrated reaction sequence (Figure 5-15). When multiple subunits are employed (dimers or hexamers), DNA molecules may get passed between the subunits and larger step sizes can be observed.

Helicase translocation is necessary but not sufficient for the unwinding of a nucleic acid. Unwinding mechanisms are classified as active or passive. In an active mechanism, the enzyme interacts directly with double-stranded DNA or RNA to destabilize the duplex structure. In a passive mechanism, the enzyme interacts only with single strands, moving onto and stabilizing them as they form through normal thermal base-pair opening

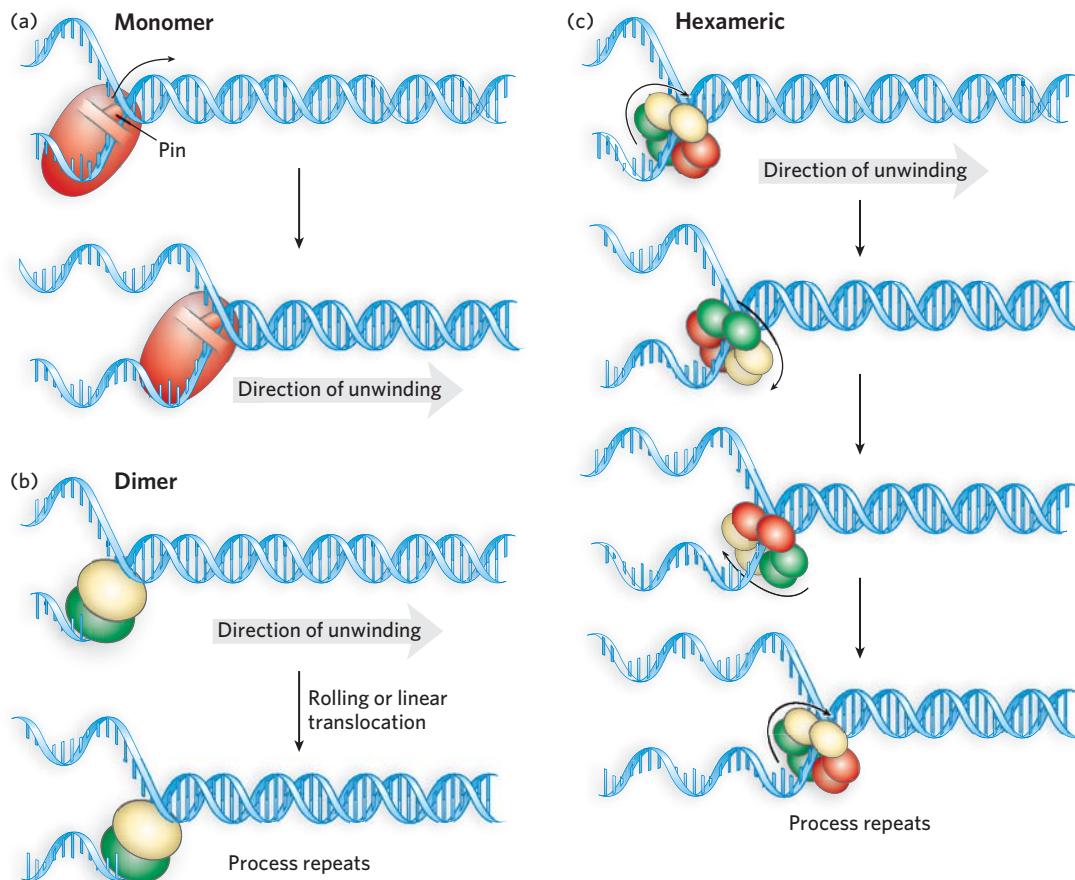


FIGURE 5-15 Unwinding reaction mechanisms for DNA helicases. (a) A monomeric helicase hydrolyzes ATP, causing conformational changes that couple unidirectional DNA translocation with duplex destabilization. Some of these enzymes have pinlike structures that function in strand separation. (b) A functional dimer model. A dimeric helicase has two identical subunits. These may alternate in binding to single- and double-stranded DNA, or translocate linearly along one strand of DNA (as shown here). In either case, the protein unwinds the DNA as it moves.

(c) Hexameric helicases move along one DNA strand of a duplex DNA, separating the strands as it moves. The bound DNA strand is usually bound alternately by three pairs of dimers within the hexameric structure. The passing of the strand from one dimer to another is coupled to conformational changes driven by ATP hydrolysis. In the diagram, the red, green, and yellow subunit pairs represent the three states; one state has bound ATP, the second has bound ADP (immediately after hydrolysis), and the third has no bound nucleotide.

and closing. Although active mechanisms are not yet well understood, they seem to be used by many helicases. In many cases, the destabilization of a double-stranded DNA or RNA is facilitated by a structure often described as a pin, a kind of wedge that prys the two strands apart. The ATPase motor pushes or pulls the paired strands past the pin (as seen in Figure 5-15a). Some dimeric helicases may employ an unwinding mechanism in which one subunit interacts with and destabilizes the duplex DNA, and the other subunit acts as a translocase on adjacent single-stranded DNA (Figure 5-15b).

A great variety of motor proteins carry out many other functions, besides nucleic acid unwinding. Many examples will arise in later chapters, and we consider just a few here. The RuvB protein (SF6 class; its name

derives from resistance to UV irradiation) is a DNA translocase involved in the movement of four-armed DNA junctions, called Holliday intermediates, that appear during DNA recombination (see Chapter 13). Acting with the RuvA protein, two hexamers of RuvB are arranged symmetrically to pump DNA, catalyzing a rapid branch migration (Figure 5-16a).

The eukaryotic Snf proteins (SF2 class; named for sucrose nonfermenting) are involved in the alteration and/or disruption of a variety of protein-DNA interactions. Each of the many Snf proteins has a particular target protein or protein complex, or set of targets, including RNA polymerase, nucleosomes, and regulatory DNA-binding proteins. Snf proteins interact directly with their target, utilizing their motor functions to

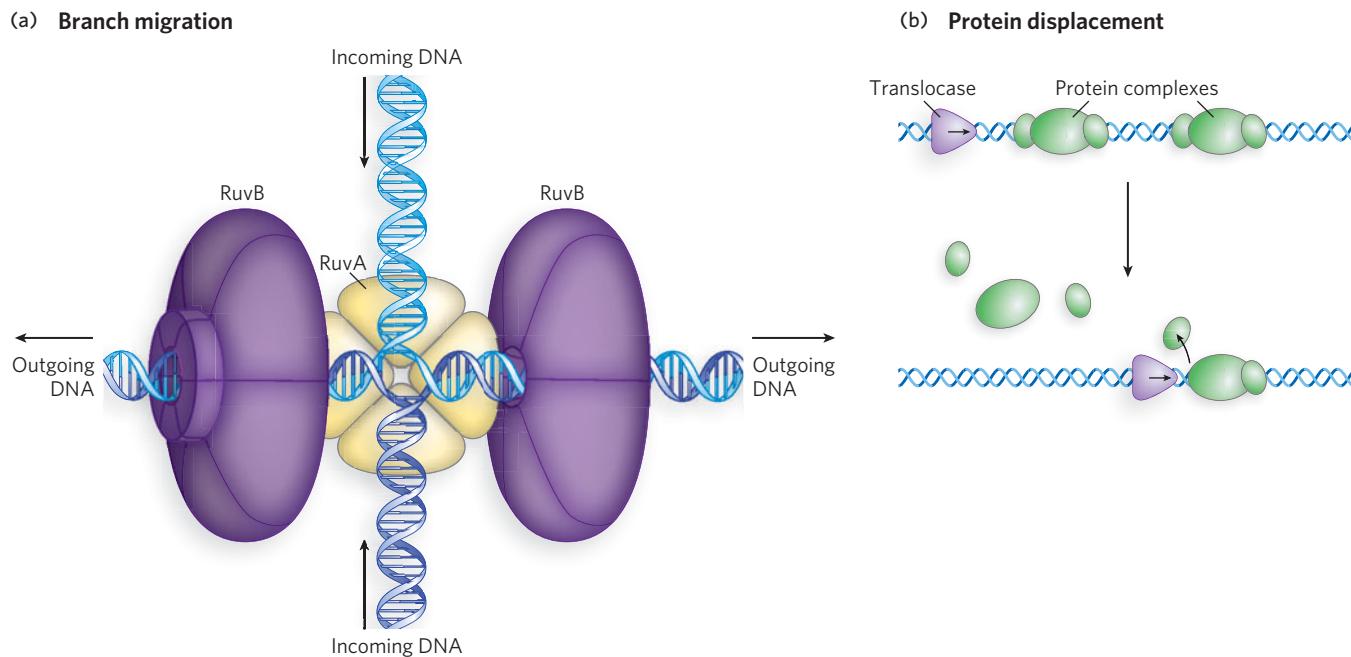


FIGURE 5-16 DNA translocases. (a) Branch migration during double-stranded DNA translocation. The bacterial RuvA and RuvB proteins form a complex that binds to a four-armed DNA junction that often forms during recombination (a Holliday intermediate; see Chapter 13), and the RuvA

protein binds to the junction itself. The RuvB protein is the translocase, propelling the DNA to the left and right, as shown (arrows). (b) Protein displacement during DNA translocation. Translocase proteins in both bacteria and eukaryotes displace proteins from nucleic acids.

displace or reposition the target proteins on the DNA (Figure 5-16b). Snf proteins are also heavily involved in chromatin remodeling (see Chapter 10). The NPH-II RNA helicase (SF2 class; NTP (nucleoside triphosphate) phosphohydrolase) is a viral helicase that unwinds RNA, but it is also very effective as a translocase that displaces proteins from single-stranded RNA. SF2 enzymes are particularly common and important in RNA metabolism. Many are critical to the maturation of large ribonucleoprotein complexes, such as ribosomes and spliceosomes.

With respect to ATP hydrolysis, motor proteins generally exhibit the standard steady-state kinetic behavior of an enzyme. An example is the Michaelis-Menten kinetics seen for ATP hydrolysis by the helicase PcrA in the presence of single-stranded DNA (Figure 5-17). PcrA is an SF1 class helicase found in gram-positive bacteria.

- Helicases are ubiquitous motor proteins involved in DNA or RNA strand separation. Helicase mechanisms are discussed in terms of oligomeric state, rate of nucleic acid unwinding or translocation, directionality, processivity, step size, and ATP-coupling stoichiometry.

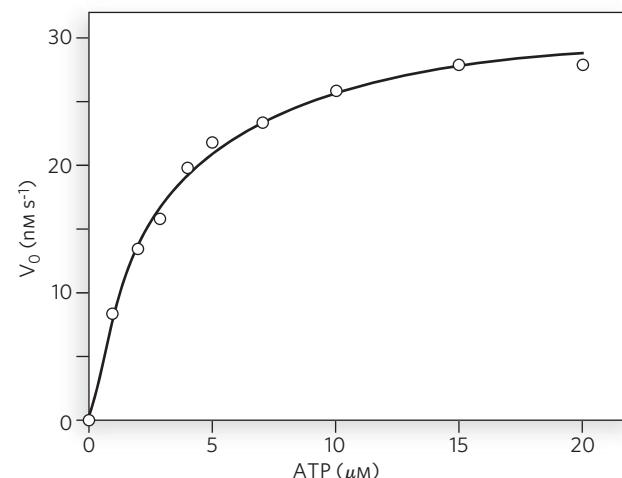


FIGURE 5-17 A plot of V_0 versus $[S]$ for the hydrolysis of ATP by the helicase PcrA. This hydrolytic reaction follows classic Michaelis-Menten kinetics. [Source: Adapted from: C. P. Toseland et al., *J. Mol. Biol.* 392: 1020–1032, 2009.]

SECTION 5.3 SUMMARY

- Motor proteins combine the functions of ligand binding and catalysis. The ligand, often RNA or DNA, is bound reversibly. Movement involves protein conformational changes triggered by ATP hydrolysis catalyzed by the motor protein.

- Many motor proteins closely related to helicases are involved in translocation along nucleic acids, movement of branches in double-stranded DNA during replication, protein displacement from nucleic acids, chromatin remodeling, and other functions.

5.4 The Regulation of Protein Function

In the flow of biological information, groups of enzymes often work together in sequential and interconnected pathways to carry out a given process, such as the replication of a chromosome or the splicing of an intron in a messenger RNA. Such processes use enormous amounts of chemical energy in the form of nucleoside triphosphates (NTPs). It is critical not only that these processes occur, but that they do so only at a specific time and in a specific place, so that resources are not wasted. In addition, the many reactions that carry out these complex processes must be precisely coordinated. Faulty coordination or poor timing could damage or alter the cellular genome. Regulation is thus an important aspect of virtually every process in molecular biology.

Most enzymes follow the Michaelis-Menten kinetic patterns described in Section 5.2. However, the timing and rate of many processes are governed by **regulatory enzymes** that exhibit increased or decreased catalytic activity in response to certain signals. The resulting regulation conserves cellular resources and prevents inappropriate alterations in the genetic material.

The activities of regulatory enzymes or binding proteins are modulated in several ways: by the noncovalent binding of allosteric modulators, autoinhibition by a seg-

ment of the protein, reversible covalent modification, or proteolytic cleavage. We touch on all of these mechanisms, while focusing on those that are most common in enzymes and proteins involved in nucleic acid metabolism.

Modulator Binding Causes Conformational Changes in Allosteric Enzymes

Allosteric enzymes or **allosteric proteins** are those having “other shapes” or other conformations induced by the binding of modulators. This is found in certain regulatory enzymes, as conformational changes induced by one or more **allosteric modulators** interconvert more-active and less-active forms of the enzyme. The modulators for allosteric enzymes may be inhibitory or stimulatory. Often the modulator is the substrate itself, and a regulatory protein or enzyme for which substrate and modulator are identical is referred to as **homotropic**. When the modulator is a molecule other than the normal ligand or substrate, the enzyme or protein is said to be **heterotropic**.

The properties of allosteric enzymes are significantly different from those of simple, nonregulatory enzymes. Some of the differences are structural. In addition to active sites, allosteric enzymes generally have one or more regulatory, or allosteric, sites for binding the modulator (Figure 5-18). Just as an enzyme’s active site is specific for its substrate, each regulatory site is specific for its modulator. Enzymes with several modulators generally have different specific binding sites for each.

Allosteric Enzymes Have Distinctive Binding and/or Kinetic Properties

Allosteric enzymes show relationships between V_0 and [S] that differ from Michaelis-Menten kinetics. They do

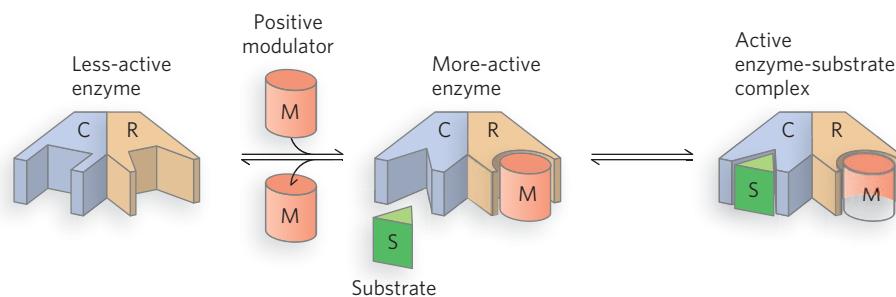


FIGURE 5-18 Allosteric enzyme interactions. In many allosteric enzymes, the substrate-binding site and the modulator-binding site(s) are on different subunits: catalytic (C) and regulatory (R) subunits, respectively. Binding of a positive (stimulatory) modulator (M) to its specific site on

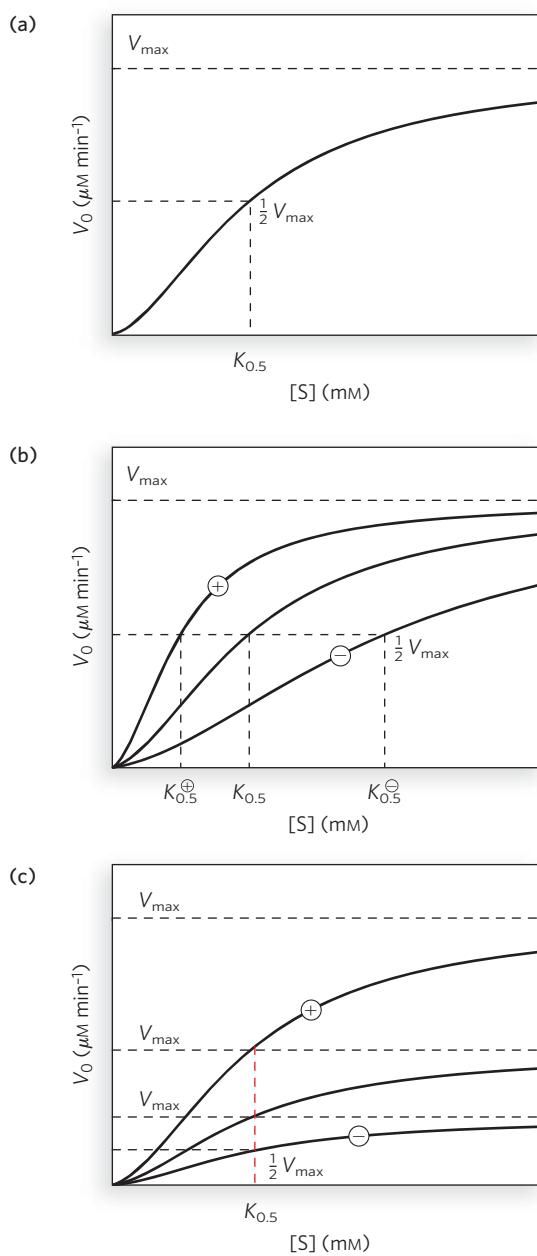
the R subunit is communicated to the C subunit through a conformational change, which renders the C subunit active and capable of binding the substrate (S) with higher affinity. On dissociation of the modulator from the regulatory subunit, the enzyme reverts to its inactive or less-active form.

FIGURE 5-19 Substrate-activity curves for allosteric enzymes. (a) The sigmoid curve of a homotropic enzyme, in which the substrate also serves as a positive (stimulatory) modulator, or activator. (b) The effects of a positive modulator (+) and a negative modulator (−) on an allosteric enzyme in which $K_{0.5}$ is altered without a change in V_{max} . The central curve shows the substrate-activity relationship without a modulator. (c) A less common type of modulation, in which V_{max} is altered and $K_{0.5}$ is nearly constant.

exhibit saturation with the substrate, but for some allosteric enzymes, plots of V_0 versus $[S]$ produce a sigmoid saturation curve, rather than the hyperbolic curve typical of nonregulatory enzymes (Figure 5-19a). On the sigmoid saturation curve we can find a value of $[S]$ at which V_0 is half-maximal, but we cannot refer to it with the designation K_m , because the enzyme does not follow the hyperbolic Michaelis-Menten relationship. Instead, the symbol $[S]_{0.5}$ or $K_{0.5}$ is often used to represent the substrate concentration giving half-maximal velocity of the reaction catalyzed by an allosteric enzyme.

For homotropic allosteric proteins or enzymes, the substrate often acts as a positive modulator (an activator), because the subunits act cooperatively. Cooperativity occurs when the binding of a substrate to one binding site alters an enzyme's conformation and affects the binding of subsequent substrate molecules. Most commonly, the binding of one molecule enhances the binding of others, an effect called positive cooperativity. This accounts for the sigmoid rather than hyperbolic change in V_0 with increasing $[S]$. One characteristic of sigmoid kinetics is that small changes in the concentration of a modulator can be associated with large changes in enzyme activity. A relatively small increase in $[S]$ in the steep part of the curve causes a comparatively large increase in V_0 (see Figure 5-19a). Much rarer are cases of negative cooperativity, where the binding of one substrate molecule impedes the binding of subsequent molecules.

For heterotropic allosteric proteins or enzymes, those whose modulators are molecules other than the normal substrate, it is difficult to generalize about the shape of the binding or substrate-saturation curve. An activator may cause the curve to become more nearly hyperbolic, with a decrease in $K_{0.5}$ but no change in V_{max} , resulting in an increased reaction velocity at a fixed substrate concentration (V_0 is higher for any value of $[S]$, as shown in Figure 5-19b, top curve). A negative modulator (an inhibitor) may produce a more sigmoid substrate-saturation curve, with an increase in $K_{0.5}$ (Figure 5-19b, bottom curve). Other heterotropic allosteric



enzymes respond to an activator by an increase in V_{max} with little change in $K_{0.5}$ (Figure 5-19c, top curve) and to an inhibitor by a decrease in V_{max} with little change in $K_{0.5}$ (Figure 5-19c, bottom curve). Heterotropic allosteric proteins and enzymes therefore show different kinds of response in their substrate-activity curves, because some have inhibitory modulators, some have activating modulators, and some have both.

Many of the allosteric effects encountered by molecular biologists are heterotropic. Numerous examples can be found among the regulatory proteins that bind to specific DNA sequences adjacent to genes, such as the **cAMP receptor protein**, or **CRP** (also called CAP, for catabolite gene activation protein; see Figure 4-18a). CRP is a dimeric

DNA-binding protein that participates in the regulation of genes involved in bacterial carbohydrate metabolism. Each subunit has separate domains that bind a modulator, cAMP (cyclic AMP), and a specific site in the DNA. The binding of cAMP exhibits negative cooperativity, in which the binding of cAMP to the modulator-binding site of one subunit reduces the affinity of the cAMP-binding site of the other subunit by two orders of magnitude. Bound cAMP also promotes conformational changes in the DNA-binding domain that facilitate the binding of CRP to its DNA binding site. This is just one example of the variety of allosteric effects observed in gene repressor and activator proteins that help modulate the sensitivity of these proteins to their environment.

Enzyme Activity Can Be Affected by Autoinhibition

Many processes of DNA and RNA metabolism are precisely targeted; they are limited to particular locations and circumstances, some of which appear transiently and unpredictably. For example, if DNA is damaged, the lesion must be repaired before the next replication cycle. The enzymes that repair DNA cleave the DNA strands near the lesion, remove damaged nucleotides, and replace them. It is essential that these enzymes be available on short notice and that they act only at DNA lesions.

One way to make such enzymes generally available but not generally active is by **autoinhibition**. A segment of the protein, sometimes an entire domain, can reduce or eliminate the activity of the enzyme. Such a protein may be present in the cell, but it functions weakly or not at all under normal circumstances. Activation of the enzyme requires self-assembly into a more functional oligomer, the binding of an auxiliary protein, or interaction with a particular ligand. In all cases, the interaction results in a conformational change that repositions the autoinhibitory protein segment. The activity of the enzyme then increases.

Autoinhibition has been documented for several proteins, including certain helicases and the bacterial RecA recombinase. In bacteria, the activity of a helicase known as Rep, involved in DNA replication, is autoinhibited by a subdomain called 2B. The bacterial RecA recombinase is autoinhibited by a short segment of polypeptide at its C-terminus (Figure 5-20). For most RecA proteins, this segment includes a high concentration of negatively charged amino acid residues, and it may interact with other parts of the RecA protein through electrostatic interactions.

Autoinhibition may be just one aspect of a broader regulatory strategy. The autoinhibited enzyme can be maintained in the cell without its activity causing unnecessary problems when it is not needed. Activation

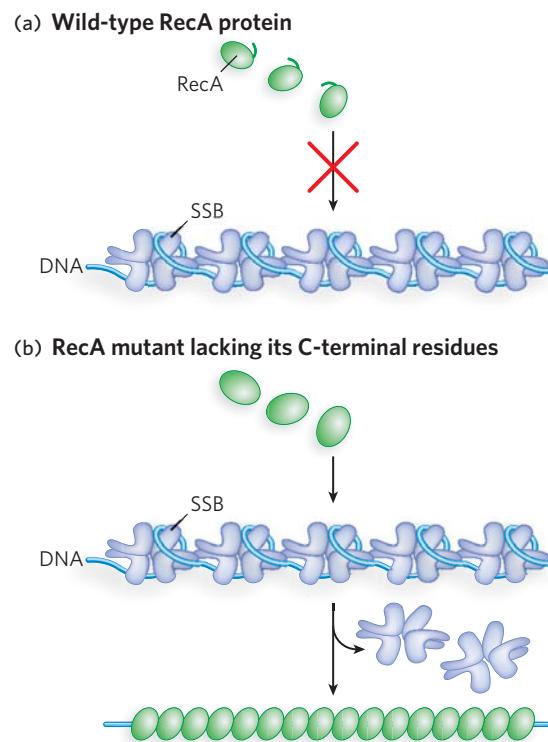


FIGURE 5-20 Autoinhibition. (a) RecA filament formation on DNA is characterized by distinct nucleation and filament extension phases. In the bacterial RecA protein, a C-terminal segment prevents efficient nucleation of binding to single-stranded DNA when single-stranded DNA-binding protein (SSB) is bound to the DNA strand. (b) If the C-terminal segment is removed to create a truncated, mutant RecA protein, nucleation and subsequent displacement of SSB to form a filament on the DNA are rapid.

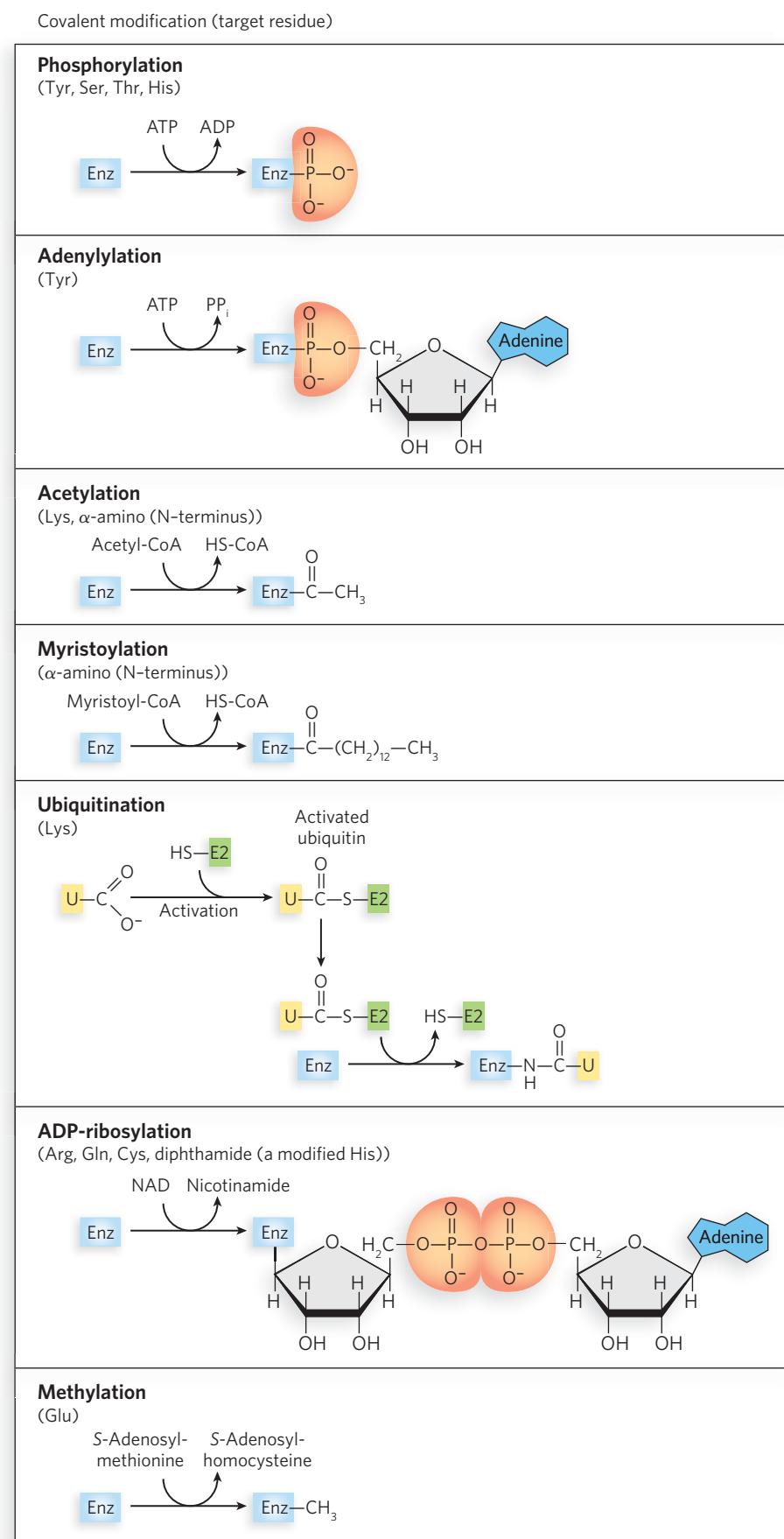
by interaction with additional proteins can occur quickly when its activity is required, and the enzyme's function can be more readily targeted to specified locations and situations.

Some Proteins Are Regulated by Reversible Covalent Modification

In another important class of regulatory mechanism, activity is modulated by **covalent modification** of one or more amino acid residues in the enzyme molecule. More than 500 different types of covalent modification have been found in proteins. Common modifying groups include phosphoryl, acetyl, adenylyl, uridylyl, methyl, amide, carboxyl, myristoyl, palmitoyl, prenyl, hydroxyl, sulfate, and adenosine diphosphate ribosyl groups. There are even entire proteins that are used as specialized modifying groups, such as ubiquitin. Some of these modifications are shown in Figure 5-21. These varied groups are generally linked to and removed from a regulated enzyme by separate enzymes. When an amino

FIGURE 5-21 Some enzyme modification reactions.

modification reactions. In the ubiquitination pathway, E2 is a carrier protein for activated ubiquitin. See text for details.



acid residue is modified, a novel amino acid with altered properties is effectively introduced. The introduction of a charge can alter the enzyme's local properties and induce a conformational change. The introduction of a hydrophobic group can trigger association with a membrane. The changes are often substantial and can be critical to the function of the altered enzyme.

The variety of protein modifications is too great to cover in detail, but we'll present a few examples. In eukaryotic cells, histones are important modification targets. As described in detail in Chapter 10, many histones and histone variants are subject to precise patterns of modification involving methylation, acetylation, phosphorylation, and ubiquitination. Such modifications play an important role in altering chromatin structure in specific regions, to facilitate gene expression and other activities.

Phosphorylation is probably the most common type of regulatory modification. It is estimated that one-third of all proteins in a eukaryotic cell are phosphorylated, and one or (often) many phosphorylation events are part of virtually every regulatory process. Some of these proteins have only one phosphorylated residue, others have several, and a few have dozens of sites for phosphorylation. This mode of covalent modification is central to a large number of regulated processes, so we'll discuss it in some detail.

Phosphoryl Groups Affect the Structure and Catalytic Activity of Proteins

The attachment of phosphoryl groups to specific amino acid residues of a protein is catalyzed by **protein kinases**; the removal of the groups is catalyzed by **protein phosphatases**. The addition of a phosphoryl group to a Ser, Thr, or Tyr residue introduces a bulky, charged group into a region that was only moderately polar. The oxygen atoms of a phosphoryl group can hydrogen-bond with one or several groups in a protein, commonly the amide groups of the peptide backbone at the start of an α helix or the charged guanidinium group of an Arg residue. The two negative charges on a phosphorylated side chain can also repel neighboring negatively charged (Asp or Glu) residues. When the modified side chain is located in a region of an enzyme critical to its three-dimensional structure, phosphorylation can have dramatic effects on enzyme conformation and thus on substrate binding and catalysis.

The Ser, Thr, or Tyr residues that are phosphorylated in regulated proteins occur within common structural motifs, called **consensus sequences**, that are recognized by specific protein kinases (Table 5-5). Some kinases are basophilic, preferentially phosphorylating a residue that has basic neighbors; others have different

Table 5-5 Consensus Sequences for Protein Kinases

Protein Kinase	Consensus Sequence and Phosphorylated Residue(s)*
Protein kinase A	-x-R-[RK]-x-[ST]-B-
Protein kinase G	-x-R-[RK]-x-[ST]-x-
Protein kinase C	-[RK](2)-x-[ST]-B-[RK](2)-
Protein kinase B	-x-R-x-[ST]-x-K-
Ca^{2+} /calmodulin kinase I	-B-x-R-x(2)-[ST]-x(3)-B-
Ca^{2+} /calmodulin kinase II	-B-x-[RK]-x(2)-[ST]-x(2)-
Myosin light chain kinase (smooth muscle)	-K(2)-R-x(2)-S-x-B(2)-
Phosphorylase b kinase	-K-R-K-Q-I-S-V-R-
Extracellular signal-regulated kinase (ERK)	-P-x-[ST]-P(2)-
Cyclin-dependent protein kinase (cdc2)	-x-[ST]-P-x-[KR]-
Casein kinase I	-[SpTp]-x(2,3)-[ST]-B-
Casein kinase II	-x-[ST]-x(2)-[EDSpYp]-x-
β -Adrenergic receptor kinase	-[DE](n)-[ST]-x(3)-
Rhodopsin kinase	-x(2)-[ST]--(E)(n)-
Insulin receptor kinase	-x-E(3)-Y-M(4)-K(2)-S-R-G-D-Y-M-T-M-Q-I-G-K(3)-L-P-A-T-G-D-Y- M-N-M-S-P-V-G-D-
Epidermal growth factor (EGF) receptor kinase	-E(4)-Y-F-E-L-V-

Sources: L. A. Pinna and M. H. Ruzzene, *Biochim. Biophys. Acta* 1314:191-225, 1996; B. E. Kemp and R. B. Pearson, *Trends Biochem. Sci.* 15:342-346, 1990; P. J. Kennelly and E. G. Krebs, *J. Biol. Chem.* 266:5,555-15,558, 1991.

*Shown here are deduced consensus sequences and (in italic) actual sequences from known substrates. The Ser (S), Thr (T), or Tyr (Y) residue that undergoes phosphorylation is in bold; all amino acid residues are shown as their one-letter abbreviations; x represents any amino acid; B, any hydrophobic amino acid; Sp, Tp, and Yp, already phosphorylated Ser, Thr, and Tyr residues. A pair of residues in square brackets (e.g., [ST]) indicates that one can substitute for the other. Numbers in parentheses indicate number of repeats—for example, x(2) means x-x; x(2,3) means x-x or x-x-x.

HIGHLIGHT 5-2 MEDICINE

HIV Protease: Rational Drug Design Using Protein Structure

Human immunodeficiency virus (HIV), the causative agent of AIDS, kills cells of the immune system. The development of a vaccine has been unsuccessful, because the surface glycoproteins targeted by antibodies change rapidly, in part due to the extremely high mutation rate of HIV (about one replication mistake in every 10,000 nucleotides of HIV genome per generation). However, there has been substantial success in the development of drugs targeting HIV-encoded enzymes that are essential for viral propagation. HIV is a retrovirus, an RNA virus that converts, or reverse transcribes, its RNA genome into DNA. Before HIV was discovered, various laboratories had already studied the life cycle of retroviruses, many of which cause cancer in humans and other animals. With the arrival of HIV, researchers had a head start—a vast amount of established research that had already identified key enzymes required for retroviral propagation. Among these enzymes is a protease that digests long, precursor polypeptides into smaller, active viral proteins.

The usual route of developing a drug that inhibits enzyme activity starts with the random screening of hundreds of thousands of chemical compounds. Possible inhibitors are then chemically optimized for potency, availability in an oral form, and low toxicity. The U.S. Food and Drug Administration's approval for the use of a drug in humans usually takes well over a dozen years. This type of process has led to drugs that inhibit certain HIV enzymes, including the reverse transcriptase. Rational drug design is another, shorter route to drug discovery. It starts with the target protein structure and designs chemicals that plug the active site and shut the enzyme down.

This process short-circuits the random, brute-force approach to searching for chemical inhibitors and has the potential to cut years off the drug-discovery process.

An astounding success in rational drug design has been achieved with the HIV protease. This is partly due to the unique architecture of the protein. HIV protease is a dimer of identical subunits, but unlike typical dimers, which contain two active sites, the HIV protease dimer has a single active site located in the central hydrophobic chamber at the dimer interface. The active site has twofold symmetry; each subunit contributes a catalytic Asp residue, and the two cooperate to hydrolyze the peptide bond (Figure 1). Information obtained from the biochemical studies and crystal structures of HIV protease made possible the rational design of chemical inhibitors. By 1996, the FDA had approved three HIV protease inhibitor drugs: indinavir (Crixivan), ritonavir (Norvir), and saquinavir (Invirase). Notably, all of these drugs are effective, in part, because they mimic the transition state of the proteolytic reaction catalyzed by the enzyme and thus bind to the enzyme virtually irreversibly (Figure 2).

Protease inhibitor drugs have helped reduce the viral titer in the plasma of HIV-infected individuals. However, because of the high mutation rate of the virus, these drugs are effective only for a limited time. Viral mutability can be countered by using different drugs at different times or in a combination cocktail with drugs that inhibit other HIV enzymes. An intensive study of the structure of HIV protease mutants that can circumvent drug action is underway to identify regions of the protein that cannot endure mutation, with the hope of designing inhibitors to which the virus cannot develop resistance.

substrate preferences, such as for a residue near a proline. Besides local amino acid sequence, the overall three-dimensional structure of a protein can determine whether a protein kinase has access to a given residue and can recognize it as a substrate. Another factor influencing the substrate specificity of certain protein kinases is the proximity of other phosphorylated residues.

To serve as an effective regulatory mechanism, phosphorylation must be reversible. Cells contain a family of phosphoprotein phosphatases that hydrolyze specific P_i -Ser, P_i -Thr, and P_i -Tyr esters (P_i is shorthand for the phosphoryl group), releasing inorganic phosphate (P_i). The phosphoprotein phosphatases we know of thus far act only on a subset of phosphoproteins, but they show less substrate specificity than protein kinases.

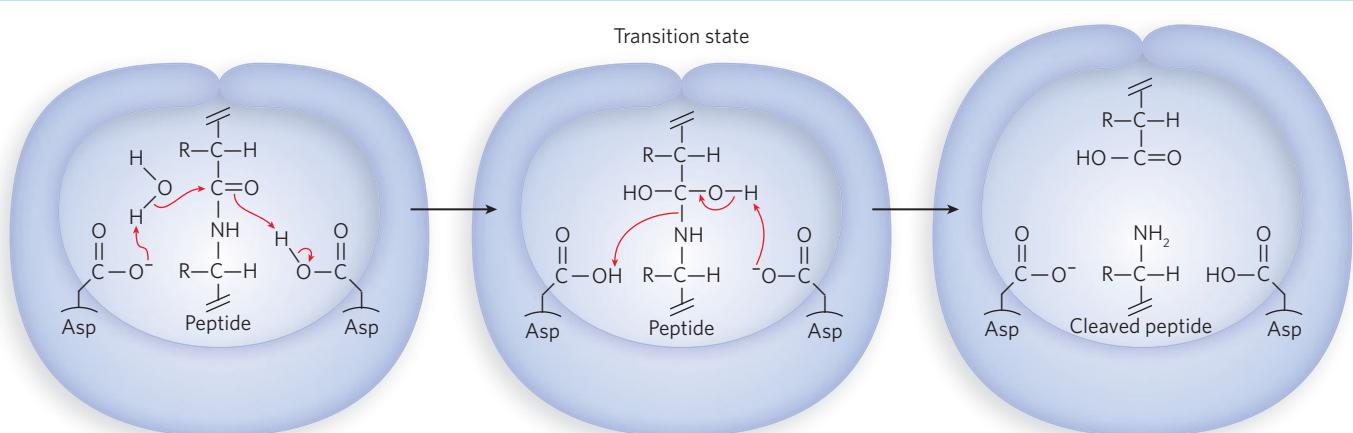


FIGURE 1 In the chemical mechanism of peptide hydrolysis by the HIV protease, one active-site Asp residue is contributed by each of the identical subunits.

One Asp activates a water molecule, while the other stabilizes the leaving group.

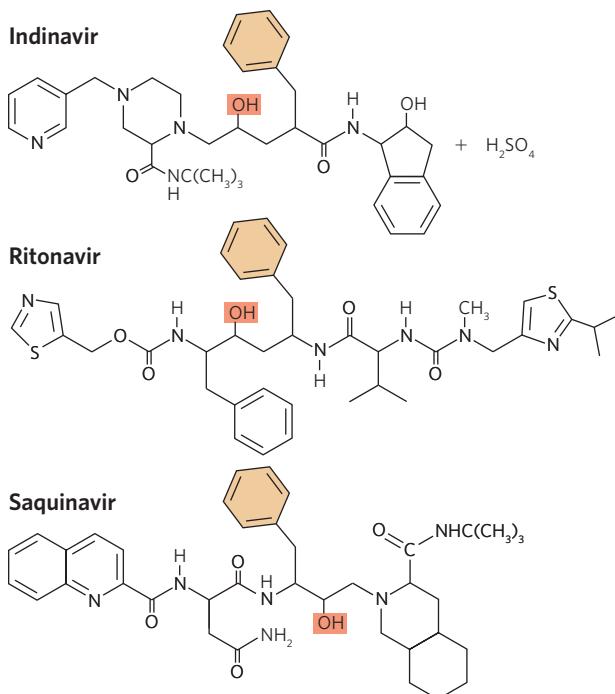


FIGURE 2 Three HIV protease inhibitors used in the treatment of HIV infection. The hydroxyl group (red) is designed to mimic the tetrahedral transition state of the enzyme-catalyzed reaction shown in Figure 1. The aromatic groups (orange) are designed to fit into other pockets on the surface of the enzyme.

Some Proteins Are Regulated by Proteolytic Cleavage

In the process of **proteolytic cleavage**, an inactive precursor protein is cleaved to form the active protein. Many eukaryotic proteases (proteolytic enzymes) are regulated in this way. A subunit of the *E. coli* DNA polymerase V (see Chapter 11) called MutD

is also activated in this way, with cleavage producing the active form, MutD^I.

The larger, uncleaved precursor proteins, before proteolytic cleavage, are generally referred to as **proteins** or **proenzymes**, as appropriate. For example, a class of proteins known as transcription factors facilitate the function of RNA polymerases in all organisms. The process of sporulation in the bacterium *Bacillus subtilis* is controlled

in part by a transcription factor called σ^E , which is synthesized as an inactive proprotein, pro- σ^E . The conversion of pro- σ^E to the mature transcription factor involves the regulated proteolytic removal of 27 amino acid residues from the N-terminus of the precursor protein.

As another example, the small number of proteins that are encoded by the genomes of eukaryotic retroviruses are generally synthesized as one large polyprotein, which must be cleaved into the individual functional proteins by a virus-encoded protease. Retroviruses, whose infection cycle is described in Figure 14-16, include the human immunodeficiency virus, or HIV. The requirement for the HIV protease to activate the viral proteins has made this enzyme an important drug target (Highlight 5-2).

SECTION 5.4 SUMMARY

- Specific enzymes, motor proteins, and other proteins are subject to various types of regulation.
- The noncovalent binding of allosteric modulators, either homotropic or heterotropic, can facilitate or inhibit the activity of nucleic acid-binding proteins or enzymes.
- Parts of a protein's own structure can reduce the overall activity of the protein in the process of autoinhibition.
- Covalent modification is a common mechanism used to alter the function of proteins and enzymes. Common modifications involve the addition and removal of phosphoryl, methyl, acetyl, ubiquitinyl, and many other types of groups.
- Ser, Thr, and Tyr residues can be phosphorylated and dephosphorylated by protein kinases and phosphatases, respectively. Kinases are specific to consensus sequences in the target protein, but phosphorylases are less specific.
- Some proteins and enzymes are regulated by proteolytic cleavage. These proteins are synthesized as larger, inactive proproteins or proenzymes and are activated by the proteolytic removal of one or more amino acid residues.

Unanswered Questions

The study of protein function is, arguably, the oldest subdiscipline in biochemistry and molecular biology. But there is still much to learn. The relatively young science of genomics keeps pointing to genes that encode

proteins about which we know little or nothing at all. Some shortcuts to functional discovery are discussed in later chapters.

1. How does protein structure relate to function?

This is an old but still very relevant question for every scientist who studies proteins. Advanced methods of structural analysis are providing more information than ever before, but many of these structural pictures are static. A clear picture of a complete binding or catalytic cycle can require a detailed knowledge of the structure of multiple protein conformations. Certain structural motifs and domains (e.g., the OB fold of single-stranded DNA-binding proteins and others, the AAA⁺ ATPase domain, and simple β-barrel structures) appear in proteins that often have seemingly unrelated functions. The manner in which particular structures are adapted to different functions is an ongoing area of investigation.

2. How do proteins function in the context of large protein assemblies?

Many proteins act only as a part of a much larger protein complex, involving anywhere from a few to many dozens of additional proteins. Unraveling the individual contributions of the subunits of these large complexes has become one of the major challenges of modern molecular biology.

3. Within the context of molecular biology, how many types of protein function remain to be discovered?

A textbook such as this one might leave a student with the impression that the fundamental protein/enzyme activities underlying information pathways are now understood. This impression is incorrect. Although major processes such as DNA replication, RNA transcription, and protein synthesis are increasingly well understood, new types of proteins with important functions are continually being discovered. Many of the newer discoveries involve proteins that have regulatory functions, or facilitate nucleoid or chromatin changes during cell division, or carry out functions in RNA metabolism. There are no boundaries to these frontiers of protein research.

4. How do proteins in the small concentrations found in cells find their interacting partners—particularly, specific sequences on very large nucleic acids—in the complex cellular environment?

This is an issue that remains of great interest to investigators in many areas of molecular biology.

How We Know

The Lactose Repressor Is One of the Great Sagas of Molecular Biology

Rickenberg, H.V., G.N. Cohen, G. Buttin, and J. Monod. 1956.

La galactoside-perméase d'*Escherichia coli*. *Ann. Inst. Pasteur* 91:829–857.



Jacques Monod,
1910–1976 (left);
André Lwoff, 1902–
1994 (middle); and
François Jacob, b.
1920 (right). [Source:
© Bettmann/Corbis.]

Jacques Monod began his scientific career in the 1930s, as a graduate student with André Lwoff, studying the capacity of *E. coli* to adapt its metabolism to different growth conditions. His approach to science was shaped in part by a 1936 trip to Thomas Hunt Morgan's laboratory at the California Institute of Technology, where he found a stimulating environment dominated by open collaboration and free discussion. Back in France, Monod's scientific career was slowed but not halted by the outbreak of World War II. While continuing his research, Monod was an active member of the French underground. His laboratory at the Sorbonne doubled as a meeting place and propaganda printing press. In the lab, he hit on the idea of an inducer, a cellular signal that would trigger the production of new enzymes needed for adapting to new metabolic circumstances. After the war, he turned back to science full-time. With the help of Lwoff, he obtained a position at the Pasteur Institute in Paris. The metabolism of lactose soon caught his attention.

The disaccharide lactose is cleaved to the monosaccharides glucose and galactose by the enzyme β -galactosidase. Monod found that when lactose was not present in the *E. coli* growth medium, β -galactosidase was barely detectable in cell extracts. When lactose was added as the sole carbon source for bacterial growth, the levels of β -galactosidase increased dramatically. Monod wondered how this might occur. In a 1940s' scientific world in which DNA sequencing, PCR (the polymerase chain reaction), and the structure of DNA were unknown, and messenger RNA still remained to be discovered, the question was not trivial.

Many bacterial enzymes of intermediary metabolism exhibited this inducible pattern. What recommended lactose metabolism as a subject of investigation? Like many other laboratories in the late 1940s, the Pasteur Institute lacked modern cold rooms and other facilities now commonly used to keep proteins active during purification. The one inducible enzyme stable enough to survive the

summer heat of an attic lab in Paris was β -galactosidase, later shown to be encoded by a gene called *lacZ*. The enzyme was also fairly easy to assay. When it was present at high levels, it would cleave an alternative substrate, 5-bromo-4-chloro-3-indolyl- β -D-galactopyranoside (denoted, more simply, X-gal), and the indole released would color the bacterial colonies blue. The Monod group initiated a study of bacterial lactose metabolism that featured a creative union of biochemistry and genetics.

Mutations were soon found that affected the induction of β -galactosidase. Colonies of these mutants turned blue on X-gal plates even in the absence of lactose. Most of the mutant cells had a mutation in a gene that came to be known as the *i* (inducer) gene, and later the *lacI* gene. In these mutations, the *lacZ* gene was expressed (produced β -galactosidase) all the time; this is known as constitutive expression (Figure 1). Moreover, the appearance of β -galactosidase activity was paralleled by the appearance of a function that transported lactose into the cell, an activity that Monod called a galactoside permease (this is encoded by the gene *lacY*). The coordinated regulation of two genes, and the loss of regulation in constitutive mutants, led to the concept that some genes regulate other genes. When Monod explained this then-revolutionary idea to his wife, a nonscientist, he was somewhat pained at her reply: "Of course, it is obvious!" Next, Monod had to figure out what the *lacI* gene was doing.

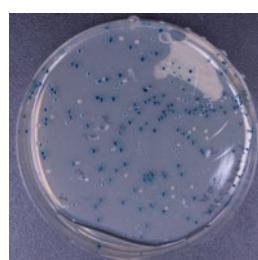


FIGURE 1 Bacterial colonies growing on an agar plate containing X-gal. Cells in the blue colonies have a mutation that results in constitutive expression of the *lac* genes. [Source: C. Mönchmeier (Benutzer: Luziferase).]

The *lacI* Gene Encodes a Repressor

Jacob, F., and J. Monod. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* 3:318–356.

Next door to Monod at the Pasteur Institute was the laboratory of François Jacob. Wounded as a member of the Free French army during the Normandy campaign, Jacob could not pursue his planned career in surgery and instead turned to science. He studied a phenomenon dubbed the “erotic induction” of bacteriophage λ . Bacteriophage λ is a prophage, a bacterial virus that integrates its DNA into its host chromosome and remains quiescent there.

Some years earlier, Joshua Lederberg had discovered the phenomenon of bacterial conjugation (bacterial sex), in which DNA is transferred from a donor to a recipient cell. The transfer is mediated by a genetic element, separate from the bacterial chromosome, known as the F plasmid. In some cells, this plasmid becomes integrated into the host chromosome, creating a strain that mediates high-frequency (Hfr) transfer of chromosomal genes. The transfer begins at the location where the F plasmid is integrated and proceeds linearly along the bacterial chromosome.

Jacob, working with Elie Wollman, used this technique to produce some of the first genetic maps, and they even deduced that the *E. coli* chromosome was circular. They also noticed that when an Hfr strain also contained an integrated prophage, the bacteriophage λ was transferred into the recipient along with the chromosomal genes. When the dormant bacteriophage λ was thus transferred to a strain that lacked an integrated prophage, the recipient cells were soon lysed. The prophage was somehow activated when it entered the recipient, leading to lytic reproduction—the erotic induction. It gradually dawned on both Monod and Jacob that the induction phenomena they were studying were closely related, and one of the great scientific collaborations of the twentieth century was born.

Monod first liked the idea that mutation in the *lacI* gene (*lacI⁻*) produced some kind of inducer that made the addition of lactose unnecessary. If this was true, then the *lacI⁻* gene should be dominant over the normal

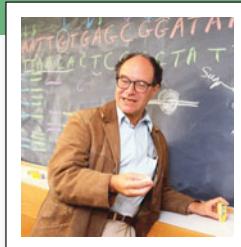
(wild-type) *lacI* gene (*lacI⁺*). To get the two variations of *lacI* (*lacI⁺* and *lacI⁻*) into the same cell, Monod worked with Jacob and Arthur Pardee (on sabbatical leave from the University of California, Berkeley) to use the bacterial conjugation method. The resulting “PaJaMo” experiments at first supported Monod’s idea. When an Hfr strain was used to introduce the *lacI⁺* and *lacZ⁺* genes into a recipient that had both the *lacI⁻* gene and another mutation that prevented β -galactosidase synthesis (*lacZ⁻*), β -galactosidase production occurred immediately after the *lacZ⁺* gene arrived in the recipient. This seemed to indicate that the *lacI⁻* gene of the recipient was dominant. However, after an hour or so, β -galactosidase production was halted as the product of the *lacI⁺* gene built up. Clearly, *lacI⁺* encoded a substance that shut off the *lacZ* gene. Continued experimentation refined the idea that *lacI⁺* encodes some kind of repressor, and that the interaction of lactose with the repressor is needed to induce the *lacZ* gene and the additional genes that are coregulated with it.

The work continued. If *lacI* was encoding a repressor, this repressor had to interact with something to shut down the lactose genes. Monod and Jacob now mutagenized cells that had two good copies of the lactose genes. It was unlikely that both copies of *lacI* would be inactivated, but inactivation of one repressor target would be enough to induce the function of one set of lactose genes. The researchers predicted that the resulting mutations would appear at a site on the chromosome distinct from *lacI* and would lead to constitutive synthesis of the lactose gene products. The mutants were found, and they defined a site that Monod and Jacob called the operator.

In a famous 1961 paper, Jacob and Monod laid out these ideas and others as part of their operon model. The concepts have guided our thinking about gene regulation ever since. Other experiments, carried out in parallel, showed that bacteriophage λ also encoded a repressor, and this repressor was needed to keep most other bacteriophage λ genes from being expressed.

The Lactose Repressor Is Found

Gilbert, W., and B. Müller-Hill. 1966. Isolation of Lac repressor. Proc. Natl. Acad. Sci. USA 56:1891-1898.



Walter Gilbert [Source: Louie Psihogios/
Science Faction/Corbis.]

By 1961, it was clear that a repressor existed, but not at all clear what it was. It could be RNA or protein, or some other kind of molecule that was synthesized by a *lacI* enzyme. By 1966, nonsense-type mutations, those that prematurely halt protein synthesis, had been found in the *lacI* gene that produced a *lacI*⁻ effect. This finding had convinced most scientists that the repressor was a protein, but it was still necessary to isolate one to prove it. The isolation work was carried out at Harvard, by physicist-turned-biologist Walter Gilbert.

To isolate a protein, you need a way to measure its presence (an assay), but that was difficult to construct. The repressor presumably bound to operator DNA, but that DNA sequence was not yet defined. The repressor also bound to the inducer (then thought to be lactose, later shown to be allolactose, a metabolic by-product of lactose metabolism). Because lactose was metabolized in the cell, it would be destroyed in crude cell extracts and thus would be of little use to the researchers. Gilbert turned to isopropyl β -D-1-thiogalactopyranoside (IPTG), a molecule known to induce the lactose operon without being metabolized in the cell.

Using a technique known as equilibrium dialysis, the researchers suspended a dialysis bag containing a bacterial cell extract in a solution containing radioactive IPTG. Pores in the dialysis bag were sufficiently small to prevent the protein molecules from escaping, but smaller molecules such as IPTG could diffuse through. If the lactose repressor was present in the extract, it would bind to IPTG, and the concentration of IPTG would increase inside the dialysis bag relative to the surrounding solution. The assay eventually worked, but only after Gilbert and his colleagues developed methods to greatly increase its sensitivity. In 1966, they reported their detection of a repressor protein that bound IPTG.

The lactose repressor was finally purified to homogeneity by the Gilbert group and several other research groups in the early 1970s. Gilbert used the repressor problem to develop several new methods to define the DNA binding sites of proteins that bound DNA specifically. For example, his group found that when dimethylsulfate (DMS) was used to modify purine residues in the DNA, the adjacent DNA backbone became more labile to cleavage in mild alkali. DMS methylates guanine residues at N-7, adenine residues at N-3. If the treatment is done briefly so that only one purine per DNA strand is labeled, on average,

then subsequent cleavage will break each DNA strand at just one position. If all the strands are radioactively labeled at the same end, a banding pattern is produced that acts as a map of the A and G residues in the DNA strand (**Figure 2**; note that G residues are methylated preferentially and generate stronger bands). If a protein is bound to the DNA prior to addition of DMS, it partially protects the purines from methylation in the region where it is bound. The resulting disruption of the base-methylation pattern helps to define the binding site of the protein; in Figure 2, the protein is the lactose repressor.

In this technique, the control lane (as seen in Figure 2) proved to be as important as the actual experiment. As Gilbert and his colleague Allan Maxam looked at one of these gels, they realized that the gel could be read to reveal the positions of all the A and G residues in the DNA strand. If they could find a technique to break the strands at C and T residues as well, they would have a new way to sequence DNA. The methods were developed and published as a new sequencing technology in 1977. The Maxam-Gilbert DNA-sequencing procedure was eventually supplanted by the Sanger sequencing method, published in the same year; that method, in turn, has given way to newer technologies (see Chapter 7). However, the effects of all these advances were profound, and they helped define the science of molecular biology in their time.

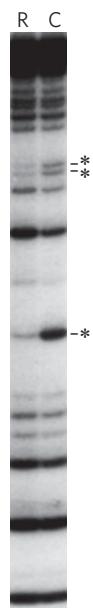


FIGURE 2 In this polyacrylamide gel, the C lane contains a control to which no repressor was bound. The darker bands pinpoint G residues; the less intense bands result from cleavage at A residues. The R lane includes added repressor. Several bands (*) exhibit diminished intensity in the R lane, defining sites to which the repressor was bound. [Source: R. T. Ogata and W. Gilbert, *J. Mol. Biol.* 132:709–728, 1979. Photo courtesy of Ronald Ogata.]

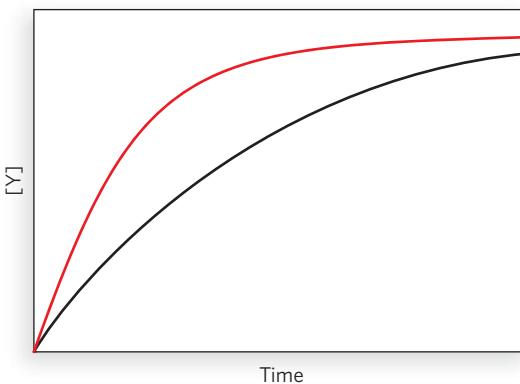
Key Terms

enzyme, p. 136	substrate, p. 146	regulatory enzyme, p. 161
ligand, p. 136	active site, p. 146	allosteric enzyme (allosteric protein), p. 161
binding site, p. 136	biochemical standard free-energy change ($\Delta G^{\circ o}$), p. 147	allosteric modulator, p. 161
induced fit, p. 136	transition state, p. 148	homotropic, p. 161
cooperativity, p. 136	maximum velocity (V_{max}), p. 149	heterotropic, p. 161
dissociation constant (K_d), p. 137	steady-state kinetics, p. 150	autoinhibition, p. 163
cofactor, p. 145	Michaelis-Menten equation, p. 150	protein kinase, p. 165
coenzyme, p. 145	Michaelis constant (K_m), p. 150	protein phosphatase, p. 165
prosthetic group, p. 145	turnover number, p. 151	
holoenzyme, p. 145	helicase, p. 157	
apoenzyme (apoprotein), p. 145		

Problems

1. Protein A has a binding site for ligand X with a K_d of 10^6 M. Protein B has a binding site for ligand X with a K_d of 10^9 M. Which protein has a higher affinity for ligand X? Explain your reasoning. Convert K_d to K_a for both proteins.
2. The lactose (Lac) repressor has a binding site for DNA and binds to a specific DNA site with a K_d of approximately 10^{-10} M. The Lac repressor also has a binding site for the galactoside allolactose. When the Lac repressor interacts with allolactose, it dissociates from the DNA. When allolactose is bound to the Lac repressor, does the K_d for the repressor's specific DNA binding site increase or decrease? Explain.
3. The binding of any protein to DNA invariably involves the displacement of other atoms or molecules. What types of atoms or molecules are most commonly displaced?
4. When a protein binds to DNA nonspecifically, with which parts of the DNA does the protein generally interact? When a protein binds to DNA at a site defined by a particular nucleotide sequence, with which parts of the DNA does the protein generally interact?
5. Which of the following situations would produce negative cooperativity (i.e., the binding of a ligand to one subunit would *decrease* the affinity of another subunit for the same ligand)? Explain your reasoning in each case.
- (a) The protein has multiple subunits, each with a single ligand-binding site. The binding of ligand to one site decreases the binding affinity of other sites for the ligand.
- (b) The protein is a single polypeptide with two ligand-binding sites, each having a different affinity for the ligand.
- (c) The protein is a single polypeptide with a single ligand-binding site. As purified, the protein preparation is heterogeneous, containing some protein molecules that are partially denatured and thus have a lower binding affinity for the ligand.
6. To approximate the actual concentration of enzymes in a bacterial cell, assume that the cell contains equal concentrations of 1,000 different enzymes in solution in the cytosol and that each protein has a molecular weight of 100,000. Assume also that the bacterial cell is a cylinder (diameter 1.0 μm , height 2.0 μm), that the cytosol (specific gravity 1.20) is 20% soluble protein by weight, and that the soluble protein consists entirely of enzymes. Calculate the *average* molar concentration of each enzyme in this hypothetical cell.
7. Which of the following effects would be brought about by any enzyme catalyzing the simple reaction $S \xrightleftharpoons[k_2]{k_1} P$, where $K_{eq} = [P]/[S]$? (a) Decreased K_{eq} ; (b) increased k_1 ; (c) increased K_{eq} ; (d) increased ΔG^\ddagger ; (e) decreased ΔG^\ddagger ; (f) more negative $\Delta G^{\circ o}$; (g) increased k_2 .
8. In the late nineteenth century, the famous chemist Emil Fischer proposed that an enzyme should possess a pocket that is complementary in shape to the substrate it interacts with in its catalytic function. This “lock and key” hypothesis was highly influential at the time. Although Fischer made many important contributions to the development of enzymology, this particular idea was largely wrong. Explain why.
9. If an irreversible inhibitor inactivates an enzyme, what is the effect on that enzyme's k_{cat} and K_m ?
10. (a) At what substrate concentration would an enzyme with a k_{cat} of 30.0 s^{-1} and a K_m of 0.0050 M operate at one-quarter of its maximum rate?
- (b) Determine the fraction of V_{max} that would be achieved at the following substrate concentrations: $[S] = \frac{1}{2}K_m$, $2K_m$, and $10K_m$.
- (c) An enzyme that catalyzes the reaction $X \rightleftharpoons{} Y$ is isolated from two bacterial species. The two enzymes

have the same V_{\max} but different K_m values for the substrate X. Enzyme A has a K_m of 2.0 μM , and enzyme B has a K_m of 0.5 μM . The plot below shows the kinetics of reactions carried out with the same concentration of each enzyme and with $[X] = 1 \mu\text{M}$. Which curve corresponds to which enzyme?



- 11.** A research group discovers a new enzyme, which they call happyase, that catalyzes the chemical reaction $\text{HAPPY} \rightleftharpoons \text{SAD}$. The researchers begin to characterize the enzyme.
- In the first experiment, with a total concentration of enzyme, $[E_t]$, at 4 nm, they find that $V_{\max} = 1.6 \mu\text{M s}^{-1}$. Based on this experiment, what is the k_{cat} for happyase? (Include the appropriate units.)
 - In another experiment, with $[E_t]$ at 1 nm and $[\text{HAPPY}]$ at 30 μM , the researchers find that $V_0 = 300 \text{ nm s}^{-1}$. What is the measured K_m of happyase for its substrate HAPPY? (Include the appropriate units.)
 - Further research shows that the supposedly purified happyase used in the first two experiments was actually contaminated with a reversible inhibitor called ANGER. When ANGER is carefully removed from the happyase preparation, and the two experiments are repeated, the measured V_{\max} in (a) increases to $4.8 \mu\text{M s}^{-1}$, and the measured K_m in (b) is 15 μM . For the inhibitor ANGER present in the original preparation, calculate the values of α and α' (see Highlight 5-1).
 - Based on the information given above, what type of inhibitor is ANGER?
- 12.** An enzyme is discovered that catalyzes the reaction $\text{A} \rightleftharpoons \text{B}$. Researchers find that the K_m for substrate A is 4 μM , and the k_{cat} is 20 min^{-1} .
- In an experiment, $[A] = 6 \text{ mM}$, and the initial velocity, V_0 , is measured as 480 nm min^{-1} . What is the $[E_t]$ used in the experiment?

(b) In another experiment, $[E_t] = 0.5 \text{ nm}$, and the measured $V_0 = 5 \text{ nm min}^{-1}$. What is the $[A]$ used in the experiment?

(c) The compound Z is found to be a very strong competitive inhibitor of the enzyme. In an experiment with the same $[E_t]$ as in part (a) but a different $[A]$, an amount of Z is added that produces an $\alpha = 10$. This reduces V_0 to 240 nm min^{-1} . What is the $[A]$ in this experiment?

- 13.** Although graphical methods are available for accurate determination of the V_{\max} and K_m of an enzyme-catalyzed reaction (see Additional Reading), sometimes these quantities can be quickly estimated by inspecting values of V_0 at increasing $[S]$. Estimate the V_{\max} and K_m of the enzyme-catalyzed reaction for which the following data were obtained.

$[S] (\text{M})$	$V_0 (\mu\text{M/min})$
2.5×10^{-6}	28
4.0×10^{-6}	40
1×10^{-5}	70
2×10^{-5}	95
4×10^{-5}	112
1×10^{-4}	128
2×10^{-3}	139
1×10^{-2}	140

- 14.** The bacterial RuvB protein, a DNA translocase, belongs to helicase superfamily 6. RuvB functions as a circular hexameric complex with a central opening. Duplex DNA is bound in the opening. RuvB is also an ATPase, and it moves along the DNA when ATP is hydrolyzed. An amino acid substitution in the ATP-binding site of RuvB generates a protein that binds to DNA but does not hydrolyze ATP and does not translocate along the DNA. When normal RuvB protein subunits are mixed with an equal amount of mutant RuvB subunits, heterohexameric complexes are formed that contain both normal and mutant subunits. These heterohexamers can hydrolyze ATP but do not move along the DNA. What conclusions can you draw from these observations?

- 15.** When eukaryotic DNA replication is prematurely halted, due to DNA damage or other causes, two proteins—ATM (ataxia telangiectasia mutated) and ATR (ATM related)—initiate a response called a checkpoint, which involves regulated changes in the functions of many cellular proteins to facilitate DNA repair. ATM and ATR are enzymes with a regulatory function. They alter the covalent structure of the proteins they regulate, increasing their measured molecular weight. Suggest what kind of enzymatic activity they might possess.

Additional Reading

General

- Kornberg, A.** 1989. Never a dull enzyme. *Annu. Rev. Biochem.* 58:1–30. An especially illuminating essay for young scientists.
- Kornberg, A.** 1990. Why purify enzymes? *Methods Enzymol.* 182:1–5.
- Kornberg, A.** 1996. Chemistry: The lingua franca of the medical and biological sciences. *Chem. Biol.* 3:3–5. This and the two articles above provide inspiration from one of the great biochemists of the past century.
- Nelson, D.L., and M.M. Cox.** 2008. *Lehninger Principles of Biochemistry*, 5th ed. New York: W.H. Freeman. See Chapters 3 through 6 for more detailed background on kinetics.
- von Hippel, P.H.** 2007. From “simple” DNA-protein interactions to the macromolecular machines of gene expression. *Annu. Rev. Biophys. Biomol. Struct.* 36:79–105.

Protein-Ligand Interactions

- Jayaram, B., and T. Jain.** 2004. The role of water in protein-DNA recognition. *Annu. Rev. Biophys. Biomol. Struct.* 33:343–361.
- Kalodimos, C.G., N. Biris, A.M.J.J. Bonvin, M.M. Levandoski, M. Guennuegues, R. Boelens, and R. Kaptein.** 2004. Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. *Science* 305: 386–389.
- Lohman, T.M., and D.P. Mascotti.** 1992. Thermodynamics of ligand-nucleic acid interactions. *Methods Enzymol.* 212:400–424.
- Raghunathan, S., A.G. Kozlov, T.M. Lohman, and G. Waksman.** 2000. Structure of the DNA binding domain of *E. coli* SSB bound to ssDNA. *Nat. Struct. Biol.* 7:648–652.

Enzymes: The Reaction Catalysts of Biological Systems

- Arabshahi, A., and P.A. Frey.** 1999. Standard free energy for the hydrolysis of adenylylated T4 DNA ligase and the apparent pK_a of lysine 159. *J. Biol. Chem.* 274:8586–8588.
- Ellenberger, T., and A.E. Tomkinson.** 2008. Eukaryotic DNA ligases: Structural and functional insights. *Annu. Rev. Biochem.* 77:313–338. A complete summary of many details gleaned from structural analysis.
- Lehman, I.R.** 1974. DNA ligase: Structure, mechanism, and function. *Science* 186:790–797.

- Liu, P., A. Burdzy, and L.C. Sowers.** 2004. DNA ligases ensure fidelity by interrogating minor groove contacts. *Nucleic Acids Res.* 32:4503–4511.

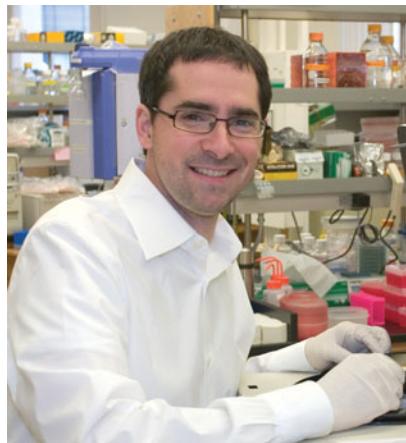
Motor Proteins

- Lohman, T.M., and K.P. Bjornson.** 1996. Mechanisms of helicase-catalyzed DNA unwinding. *Annu. Rev. Biochem.* 65:169–214.
- Lohman, T.M., E.J. Tomko, and C.G. Wu.** 2008. Non-hexameric DNA helicases and translocases: Mechanisms and regulation. *Nat. Rev. Mol. Cell Biol.* 9:391–401.
- Pyle, A.M.** 2008. Translocation and unwinding mechanisms of RNA and DNA helicases. *Annu. Rev. Biophys.* 37: 317–336
- Singleton, M.R., M.S. Dillingham, and D.B. Wigley.** 2007. Structure and mechanism of helicases and nucleic acid translocases. *Annu. Rev. Biochem.* 76:23–50.

Regulation of Protein Function

- Bialik, S., and A. Kimchi.** 2006. The death-associated protein kinases: Structure, function, and beyond. *Annu. Rev. Biochem.* 75:189–210.
- Elphick, L.M., S.E. Lee, V. Gouverneur, and D.J. Mann.** 2007. Using chemical genetics and ATP analogues to dissect protein kinase function. *ACS Chem. Biol.* 2:299–314.
- Gelato, K.A., and W. Fischle.** 2008. Role of histone modifications in defining chromatin structure and function. *Biol. Chem.* 389:353–363.
- Martin, C., and Y. Zhang.** 2005. The diverse functions of histone lysine methylation. *Nat. Rev. Mol. Cell Biol.* 6: 838–849.
- Millar, C.B., and M. Grunstein.** 2006. Genome-wide patterns of histone modifications in yeast. *Nat. Rev. Mol. Cell Biol.* 7:657–666.
- Moorhead, G.B.G., L. Trinkle-Mulcahy, and A. Ulke-Leme. 2007. Emerging roles of nuclear protein phosphatases. *Nat. Rev. Mol. Cell Biol.* 8:234–244.**
- Shahbazian, M.D., and M. Grunstein.** 2007. Functions of site-specific histone acetylation and deacetylation. *Annu. Rev. Biochem.* 76:75–100.
- Tonks, N.K.** 2006. Protein tyrosine phosphatases: From genes, to function, to disease. *Nat. Rev. Mol. Cell Biol.* 7:833–846.

DNA and RNA Structure



Jamie Cate [Source: Michael Barnes, UC Berkeley College of Chemistry.]

got one small perfect-looking crystal to grow from a purified ribosome sample. We took the crystal to a synchrotron x-ray beamline and saw the first diffraction pattern indicating that the crystal indeed contained ribosomes. That was so exciting! But just to be sure, we recovered the crystal from the x-ray diffraction apparatus, dissolved it in water, and checked the contents on an agarose gel—and there was the ribosomal RNA, clear and pristine.

We figured out how to grow more of those crystals and eventually we solved the crystal structure of the complete ribosome. Staring in awe at the electron density map, the RNA helices wove through the molecule like great curving spiral staircases. I felt chills down my spine, realizing I was the first person to see such incredible molecular beauty. This was the culmination of six years of challenging—and at times frustrating—experiments. I felt the sweet joy of success, and also contemplated the many new discoveries that would result from this work.

—**Jamie Cate**, on determining the molecular structure of the bacterial ribosome

Moment of Discovery

When I first started my lab at the Whitehead Institute, my dream was to *crystallize the bacterial ribosome and to solve its molecular structure at high resolution*. Although a lot had been learned about RNA structure from work on catalytic RNAs and the individual subunits of the ribosome, the possibility of seeing the complete structure of the protein-synthesizing machinery was irresistible.

Working closely with graduate student Steve Santoso, we eventually

- 6.1 The Structure and Properties of Nucleotides 177**
- 6.2 DNA Structure 185**
- 6.3 RNA Structure 194**
- 6.4 Chemical and Thermodynamic Properties of Nucleic Acids 200**

Discovered in the nineteenth century, DNA (deoxyribonucleic acid) was proposed, by the early twentieth century, as the molecule that stores biological information (see Chapter 2). At that time, however, the way in which the particular properties of its molecular structure could produce traits and behaviors in living organisms was unimaginable. Hoping to determine how DNA carried genetic messages that are faithfully transmitted to the next generation when cells divide, researchers in several laboratories, in the 1950s, made it their goal to solve the molecular structure of DNA. In 1953, James Watson and Francis Crick, at Cambridge University, used x-ray diffraction data obtained by Rosalind Franklin to deduce DNA's simple and beautiful double-helical structure (Figure 6-1). This landmark discovery, for which Watson and Crick (together with Maurice Wilkins, for his work on the x-ray diffraction) received the Nobel Prize in Physiology or Medicine in 1962, gave rise to all of modern molecular biology. It



James Watson [Source: Associated Press.]

Francis Crick, 1916–2004 [Source: Associated Press.]

was immediately apparent to scientists how this unique structure of DNA could allow biological information to be easily and faithfully duplicated and transmitted from generation to generation.

Like DNA, RNA (ribonucleic acid) was first isolated in the nineteenth century from the nuclei of cells. Scientists later recognized that RNA is chemically distinct from DNA, because it contains a different kind of sugar in its nucleotide building blocks (see Chapter 3). As described in Chapter 2, ribosomal RNAs (rRNAs) were found to be components of ribosomes, the complexes that carry out protein synthesis. Messenger RNAs (mRNAs) were known to be intermediaries, carrying genetic information from genes to ribosomes. And transfer RNAs (tRNAs) had been identified as adaptor molecules that translate the information in mRNA into a specific sequence of amino acids. We now know that RNA molecules have many other biological functions as well. For example, they comprise the genomes of certain viruses, such as the human immunodeficiency virus (HIV) and hepatitis C virus (HCV). Some RNA molecules have the ability to work as catalysts—a discovery that provided, for the first time, a plausible scenario for the evolution of early life forms based on self-replicating RNA. (The diversity of functional RNAs and their roles in evolution are discussed in Chapters 15 and 16.) In the quest to understand how RNA could perform such a range of functions, researchers have determined the structures of numerous types of RNA molecules and RNA-protein complexes, including the structure of the ribosome itself. Unlike DNA, RNA molecules are almost always single-stranded and consist of much shorter chains of nucleotides. They also have a propensity to

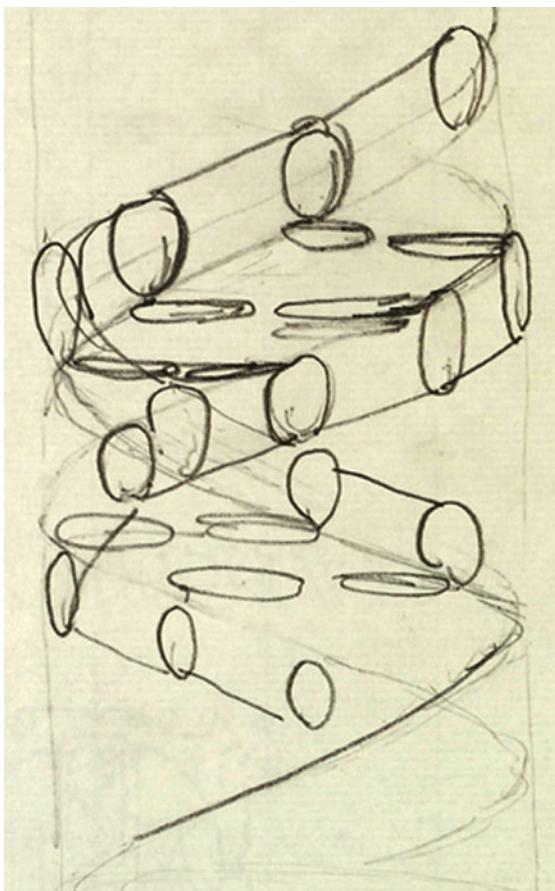


FIGURE 6-1 Francis Crick's first drawing of DNA structure.

Two base-paired strands of DNA form a helical structure in which the phosphate and sugar groups are on the outside and the bases are on the inside. The helix twists in a right-handed direction. [Source: © Photo Researchers/Alamy.]

fold back on themselves, creating many discrete double-helical regions that can assemble into complex three-dimensional structures.

As we'll see, there is no single generic structure of DNA or RNA. Rather, a huge number of variations on a common structural theme result from the chemical and physical properties of the polynucleotide chain. Indeed, the structural stability of DNA and the structural diversity of RNAs explain why these molecules have evolved to function in all aspects of maintaining and transmitting biological information. In this chapter, we first explore the general properties of nucleotides, then turn to the structures of DNA and RNA. We conclude by looking at the chemical behavior of nucleic acids under biological conditions.

6.1 The Structure and Properties of Nucleotides

All nucleic acids are chemically linked chains of nucleotides, the basic building blocks of DNA and RNA. To understand the structures, functions, and replication of nucleic acids, we first need to understand the structure of their nucleotide components and how they behave in the context of a DNA or RNA polymer. We therefore begin our discussion of DNA and RNA by considering the nature of the nucleotide.

Nucleotides Comprise Characteristic Bases, Sugars, and Phosphates

A **nucleotide** is a molecule consisting of three characteristic components: a heterocyclic base, a five-carbon sugar called a pentose, and a phosphate group. The same molecule without the phosphate group is called a **nucleoside**. Each base is a derivative of one of two parent compounds, a **purine** or a **pyrimidine** (Figure 6-2a), which are nitrogenous bases. They are called bases because free purines and pyrimidines are weakly basic compounds. The carbon and nitrogen atoms in the parent structures are numbered according to convention to facilitate the naming and identification of the many derivative compounds. The carbon atoms in the pentose are also numbered; in nucleotides and nucleosides, these numbers are given a prime ('') designation to distinguish the sugar atoms from the numbered atoms of the nitrogenous bases.

In nucleosides, the covalent joining of a base (at N-9 of purines and N-1 of pyrimidines) to the 1' carbon (C-1') of the pentose forms a **glycosidic bond** (specifically, an *N*- β -glycosyl bond), which involves the loss of a molecule of water. To form a nucleotide, a phosphate

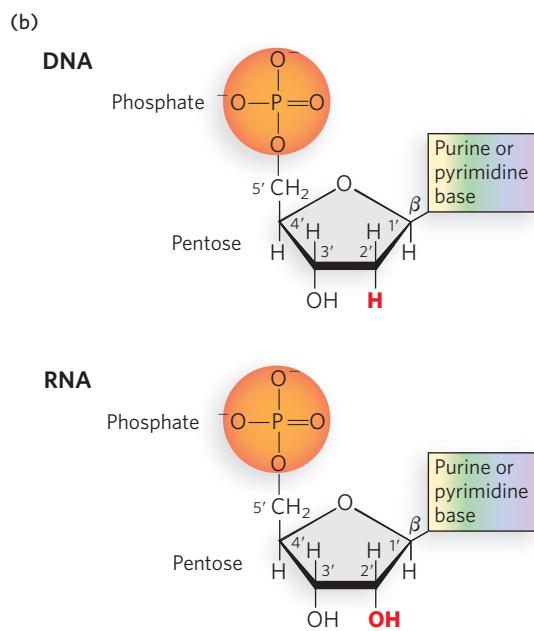
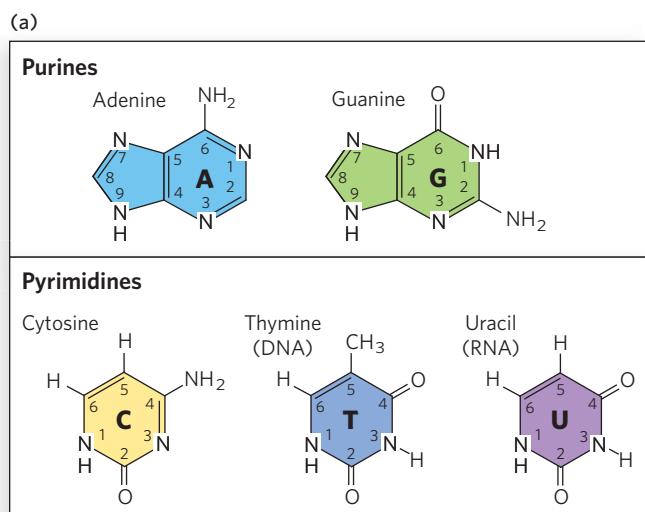


FIGURE 6-2 The chemical composition of nucleotides.

(a) The bases are purines, with nine-membered rings, or pyrimidines, with six-membered rings, with numbering systems as shown. In DNA and RNA, the purines are adenine and guanine; in DNA, the pyrimidines are cytosine and thymine; in RNA, the pyrimidines are cytosine and uracil.

(b) Nucleotides consist of a phosphate, a pentose sugar, and a heterocyclic base; carbons in the pentose rings are numbered as shown, with numbers followed by a prime ('') to distinguish them from the numbered atoms of the bases. In DNA, the pentose is 2'-deoxyribose, which has no hydroxyl group on the 2' carbon (red); in RNA, the sugar is ribose, which includes a 2' hydroxyl. A glycosidic bond links the 1' carbon of ribose or deoxyribose to the base; the β indicates its direction relative to the pentose ring.

group is covalently joined to the 5' carbon (C-5') of the pentose to form an ester, also with the concomitant loss of a water molecule (**Figure 6-2b**).

Four different bases are found in DNA: two are purines, **adenine (A)** and **guanine (G)**, and two are pyrimidines, **cytosine (C)** and **thymine (T)**. RNA also contains four types of bases. The two purines are the same as those in DNA: adenine and guanine; and, as in DNA, one of the pyrimidines is cytosine. However, the second major pyrimidine in RNA is **uracil (U)** instead of thymine. Only rarely does thymine occur in RNA, or uracil in DNA. The structures of the five major bases are shown in Figure 6-2a; the nomenclature of their corresponding nucleotides and nucleosides is summarized in Table 3-1.

Nucleic acids have two kinds of pentoses. The recurring nucleotide units of DNA contain 2'-deoxy-d-ribose, whereas the nucleotide units of RNA contain d-ribose. The d-ribose has a hydroxyl group attached to the 2' carbon, whereas 2'-deoxy-d-ribose lacks this functional group (see Figure 6-2b). In nucleotides, both types of pentoses are in their β -furanose (closed five-membered ring) form (**Figure 6-3a**). As **Figure 6-3b** shows, the pentose ring is not planar, but exists in one of a variety of conformations generally described as “puckered.” The predominant type of sugar pucker that characterizes DNA differs from that found in RNA, resulting in the different shapes and geometries of the DNA and RNA double helices, as we'll see later in this chapter.

Because of their different pentose components, the structural units of DNAs and RNAs are known as **deoxyribonucleotides** (deoxyribonucleoside 5'-monophosphates) and **ribonucleotides** (ribonucleoside 5'-monophosphates), respectively (**Figure 6-4**). Although the major purine and pyrimidine nucleotides are the most common, both DNA and RNA molecules also contain some minor bases. In DNA, the minor bases are usually methylated forms of the major bases. These unusual bases in DNA molecules often have roles in regulating or protecting the genetic information. Minor bases of many types are also found in RNA molecules, particularly in tRNAs, rRNAs, and other RNAs whose function requires a specific three-dimensional structure. In cells, minor bases in RNA can be formed by enzymatic modification of one of the common nucleotides to add or remove a functional group, or by complete replacement of a standard base with a less common one. Chemical modifications of DNA and RNA and their effects on nucleotide structure and function are discussed in Section 6.4.

Cells also contain nucleotides with phosphate groups in positions other than on the 5' carbon

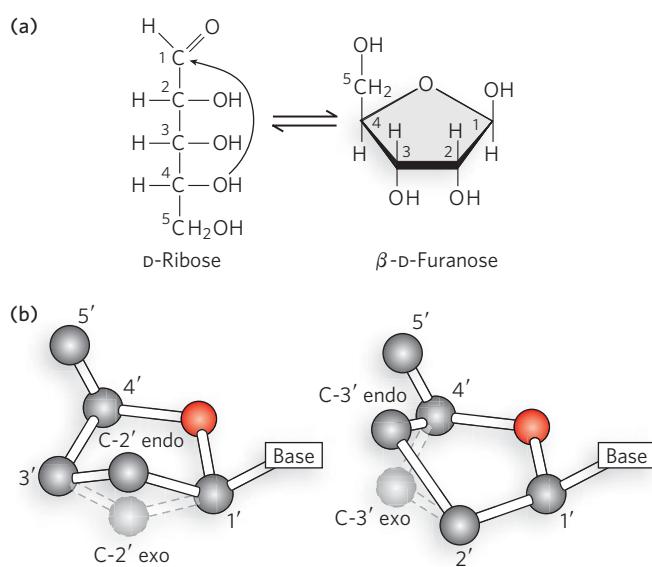


FIGURE 6-3 Pentose ring structures in nucleic acids.

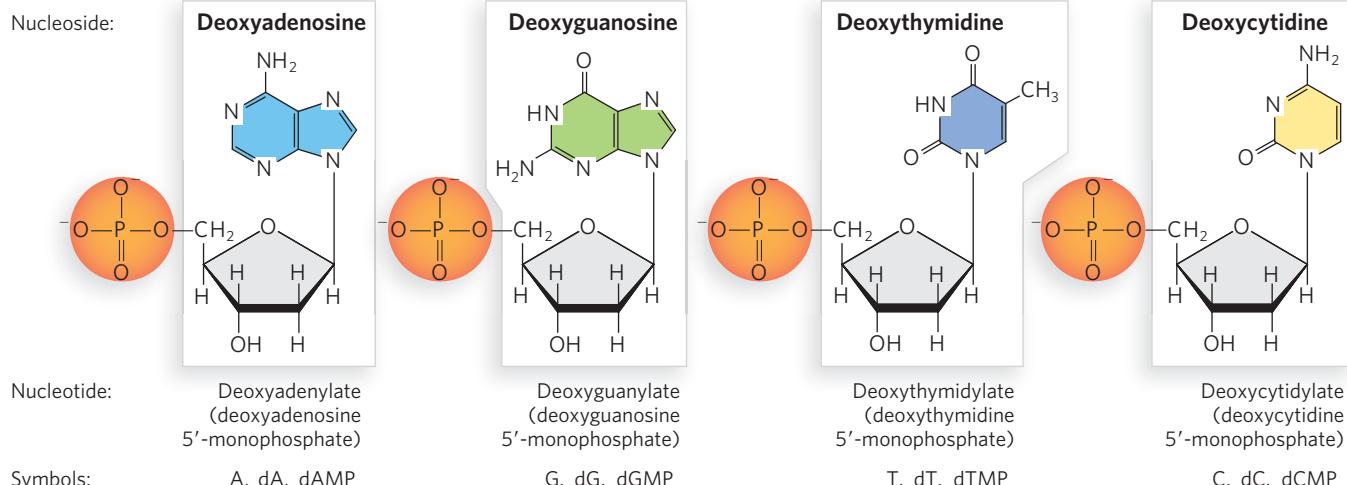
(a) The linear and closed-ring forms of ribose are in equilibrium in solution. When incorporated into nucleosides, nucleotides, or polynucleotides, the pentose exists only in the ring form. The pentose ring is formed by reaction of the hydroxyl group on C-4 with the aldehyde at C-1. (b) The pentose rings in nucleosides and nucleotides can exist in four distinct puckered conformations. In each case, four of the five ring atoms lie in a plane, but the fifth ring atom, either C-2' or C-3', is out of the plane. The C-2' endo configuration, in which the C-2' atom points in the same direction as the C-5' atom, predominates in DNA. The C-3' endo configuration, in which the C-3' atom points in the same direction as the C-5' atom, predominates in RNA.

(**Figure 6-5**). For example, **ribonucleoside 2',3'-cyclic monophosphates** are stable intermediates, and **ribonucleoside 2'-monophosphates** or **ribonucleoside 3'-monophosphates** are the end products of the hydrolysis of RNA by enzymes called **ribonucleases**. Other variations are adenosine 3',5'-cyclic monophosphate (cAMP) and guanosine 3',5'-cyclic monophosphate (cGMP), which are important chemical signals of the metabolic state of the cell (further discussed below).

Phosphodiester Bonds Link the Nucleotide Units in Nucleic Acids

The successive nucleotides of DNA and RNA are covalently joined through phosphate group “connectors” in which the 5'-phosphate group of one nucleotide unit is linked to the 3'-hydroxyl group of the next nucleotide, creating a **phosphodiester bond** (**Figure 6-6**); this involves the loss of water, and the joined nucleotides are

(a) Deoxyribonucleotides



(b) Ribonucleotides

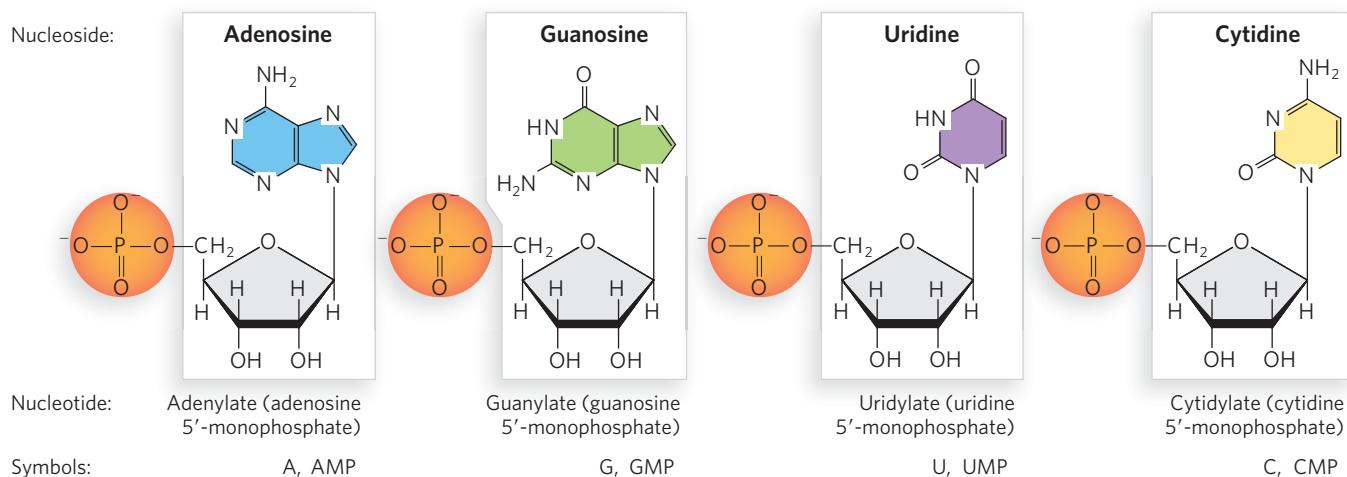


FIGURE 6-4 Deoxyribonucleotides and ribonucleotides of nucleic acids. All nucleotides are shown in their predominant form at neutral pH. (a) Deoxyribonucleotides of DNA. (b) Ribonucleotides of RNA.

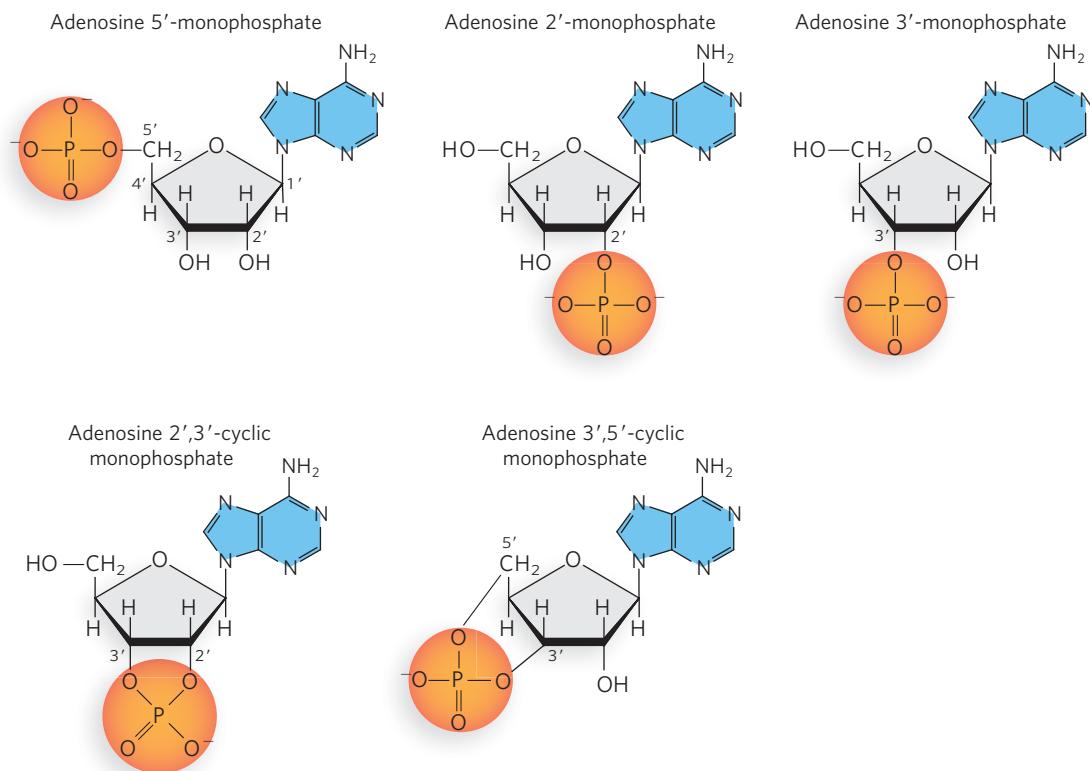
therefore referred to as “residues.” As we’ll see, these 5'-to-3' links give every DNA or RNA chain a directionality, or polarity. The alternating phosphate and sugar residues form the backbone of the nucleic acid, and the bases can be viewed as side groups joined to this sugar-phosphate backbone at regular intervals.

KEY CONVENTION

The polarity of a single DNA or RNA chain is defined by the chemical groups—the free 5' phosphate or 3' hydroxyl—at the termini of the chain, not by the 5' and 3' oxygens of internal phosphodiester bonds. Linear DNA and RNA molecules each have a single unique 5' terminus and 3' terminus.

The backbone of both DNA and RNA is hydrophilic. The hydroxyl groups of the sugar residues form hydrogen bonds with water. The phosphate groups, with a pK_a near 2, are completely ionized and negatively charged at pH 7, and the negative charges are generally neutralized by ionic interactions with positive charges on proteins, metal ions, or short, linear organic molecules called polyamines, which contain two or more amine groups.

The covalent backbone of DNA and RNA is subject to slow, nonenzymatic hydrolysis of the phosphodiester bonds. In the test tube, RNA is hydrolyzed rapidly under alkaline conditions, but DNA is not; the 2'-hydroxyl group on the sugars in RNA is directly involved in the hydrolytic process. Cyclic 2',3'-monophosphates are the

**FIGURE 6-5 Examples of adenosine monophosphates.**

Adenosine 5'-monophosphate, with a phosphate group on the C-5', is the most common adenine-containing nucleotide, and the one found in RNA. Adenosine 2'-monophosphate, adenosine 3'-monophosphate, and adenosine 2',3'-cyclic

monophosphate are formed during enzymatic or alkaline hydrolysis of RNA. Adenosine 3',5'-cyclic monophosphate (cAMP) is a signaling molecule that accumulates when the cell has a limited supply of nutrients.

first products of the action of alkali on RNA, and these are rapidly hydrolyzed further to yield a mixture of nucleoside 2'- and 3'-monophosphates (Figure 6-7). The sugar component of DNA does not have a 2'-hydroxyl group and is not as easily hydrolyzed in alkaline conditions, making the DNA backbone inherently more stable than that of RNA.

KEY CONVENTION

The structure of a single strand of nucleic acid is always written with the 5' terminus at the left and the 3' terminus at the right—that is, in the 5' → 3' direction. When a double-stranded sequence is shown, the top strand is written in the 5' → 3' direction. The various representations of a nucleotide sequence, using a pentanucleotide as example, are: 5'-ACGTA-3', ACGTA, pA-C-G-T-A_{OH}, pApCpGpTpA, and pACGTA, where *p* denotes a monophosphate, and a subscript OH denotes a 3'-hydroxyl group.

A short nucleic acid containing 50 or fewer nucleotides is generally called an **oligonucleotide**; a longer nucleic acid is a **polynucleotide**.

Nucleotide Bases Affect the Three-Dimensional Structure of Nucleic Acids

Purines and pyrimidines have a variety of chemical properties that affect the structure, and ultimately the function, of nucleic acids. The purine and pyrimidine bases common in DNA and RNA are conjugated ring systems, with alternating single and double bonds between ring atoms (see Figure 6-2). Resonance among atoms in the rings gives most of the bonds a partial double-bond character. One result is that pyrimidines are planar molecules, and purines are very nearly planar, with just a slight pucker. Free pyrimidine and purine bases can exist in two or more forms, called tautomers, depending on pH (Figure 6-8). The structures shown on the left in Figure 6-8 are the tautomers that predominate at physiological pH (pH ~ 7). As a

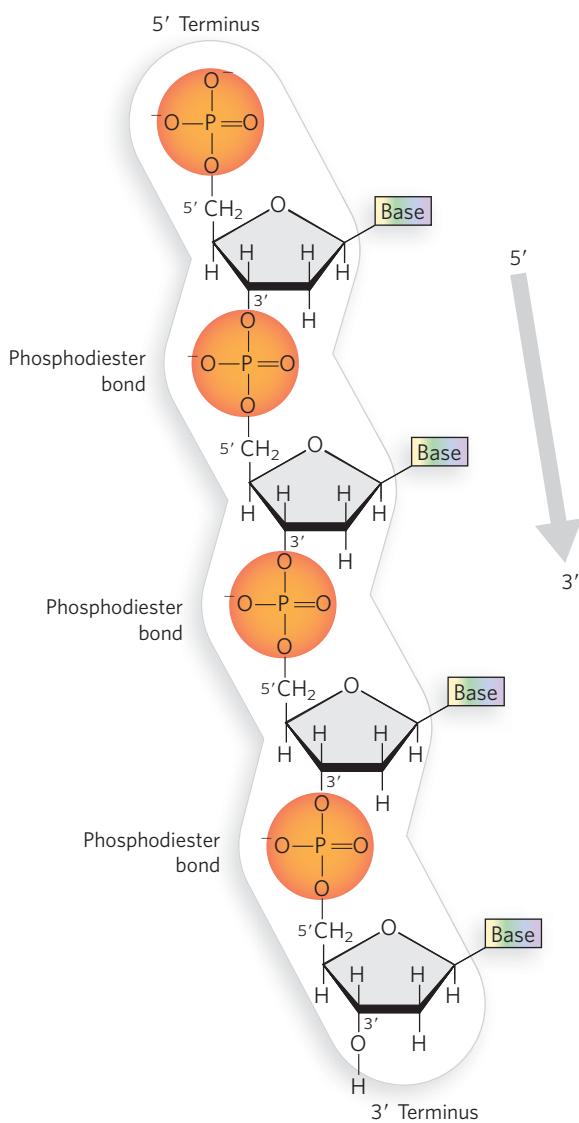


FIGURE 6-6 The phosphodiester linkages in nucleic acids.

Phosphodiester bonds covalently connect the nucleotide units in DNA and RNA. The backbone of alternating sugars and phosphate groups is highly negatively charged. All the phosphodiester linkages in a polynucleotide chain have the same orientation, giving each linear nucleic acid strand a specific polarity and distinct 5' and 3' termini.

result of resonance, delocalized electrons in the conjugated rings are available to absorb ultraviolet (UV) light at wavelengths near 260 nm (Figure 6-9). Ultraviolet light absorbance is used as a method for detecting nucleic acids (see Section 6.4).

The chemical properties of the purines and pyrimidines also give rise to two important modes of interaction between bases in nucleic acids. The first, called

hydrophobic stacking, arises because the bases are hydrophobic and thus relatively insoluble in water at the near-neutral pH of the cell. As a result, the bases align such that two or more are positioned with the planes of their rings parallel, like a stack of coins (Figure 6-10 on page 184). Base stacking helps minimize the contact of the bases with water, and base-stacking interactions are very important in stabilizing the three-dimensional structure of nucleic acids. Such stacking also involves a combination of van der Waals and electrostatic interactions among the bases.

The second important mode of base interaction in nucleic acids is **base pairing**, which results from the hydrogen-bonding capacity of the ring nitrogens, ring carbonyl groups, and exocyclic (i.e., outside the ring structure) amino groups of the pyrimidines and purines. Hydrogen bonds between bases, involving the amino and carbonyl groups, permit a complementary association of two (and occasionally three or four) nucleic acid strands. The most important hydrogen-bonding patterns are those defined by Watson and Crick in 1953, in which A hydrogen-bonds specifically with T (or U), and G with C (Figure 6-11 on page 184). These two types of base pairs predominate in double-stranded DNA and RNA (and thus are considered the canonical base pairs). The purine and pyrimidine tautomers that predominate at physiological pH, shown on the left in Figure 6-8, readily adopt these hydrogen-bonding patterns. It is this specific pairing of bases in the double-stranded DNA helix that permits the duplication of genetic information.

Nucleotides Play Additional Roles in Cells

Nucleotides have functions in cells beyond providing the building blocks for DNA and RNA. The phosphate group covalently linked to the 5' hydroxyl of a nucleoside may have one or two additional phosphates attached. The resulting molecules are referred to as nucleoside mono-, di-, and triphosphates (Figure 6-12 on page 184). Starting from the phosphate closest to the ribose, the three phosphates are generally labeled α , β , and γ . Nucleoside triphosphates are the activated precursors of DNA and RNA synthesis (see Chapters 11 and 15). Furthermore, hydrolysis of nucleoside 5'-triphosphates, primarily ATP, provides the chemical energy to drive a wide variety of cellular reactions (see Chapter 3).

The nucleoside adenosine also forms part of the structure of otherwise unrelated enzyme cofactors that perform a wide range of chemical functions (Figure 6-13 on page 185). For example, nicotinamide adenine dinucleotide (NAD^+) plays a crucial role in

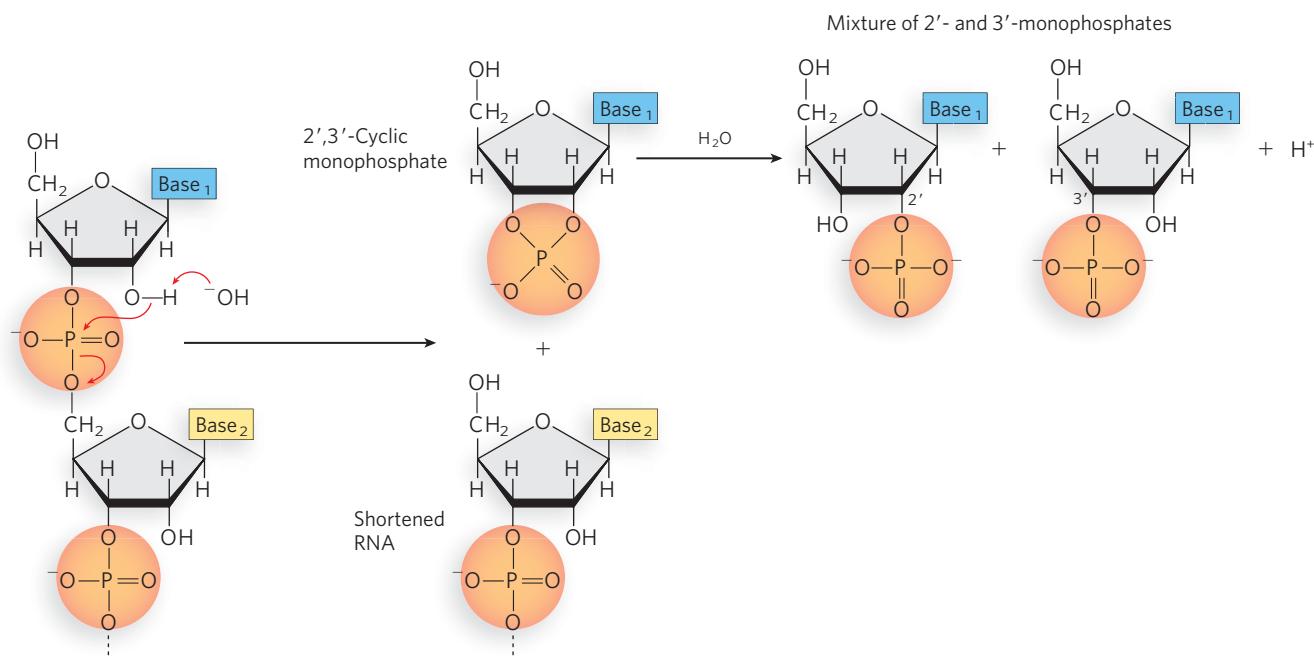


FIGURE 6-7 The hydrolysis of RNA. The 2'-hydroxyl group can be activated as a nucleophile under alkaline ($pH > 7$) conditions or by ribonucleases. The 2',3'-cyclic monophosphate product is further hydrolyzed to a mixture of 2'- and 3'-monophosphates.

cellular energy production in both animal cells and plant cells. A related cofactor, $NADP^+$, contributes to the synthesis of lipids and nucleic acids, and participates in photosynthesis. Flavin adenine dinucleotide (FAD) is the active form of vitamin B₂ (riboflavin), which transfers electrons in some biosynthetic reactions. Enzymatic reactions that involve the transfer of a methyl group from one molecule to another often involve the substrate S-adenosylmethionine (adoMet), which consists of an adenosine linked to a methionine.

In these adenosine-containing compounds, the adenosine portion does not participate directly in the molecule's primary function. Instead, it seems to be a molecular "handle" that allows the cofactor or substrate to bind tightly in an enzyme active site. Adenosine may have taken on this role partly because of its abundance in the environment of the early Earth (see Chapter 1, How We Know). The importance of adenosine probably lies not so much in some special chemical characteristic as in the evolutionary advantage of using one compound for multiple purposes. Once ATP became the universal source of chemical energy, biological systems developed to synthesize ATP in greater abundance

than the other nucleotides; because adenosine was abundant, it became the logical choice for incorporation into a wide variety of structures. This economy also extends to protein structure. For example, the Rossmann fold, a protein domain that binds adenosine (see Figure 4-15), is found in many enzymes that bind ATP and enzyme cofactors.

Some nucleotides function as regulatory molecules. One of the most common is **adenosine 3',5'-cyclic monophosphate (cyclic AMP, or cAMP)** (see Figure 6-5), formed from ATP in a reaction catalyzed by adenylyl cyclase—an enzyme whose activity is closely linked to the metabolic state of the cell. Cyclic AMP performs regulatory functions in virtually every cell outside the plant kingdom. **Guanosine 3',5'-cyclic monophosphate (cGMP)** occurs in many cells and also has regulatory functions.

Both cAMP and cGMP are called **second messengers**, because they are produced or degraded in response to the interaction of extracellular chemical signals ("first messengers") with receptors on the cell surface. The second messengers induce adaptive changes in the cell interior. In this way, cells can respond quickly to environmental changes by taking cues

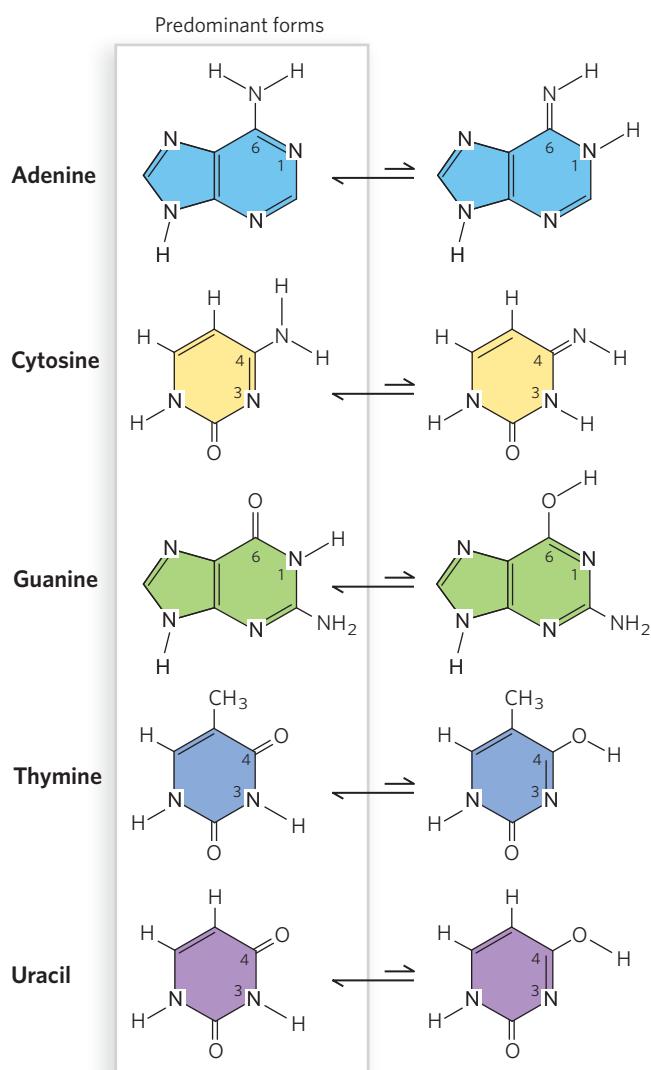


FIGURE 6-8 Tautomers of pyrimidine and purine bases.

Each purine or pyrimidine can exist as one of several isomers that differ in the placement of a hydrogen atom and a double bond (tautomers). Shown here are two tautomers for each of the common bases found in nucleic acids. The predominant tautomer of each base at physiological pH is on the left. The predominant tautomeric forms are found in DNA and RNA and participate in Watson-Crick base pairing (see Figure 6-11).

from hormones or other external chemical signals. For example, light entering the photoreceptors of the human eye activates an enzyme that degrades cGMP, causing sodium channels in the photoreceptor cell membrane to close and thereby triggering visual information to be sent to the brain. Another regulatory nucleotide, guanosine tetraphosphate (ppGpp), is produced in bacteria in response to a slowdown in protein synthesis during amino acid starvation. This nucleotide inhibits the synthesis of the rRNA and tRNA molecules needed for protein synthesis, preventing the unnecessary production of nucleic acids.

SECTION 6.1 SUMMARY

- A nucleotide consists of a nitrogenous base (a purine or a pyrimidine), a pentose sugar, and one or more phosphate groups. Nucleic acids are polymers of nucleotide units, joined by phosphodiester linkages between the 3'-phosphate group of one unit and the 5'-hydroxyl group of the next. Polynucleotides have a directionality defined by a 5' terminus and a 3' terminus.

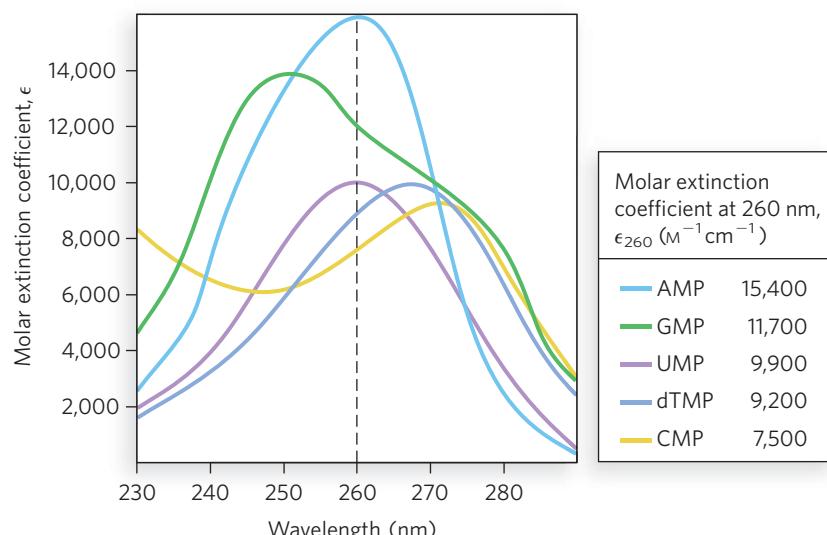


FIGURE 6-9 Absorption spectra of the common nucleotides.

The plots show molar extinction coefficients at pH 7.0 as a function of wavelength for the nucleoside 5'-monophosphates. The molar extinction coefficient, ϵ (units $\text{M}^{-1}\text{cm}^{-1}$), measures the amount of light absorbed by a 1 M solution with a light path length of 1 cm. The table shows the molar extinction coefficients at 260 nm for the plotted nucleotides.

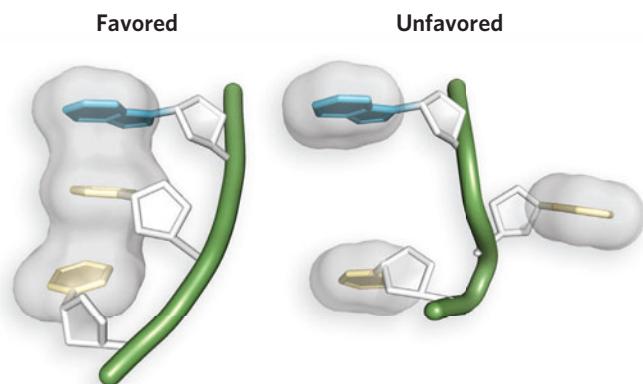


FIGURE 6-10 Base stacking in nucleic acids. Hydrophobic, van der Waals, and electrostatic interactions favor the alignment of bases in an aqueous solution or within a polynucleotide chain (three nucleotides in an RNA chain are shown here); the unstacked orientation is disfavored. Van der Waals radii are shown in gray.

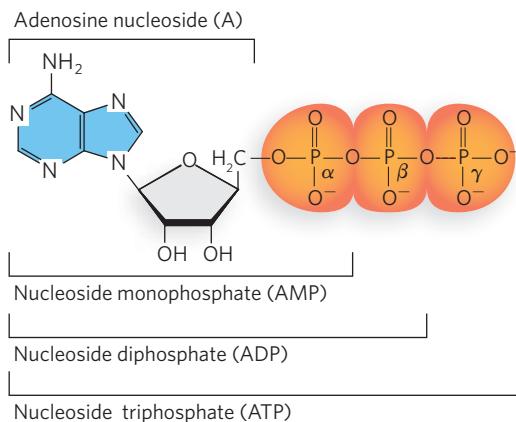


FIGURE 6-12 Nomenclature for nucleotides. The phosphate group covalently linked to the 5' hydroxyl of a nucleoside may have one or two additional phosphates attached; the resulting molecules are referred to as nucleoside mono-, di-, and triphosphates. Starting from the phosphate closest to the ribose, the three phosphates are designated α , β , and γ .

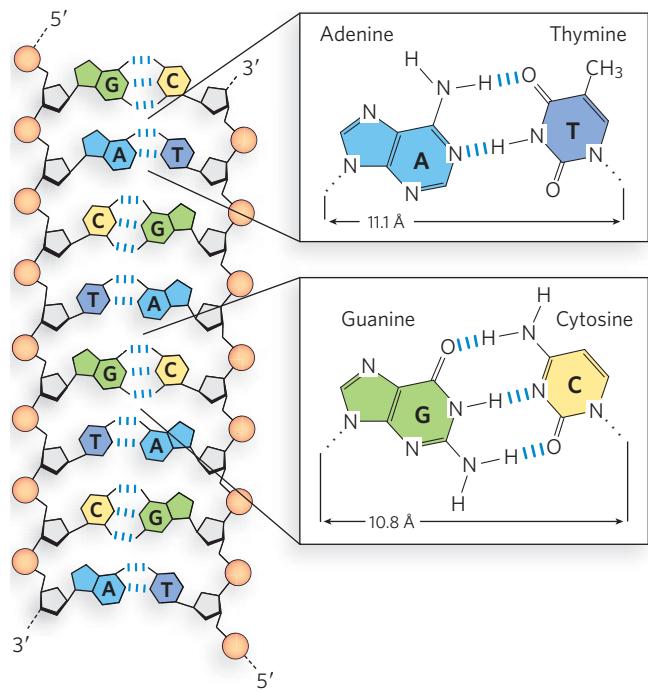


FIGURE 6-11 Hydrogen-bonding patterns in Watson-Crick base pairs. Hydrogen bonds are represented by three blue lines.

- DNA and RNA are two types of nucleic acids. The nucleotides in DNA contain 2'-deoxy-D-ribose, whereas the nucleotides in RNA contain D-ribose. The hydroxyl group at the 2' position in D-ribose makes the RNA backbone more susceptible to hydrolysis than DNA.

- Both DNA and RNA contain four different bases, two purines and two pyrimidines. The purines in DNA and RNA are the same: adenine and guanine. DNA contains the pyrimidines cytosine and thymine, and RNA contains the pyrimidines cytosine and uracil.
- In addition to A, C, G, T, and U, numerous minor bases occur in nature, often differing from the canonical bases by the presence of a functional group at a particular position on the base; these bases can play central roles in nucleic acid structure and biochemical function.
- The chemical properties of the nitrogenous bases affect nucleotide and nucleic acid structure. As a result of resonance, the bases in a nucleotide chain are planar and tend to stack. The hydrogen-bonding capabilities of the conjugated rings allow the formation of specific base-pair interactions: A pairs with T (or U) and C pairs with G.
- Adenosine is a building block for some important enzyme cofactors, such as nicotine adenine dinucleotide (NAD^+) and flavin adenine dinucleotide (FAD). The presence of an adenosine component in a variety of cofactors enables recognition by enzymes that share common structural features.
- Cyclic AMP, formed from ATP in a reaction catalyzed by adenylyl cyclase, is a common second messenger produced in response to hormones and other chemical signals.

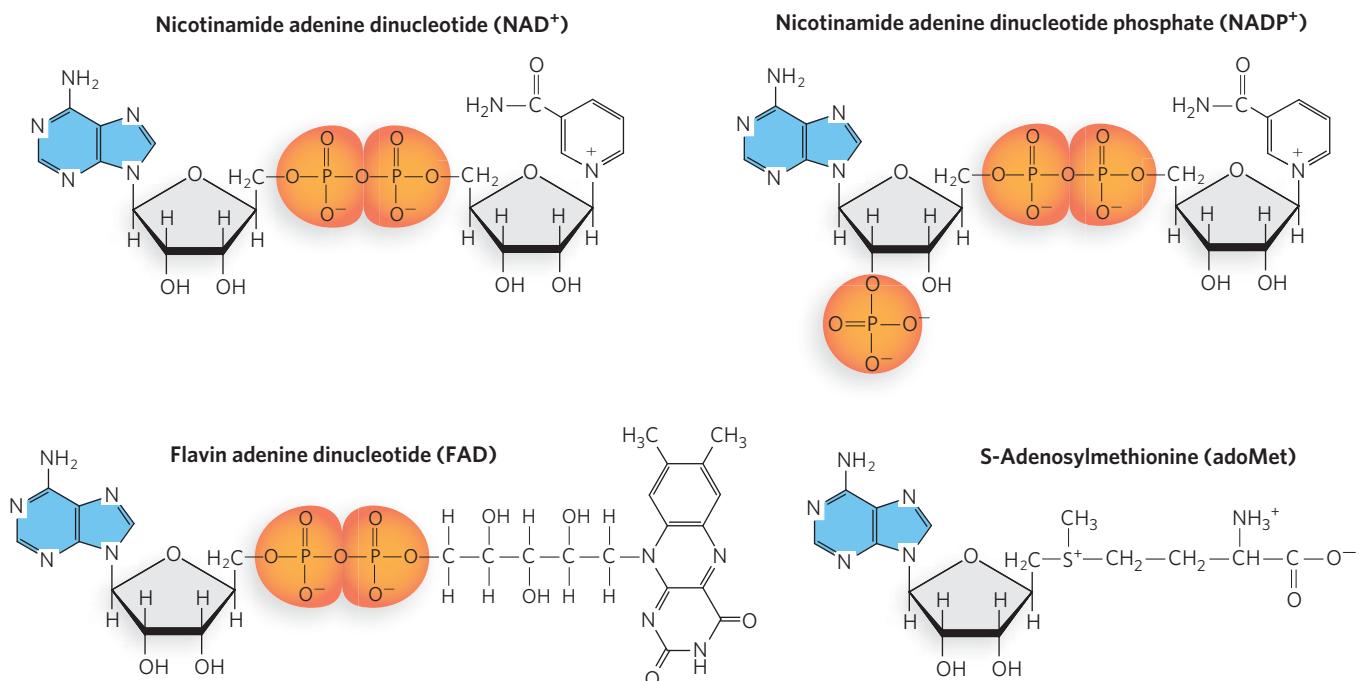


FIGURE 6-13 Some cofactors and a substrate containing adenosine. The adenine base (blue) and the ribose attached

to it form the adenosine portion of each molecule. See text for details.

6.2 DNA Structure

In Chapter 2 we discussed the experiments that revealed DNA as the genetic material in cells. Recognizing that DNA is the primary carrier of biological information in cells and viruses, researchers were motivated to determine its molecular structure. As we'll see, what they learned about DNA's structure explained how it functions as the key molecule of inheritance, thereby paving the way for many investigations into the mechanisms of DNA replication and metabolism.

DNA Molecules Have Distinctive Base Compositions

In the 1940s, Erwin Chargaff and his colleagues made an important discovery that provided clues to the structure of DNA. Using DNA samples isolated from many different organisms, they observed that the four bases of DNA occur in different ratios that are characteristic of each species. They also observed that for specific pairs of bases, the amounts of each base are closely related. Their data showed that:

1. The base composition of DNA generally varies from one species to another.
2. DNA specimens isolated from different tissues of the same species have the same base composition.

3. The base composition of DNA in a given species does not change with an organism's age, nutritional state, or environment.

4. In all cellular DNAs, regardless of the species, the number of adenine residues equals the number of thymidine residues (i.e., $A = T$), and the number of guanosine residues equals the number of cytidine residues ($G = C$). From these relationships it follows that the sum of the purine residues equals the sum of the pyrimidine residues: $A + G = T + C$.

Referred to as **Chargaff's rules**, these quantitative relationships were confirmed by many subsequent researchers. Not only were these findings a key to establishing the three-dimensional structure of DNA, they also yielded clues about how genetic information is encoded in DNA and passed along from one generation to the next.

DNA Is Usually a Right-Handed Double Helix

Chargaff's discoveries imposed important constraints on possible models for the structure of DNA. At the same time, Rosalind Franklin and Maurice Wilkins were using the powerful method of x-ray diffraction to analyze DNA fibers (see How We Know). They showed that DNA produces a characteristic x-ray diffraction pattern. From this

pattern, Watson and Crick deduced that DNA molecules are helical, with two periodicities along their long axis: a primary one of 3.4 Å and a secondary one of 34 Å. The challenge then was to formulate a three-dimensional model of the DNA molecule that could account not only for the x-ray diffraction data but also for the specific A = T and G = C base equivalences discovered by Chargaff.

Using the x-ray diffraction data obtained by Franklin and Wilkins, Watson and Crick proposed that DNA is composed of two polynucleotide strands entwined in the form of a right-handed double helix (Figure 6-14). Alternating 2'-deoxy-D-ribose and phosphate units make up the backbone of each strand, from which the bases project inward, toward the center of the helix. The bases are thus positioned to form hydrogen-bonding interactions with each other according to the preferred pairings of A with T and C with G. The two unequal surfaces formed by the twist of the helix are called the **major groove** and the **minor groove**. DNA strands always have a defined directionality, or polarity, due to the asymmetric shape and chemical linkage of the component nucleotides. In the double helix, the two strands have opposite directionality and the helix is said to be **antiparallel**. In chemical terms, this means that one strand runs in the 5' → 3' direction and the other runs in the 3' → 5' direction. The antiparallel orientation of the DNA strands is more energetically favorable than the parallel configuration, due to the geometry of the component bases. Furthermore, the DNA double helix almost always twists in a right-handed direction (see Figure 4-6b). Rarely, left-handed helices are observed, in which the twist is to the left. By convention, helices are assumed to be right-handed unless otherwise specified.

Watson and Crick's double-helical model of DNA makes chemical sense, accounting for the known

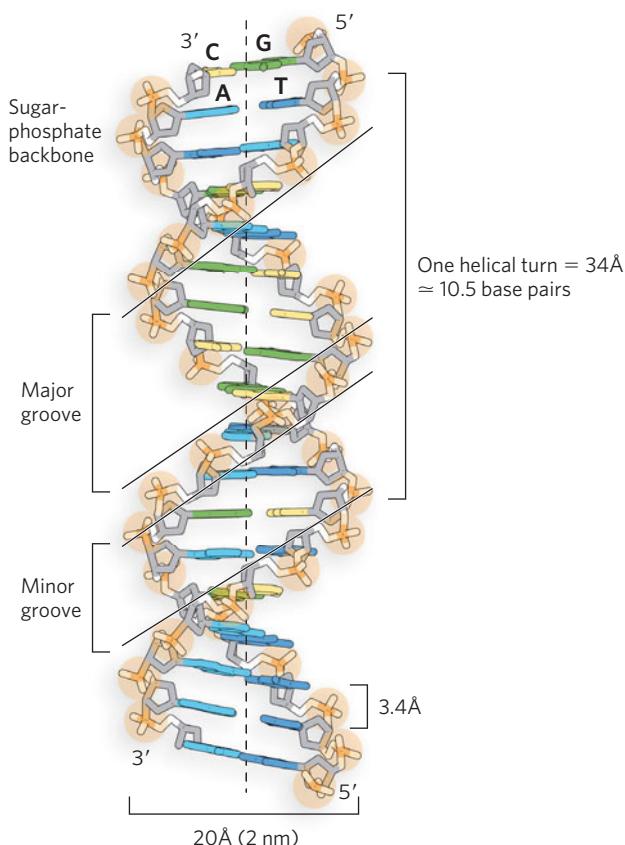


FIGURE 6-14 The double-helical structure of DNA. The original model proposed by Watson and Crick had 10 base pairs and a length of 34 Å per turn of the helix. Later measurements of DNA in solution (as opposed to in a crystal or fiber) showed 10.5 base pairs per helical turn. The major and minor grooves, where most interactions with proteins or other nucleic acids occur, are shown. See text for details.



Rosalind Franklin, 1920–1958

[Source: © National Portrait Gallery, London.]



Maurice Wilkins,

1916–2004 [Source: Associated Press.]

properties of the component nucleotides. The hydrophilic backbones of alternating sugar and phosphate groups are on the outside of the helix, facing the surrounding water. The pentose ring of each deoxyribose is in the C-2' endo conformation (see Figure 6-3), and this sugar pucker defines the distance between adjacent phosphate groups in the DNA backbone. The purine and pyrimidine bases of both strands are stacked inside the double helix, with their hydrophobic and nearly planar ring structures very close together and relatively perpendicular to the long axis. Each nucleotide base of one strand is paired in the same plane with a base from the other strand. Watson and Crick found that the hydrogen-bonded base pairs, G with C and A with T, are those that agreed best with the x-ray diffraction data, providing a rationale for Chargaff's rule that in any DNA, G = C and A = T. It is important to note that three hydrogen bonds can form between G and C, symbolized G≡C, but only two can form between A and T,

symbolized A=T. By always pairing a purine (A or G) with a pyrimidine (T or C), consistent spacing is maintained between the two antiparallel DNA backbones, giving a regular, uniform shape to the double helix. This has significant consequences for the stability of any double-stranded DNA sequence (see Section 6.4).

The double-helical structure of DNA also explains the periodicities observed in the x-ray diffraction patterns of DNA fibers. The vertically stacked bases inside the double helix are 3.4 Å apart; the secondary repeat distance of about 34 Å is accounted for by the presence of 10 base pairs in each complete turn of the double helix. In aqueous solution the structure differs slightly from that in fibers, having 10.5 base pairs per helical turn.

The stability of the DNA double helix arises primarily from the hydrophobic base-stacking interactions, which are largely nonspecific with respect to sequence. The configuration of planar purine-pyrimidine base pairs at the center of the helix allows their flat surfaces to stack on top of each other through shared electrons (see Figure 6-10). This energetically favorable situation stabilizes the double helix relative to single-stranded DNA by minimizing contact of the hydrophobic purines and pyrimidines with water. Furthermore, extensive networks of weak bonds in double-stranded DNA, such as van der Waals interactions and hydrogen bonds, are arranged so that for most of these bonds, they cannot break without simultaneously breaking many others. Consequently, DNA double helices 10 or more base pairs in length are stable at room temperature. In fact, DNA can persist in fossil samples over long periods of time, making possible the sequencing of DNA samples from long-extinct species—including Neandertal hominids and woolly mammoths!

By far the most significant property of the double helix as an information carrier is the hydrogen bonding between the bases. Because adenine is always hydrogen-bonded to thymine, and guanine is always hydrogen-bonded to cytosine, exact copies of encoded information can be replicated. This specific base pairing gives the two helical strands a complementary relationship in which the base sequence of one strand defines the sequence of its partner. For example, the sequence 5'-GTAACGC-3' on one strand specifies the complementary sequence 5'-GCGTTAC-3' on the other strand.

Thus, the discovery of the DNA double helix immediately suggested a mechanism for the transmission of genetic information. As Watson and Crick proposed, this structure could logically be reproduced by separating the two strands and synthesizing a complementary strand for each. Because nucleotides in each new strand are joined in a sequence specified by the base-pairing

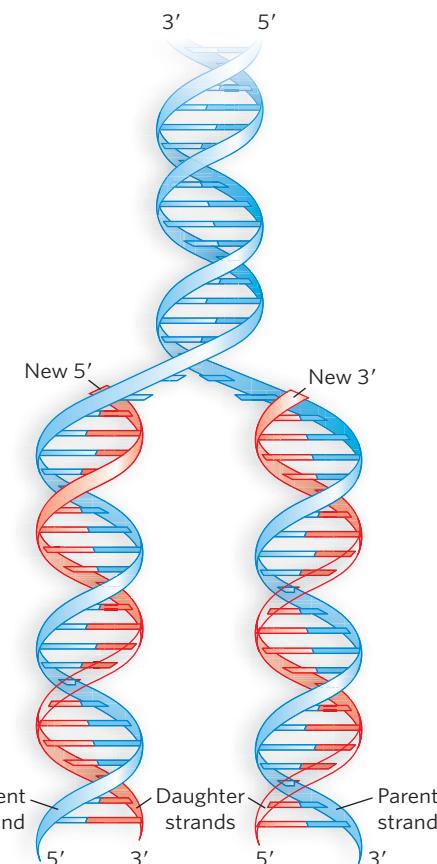


FIGURE 6-15 The mechanism for DNA replication. The newly synthesized complementary strands (“daughter” strands) are shown in red.

rules stated earlier, each preexisting strand functions as a template to guide the synthesis of a complementary strand (Figure 6-15). These expectations were experimentally confirmed, inaugurating a revolution in our understanding of biological inheritance. (DNA replication is discussed in detail in Chapter 11.) The accuracy of base pairing can also be used for computation, raising the possibility of future computers based on DNA (Highlight 6-1).

DNA Adopts Different Helical Forms

Nucleic acids are inherently flexible molecules. Numerous bonds in the sugar-phosphate backbone can rotate, and thermal fluctuation can lead to bending, stretching, and unpairing of the two strands. As a result, cellular DNA contains significant deviations from the Watson-Crick DNA structure, some or all of which may play important roles in DNA metabolism. Generally, such structural variations do not affect the key properties of strand complementarity: antiparallel strands and the requirement for A=T and G≡C base pairs.

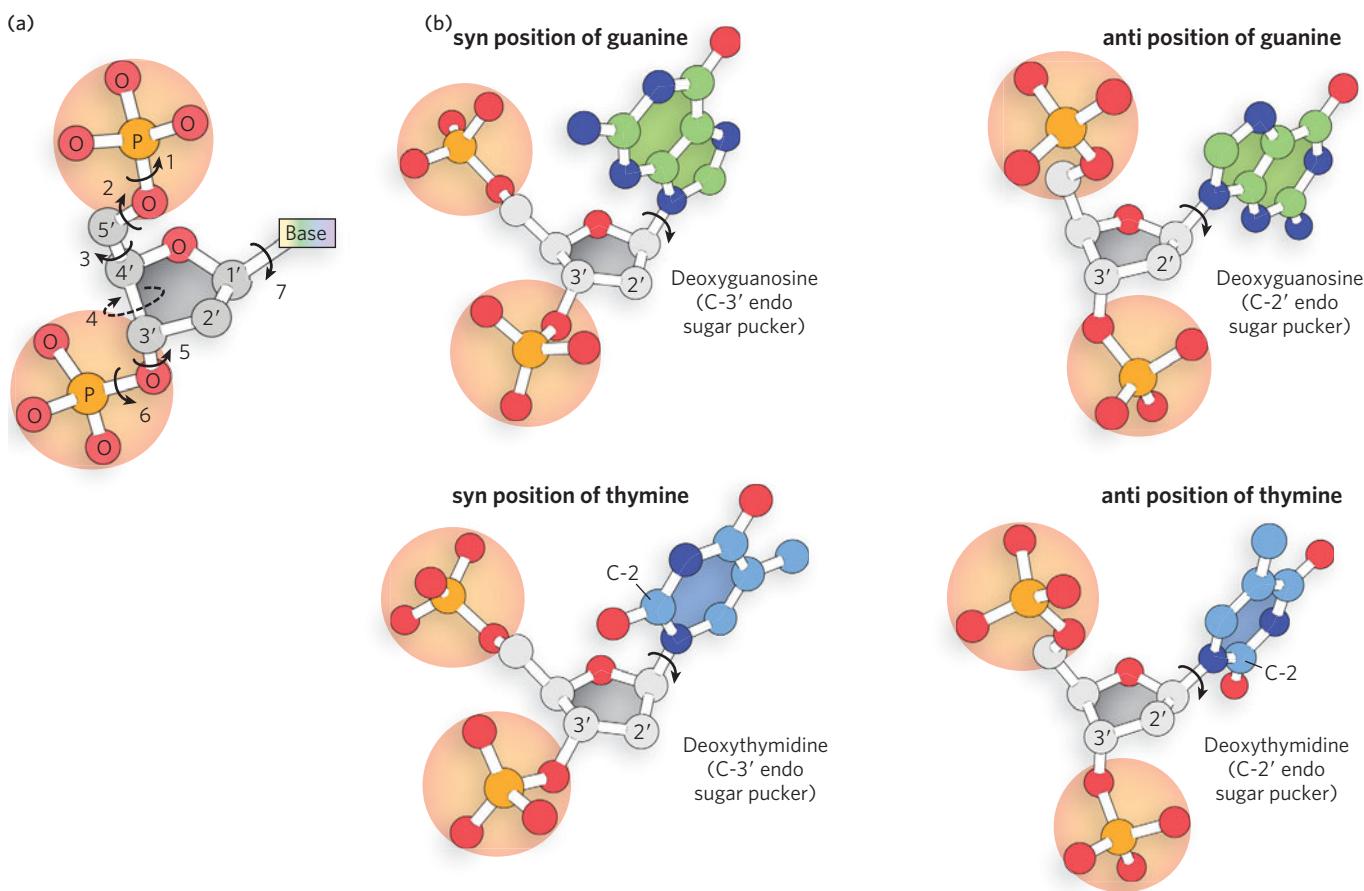


FIGURE 6-16 Factors contributing to structural variation in DNA. (a) DNA nucleotide conformation is affected by rotation about seven different bonds. Six of the bonds rotate freely; rotation about bond 4 is constrained by the sugar ring, giving rise to the sugar pucker. (b) The syn and anti positions

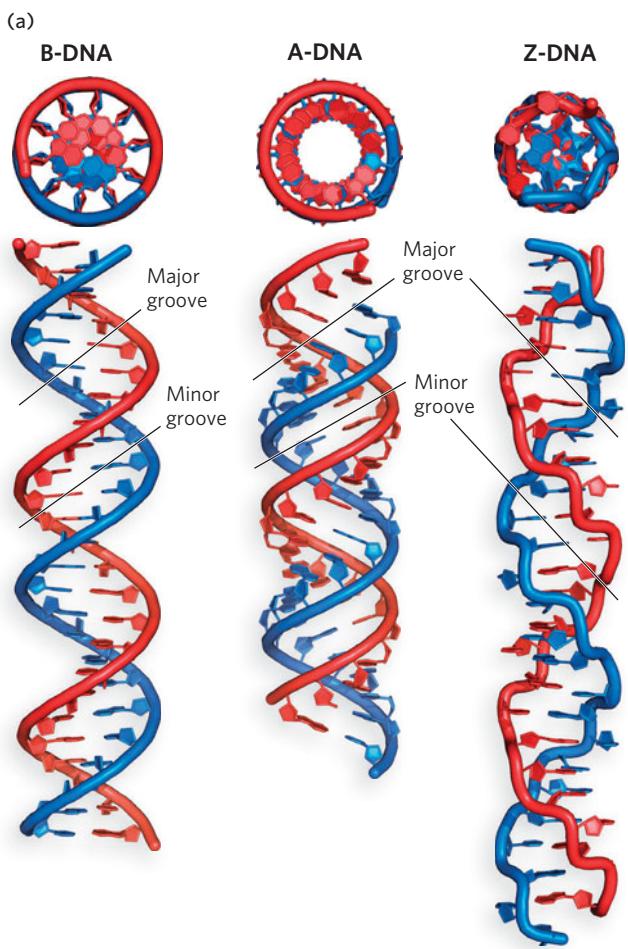
of deoxyguanosine, and the syn and anti positions of deoxythymidine. Note that pyrimidines are restricted to the anti position; the carbonyl at C-2 causes steric clash in the syn conformation.

Variation in the three-dimensional structure of DNA reflects three things: the different possible conformations of the deoxyribose (see Figure 6-3), rotation about the contiguous bonds that make up the sugar-phosphate backbone (Figure 6-16a), and free rotation about the glycosidic bond. Because of steric constraints, purines in purine nucleotides are restricted to two stable conformations with respect to deoxyribose, called syn and anti (Figure 6-16b). Pyrimidines are generally restricted to the anti conformation, because of steric interference between the sugar and the carbonyl oxygen at C-2 of the pyrimidine.

The Watson-Crick structure is known as **B-form DNA**, or **B-DNA**. As the most stable structure for a random-sequence DNA molecule under physiological conditions, B-DNA is the standard point of reference in any study of the properties of DNA. Structures of short B-DNA duplexes have been studied in depth, revealing many details about the double helix (see How We

Know). Two structural variants well characterized by x-ray crystallography are **A-form DNA (A-DNA)** and **Z-form DNA (Z-DNA)**. These three DNA conformations are shown in Figure 6-17, with a summary of their properties.

A-DNA is favored in many solutions that are relatively devoid of water. In this case, the DNA is still arranged in a right-handed double helix, but the helix is wider and the number of base pairs per helical turn is 11, rather than 10.5 as in B-DNA. Whereas base pairs in B-DNA tilt slightly in a negative direction—that is, below the plane—with respect to a plane perpendicular to the helical axis, the base pairs in A-DNA are tilted above the plane by about +20°. In addition, the distance between adjacent phosphates in the polynucleotide chain, a direct consequence of the sugar pucker, changes from 7 Å in B-form helices to 5.9 Å in A-form helices (Figure 6-18 on page 190). These structural changes deepen the major groove while making the minor groove



(b)

	B-DNA	A-DNA	Z-DNA
Helix sense	Right-handed	Right-handed	Left-handed
Diameter	~20 Å	~26 Å	~18 Å
Base pairs per helical turn	10.5	11	12
Helix rise per base pair	3.4 Å	2.6 Å	3.7 Å
Base tilt in relation to the helix axis	-6°	+20°	-7°
Sugar pucker conformation	C-2' endo	C-3' endo	C-2' endo for pyrimidines; C-3' endo for purines
Glycosyl bond conformation	Anti	Anti	Anti for pyrimidines; syn for purines

FIGURE 6-17 A comparison of the B, A, and Z forms of DNA.

(a) In each case, the sugar-phosphate backbones wind around the exterior of the helix (red and blue), with the bases pointing inward. The same 25-base-pair DNA sequence is shown in all three forms. Differences in helical diameter can be seen in end-on views (top); differences in helical rise and groove shape are apparent in the side views (bottom). B-DNA, the most common form in cells, has a wide major groove and a narrow minor groove. A-form helices, common for RNA and certain DNA structures, are more compact than B-DNA. The major groove is deeper and the minor groove is shallower than in B-DNA. Z-DNA, which forms only under high salt conditions or with C≡G-rich DNA sequences, is left-handed, and its backbone has a zigzag pattern. It is less compact than B-DNA, with a very shallow major groove and a narrow and deep minor groove. (b) The table summarizes some properties of the three forms of DNA.

shallower. The reagents used to promote crystallization of DNA tend to dehydrate it, and thus most short DNA molecules tend to crystallize in the A form.

Z-DNA is a more radical departure from B-DNA; the most obvious distinction is the left-handed helical rotation. There are 12 base pairs per helical turn, and the structure appears more slender and elongated. The DNA backbone takes on a zigzag appearance (hence the Z designation). Certain nucleotide sequences fold into left-handed Z helices much more readily than others. Prominent examples are sequences in which pyrimidines alternate with purines, especially alternating C and G residues or 5-methyl-C and G residues

(methylated bases are discussed in Section 6.4). To form the left-handed helix in Z-DNA, the purine residues flip to the syn conformation, alternating with pyrimidines in the anti conformation. The major groove is barely apparent in Z-DNA, and the minor groove is narrow and deep.

Whether A-DNA occurs in cells is uncertain, but there is evidence for some short stretches of Z-DNA in the chromosomes of both bacteria and eukaryotes. The evidence comes in part from experimentally prepared antibodies against short Z-form DNA segments, which can selectively bind to sequences in chromosomal DNA. These potential Z-DNA tracts correlate with

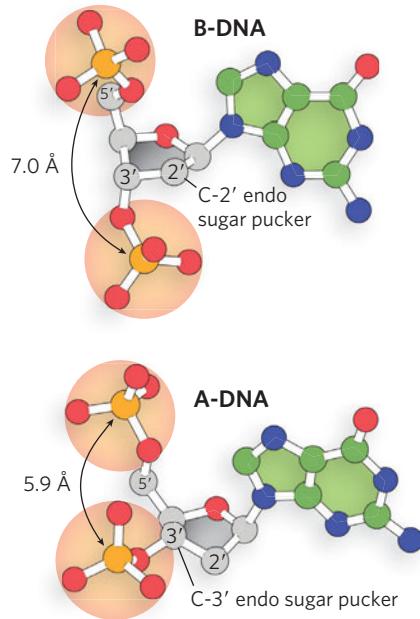


FIGURE 6-18 The effect of sugar pucker on the distance between phosphates in nucleic acids. The configuration of the pentose ring, known as the sugar pucker, is different in the backbones of B-DNA and A-DNA. As a result, the phosphate groups attached to the 5' and 3' positions of each nucleotide are different distances apart in the two forms, giving rise to distinct helical geometries.

actively transcribed regions of the genome and could play a role (as yet undefined) in genetic recombination or in the regulation of gene expression.

Certain DNA Sequences Adopt Unusual Structures

Other sequence-specific DNA structures have been detected, within larger chromosomes, that may affect the function and metabolism of the DNA segments in their immediate vicinity. For example, certain repetitive sequences can bend the DNA helix in a distinct way. The stability and geometry of base-pair stacking influences the preferred direction of DNA bending. This was first observed in repetitive-sequence DNA isolated from trypanosomes, the protozoa that cause African sleeping sickness. Typical sequences that cause pronounced bending contain stretches of four to six A and T residues separated by C- and G-rich segments. The A- or T-tracts, each corresponding to a half-turn of the double helix, are spaced such that the geometry of the repetitive base pairs tends to curve the DNA helix in one direction (Figure 6-19). This DNA bending helps certain proteins—such as transcription factors, which promote

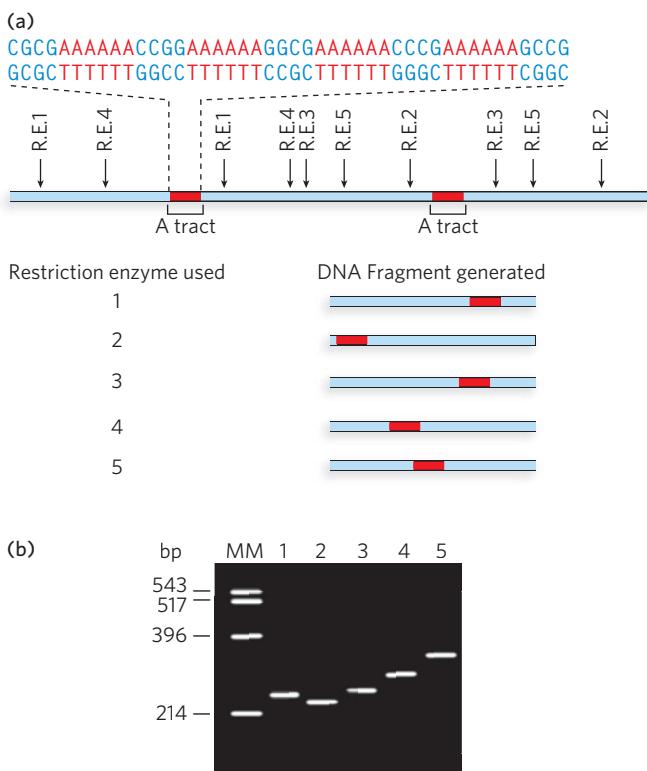


FIGURE 6-19 DNA bending at A-tracts. Bending in DNA segments containing A-tracts of six adenines in a row can be detected by using enzymes to cut the DNA at specific sites. (a) The DNA sequence used in this experiment contains two A-tracts. Arrows point to sites that can be cleaved by various restriction enzymes (R.E. 1–R.E. 5) to generate DNA fragments of equal length that contain an A-tract at the end or in the middle of the fragment. (b) DNA fragments are analyzed by gel electrophoresis. Even though all DNA fragments are the same size (~215 base pairs), their rate of migration through the gel depends on the relative location of the A-tract. When the A-tract is located in the middle, the DNA fragment is more bent and migrates slowly; when the A-tract is at the end, the fragment is less bent and migrates faster. The lane on the left, MM, contains molecular markers that provide a reference of DNA fragment size in base pairs (bp). [Source: (b) Adapted from Asayama et al., *J. Biochem.* (Tokyo) 125:460–468, 1999.]

the synthesis of mRNAs—bind to their target DNA binding sites.

Regions of DNA where the two complementary strands have the same sequence when read in the 5' → 3' or the 3' → 5' direction occur relatively frequently in chromosomal DNA and are called **palindromes**. In language, a palindrome is a word, phrase, or sentence that is spelled identically when read either forward or backward; two examples are ROTATOR and NURSES RUN. In biology, the term applies to double-stranded regions of DNA where one strand's sequence is identical

HIGHLIGHT 6-1 TECHNOLOGY

DNA Computing

The ability of DNA to form a variety of helical structures, at least under controlled laboratory conditions, has led many scientists to wonder whether DNA could function in ways other than carrying genetic information. Although there is no evidence for this in biological systems, DNA is proving to be a versatile material for making very small molecular structures,

enzymes, and even computers. Though DNA computers aren't on the shelves of your local electronics shop yet, the technology is under development.

The concept of DNA computing originated in 1994, when Leonard Adleman, a computer scientist at the University of Southern California, came up with the idea after reading James Watson's



Leonard Adleman [Source: Courtesy of University of Southern California.]

textbook *The Molecular Biology of the Gene* (first published in 1965). Adleman realized that in the way it stores genetic information, DNA is similar to a computer hard drive. As reported in *Science* magazine in 1994, Adleman used DNA to solve the well-known mathematical problem of finding the shortest distance for a salesman traveling between seven cities and going through each city only once. Using a different, unique DNA sequence to represent each city and possible travel route, Adleman mixed the DNA strands together in a test tube and allowed them to anneal. Each double-stranded fragment, which forms in a few seconds, represented a possible answer to the problem. DNA fragments containing mismatched base pairs were eliminated with enzymes, leaving behind only the travel routes connecting all seven cities.

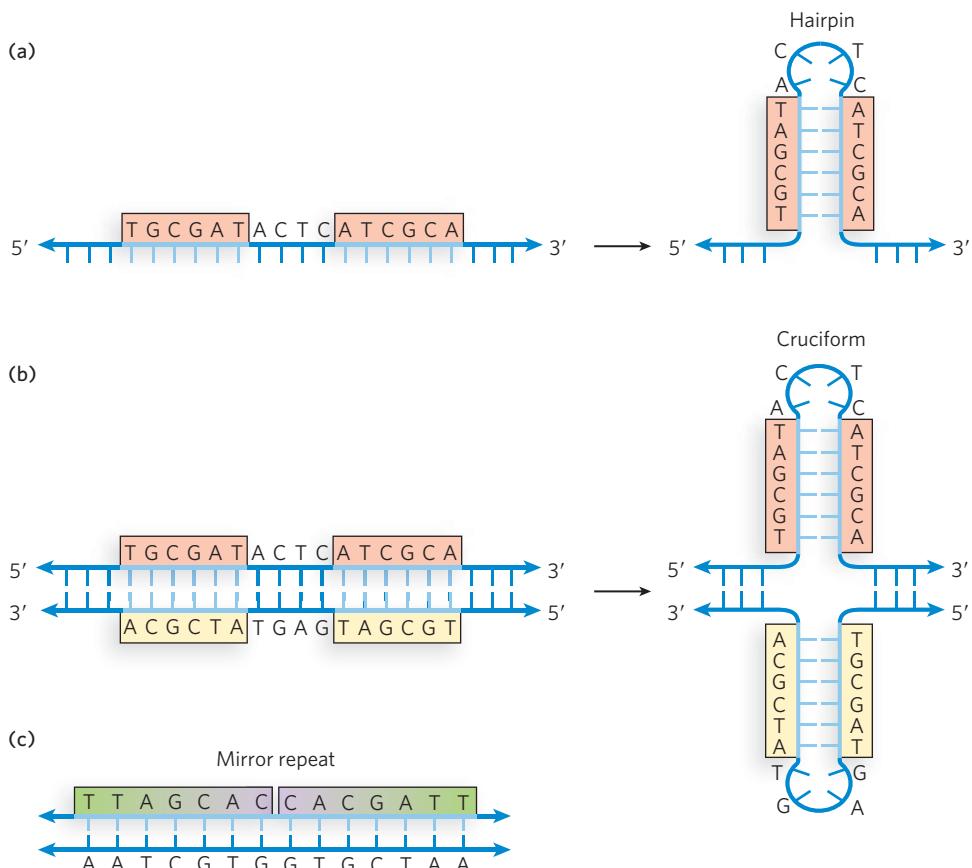
Based on this modest but elegant initial challenge, DNA molecules are now being used to solve more complex computing problems, leading the way toward diverse biotechnology and security applications, including detecting and tracing minute amounts of toxic or dangerous chemicals.

to its complement; for example, 5'-GAATTC-3' is a palindrome because its complementary sequence is also 5'-GAATTC-3'. Palindromes are formed from adjacent **inverted repeats**, which can occur within one strand of DNA or over the two strands of the double helix (Figure 6-20a, b). A related arrangement is a **mirror repeat**, in which the inverted repeat sequence is non-palindromic (Figure 6-20c).

These sequences play important biological roles, such as slowing or blocking protein synthesis by the ribosome—a process called translation attenuation (see Chapter 18)—or forming recognition sites for restriction enzymes, which catalyze double-stranded DNA cleavage (see Chapter 7). Inverted repeats over two DNA strands are self-complementary within each strand and therefore have the potential to form **hairpin** or **cruciform** (cross-shaped) structures (see Figure 6-20a, b). Such sequences are found in virtually every large DNA molecule and can encompass a few base pairs or thousands. The extent to which inverted repeats occur as cruciforms in cells is not known, although some cruciform structures have been demonstrated *in vivo* in *Escherichia coli*. These structures form transiently during recombination of DNA molecules, in which genetic

information on two chromosomes is exchanged during cell division. Self-complementary sequences cause isolated single strands of DNA (or RNA) in solution to fold into complex structures containing multiple hairpins.

Several unusual DNA structures involve three or even four DNA strands. Although rare, these structural variants merit investigation because there is a tendency for them to form at sites where important events in DNA metabolism (replication, recombination, transcription) are initiated or regulated. Nucleotides participating in Watson-Crick base pairing have the potential to form additional hydrogen bonds, particularly with functional groups arrayed in the major groove. For example, a thymidine can pair with the adenine of an A=T nucleotide pair, and a cytidine (if protonated) can pair with the guanosine of a G≡C pair, forming “base triples” (Figure 6-21a on page 193). The N-7, O⁶, and N⁶ of purines, the atoms that participate in this additional hydrogen bonding, are often referred to as **Hoogsteen positions**, and the non-Watson-Crick pairing is called **Hoogsteen pairing**. Karst Hoogsteen, in 1963, was the first to recognize the potential for these unusual pairings. Hoogsteen pairing allows the formation of **triplex DNAs** (Figure 6-21b).

**FIGURE 6-20** Palindromes and cruciforms in DNA.

(a) Inverted repeats within a single strand of DNA can be converted into a hairpin, which is double-stranded in the stem region. (b) Inverted repeats within a double-

stranded DNA sequence can form a cruciform (double-hairpin) structure. (c) DNA can also contain mirror repeats, such as TTAGCACCACGATT, which are not palindromic.

The triplexes form most readily within long sequences containing only pyrimidines or only purines in a given strand. Some triplex DNAs contain two pyrimidine strands and one purine strand; others contain two purine strands and one pyrimidine strand. DNA triplex formation can be highly sequence-specific. For example, triplex formation between a small section of chromosomal DNA and a chemically modified single-stranded DNA enabled Peter Dervan and his colleagues to cleave a human chromosome at a single site. They did this by synthesizing short oligonucleotides that could form triplex interactions with various segments of the chromosomal DNA. Because each oligonucleotide had a chemical modification at one end that triggered DNA strand scission, the researchers were able to fragment the DNA at specific sites—which helped in mapping the location of the gene for the inherited neurological disorder Huntington disease.

Four DNA strands can also associate to form a tetraplex (or quadruplex), but this occurs readily only

for DNA sequences with a very high proportion of G residues (Figure 6-21c). The guanosine tetraplex, or **G tetraplex**, is quite stable over a wide range of conditions. The DNA regions at the ends of linear chromosomes, called telomeres, typically consist of G-rich segments that have a propensity to form tetraplex structures when tested in the laboratory. Whether such structures contribute to the stability and recognition of telomeres *in vivo* is not known.

In the DNA of living cells, sites recognized by many sequence-specific DNA-binding proteins are arranged as palindromes, and polypyrimidine or polypurine sequences that can form triple helices are found within regions involved in regulating the expression of some eukaryotic genes. In principle, synthetic DNA strands designed to pair with these sequences to form triplex DNA could disrupt gene expression. This approach to controlling cellular metabolism is of growing commercial interest for its potential application in medicine and agriculture. Unusual DNA structures can also be engineered, raising

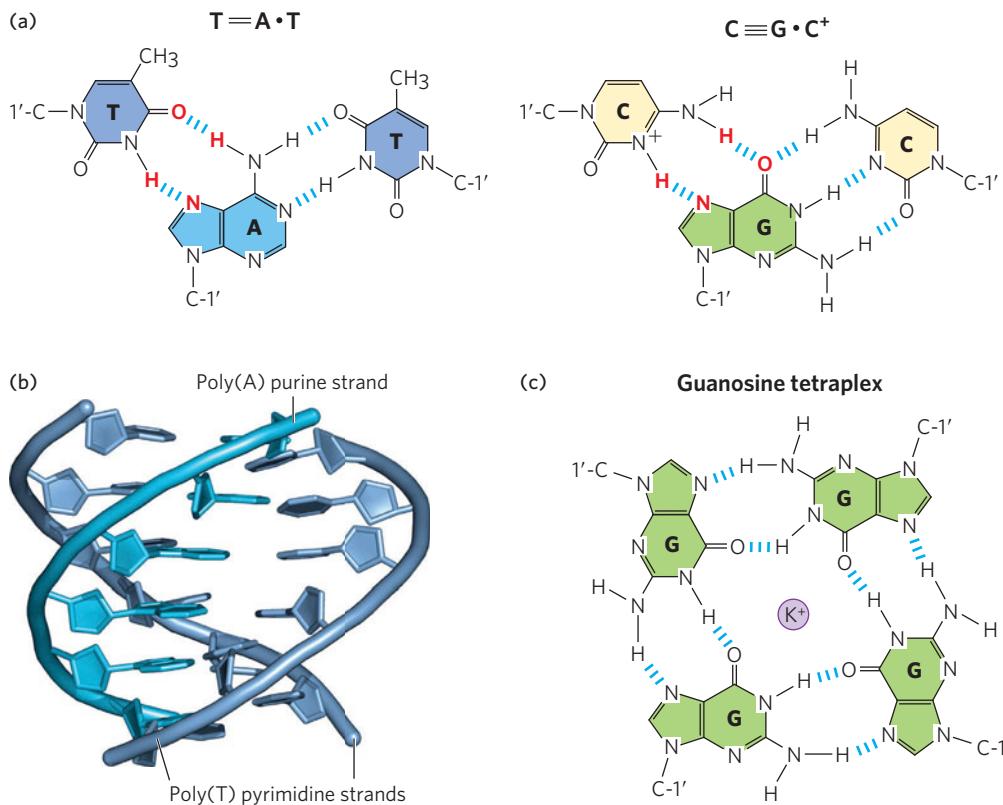


FIGURE 6-21 Three- and four-stranded DNA structures.

(a) Base pairing in triplex DNA. Atoms participating in Hoogsteen pairing are in red; traditional Watson-Crick base pairs are in black. (b) Side view of a triple-helical DNA containing two poly(T) strands and one poly(A) strand. The dark blue and light blue strands in the foreground are antiparallel and engage in normal Watson-Crick base pairing.

The poly(T) strand in the background is parallel to the poly(A) strand and is paired through Hoogsteen hydrogen bonding. (c) One layer of a guanosine tetraplex structure, showing hydrogen bonding of the bases. A K^+ ion in the center of the tetraplex stabilizes the structure by coordinating the bases' functional groups. [Source: (b) PDB ID 1BCE.]

the possibility of using DNA as a container to deliver drugs or proteins (Highlight 6-2).

SECTION 6.2 SUMMARY

- Chargaff's rules state that in double-stranded DNA, the number of adenine nucleotides equals the number of thymine nucleotides ($A = T$), and the number of cytosine nucleotides matches the number of guanosine nucleotides ($G = C$).
- Using Franklin's and Wilkins's x-ray diffraction data from DNA fibers, Watson and Crick proposed that native DNA consists of two antiparallel chains in a right-handed double-helical arrangement. The hydrophilic sugar-phosphate backbone of each strand is on the outside of the helix, and the planar purine and pyrimidine bases project inward, perpendicular to the backbone axis. Complementary base pairs, $A = T$ and $G \equiv C$, are formed by hydrogen bonding within the helix, consistent with Chargaff's rules. The helical structure is further stabilized by shared electrons between the stacked planar base pairs.
- In B-DNA, the most common form of DNA in cells, the base pairs are stacked nearly perpendicular to the long axis of the double helix, 3.4 \AA apart, with 10.5 base pairs per turn.
- Two other variations on DNA structure are A-DNA and Z-DNA. Like B-DNA, A-DNA is right-handed, but it is more compact, with 11 base pairs per turn. A-DNA is favored in solutions that lack water, such as reagents used to crystallize DNA. Z-DNA forms a left-handed helix that contains 12 base pairs per turn and occurs only in sequences rich in C and G residues. Evidence suggests eukaryotic DNA contains short stretches of Z-DNA, which might function in genetic recombination or the regulation of gene expression.

HIGHLIGHT 6-2 TECHNOLOGY

The Design of a DNA Octahedron

The ability of a single strand of DNA to base-pair specifically with its complementary sequence is essential for the accurate replication of encoded information. Scientists have also recognized that base pairing is an extremely useful property for assembling various three-dimensional structures from DNA, for purposes ranging from encapsulation and delivery of pharmaceuticals to DNA-based computing. The trick is to control the base-pairing interactions to favor the formation of desired shapes. A particular challenge is to design sequences that can be cloned and copied by DNA polymerase enzymes in cells and that are also capable of self-assembling into specific three-dimensional shapes.

In 2004, Gerald Joyce and his colleagues at the Scripps Research Institute succeeded in designing a DNA sequence that could fold up into an octahedron in the presence of short complementary oligonucleotides. The research group produced a 1,669-nucleotide, single-stranded DNA molecule containing many such self-complementary sequences to enable specific base pairing between segments (Figure 1).



Gerald Joyce [Source: Courtesy of Gerald F. Joyce, MD, PhD/The Scripps Research Institute.]

To induce three-dimensional folding, five 40-nucleotide oligonucleotides were added to form the struts of an octahedron by base pairing to distinct sites within the 1,669-nucleotide DNA. DNA mixtures were heated to disrupt base pairing and make single strands

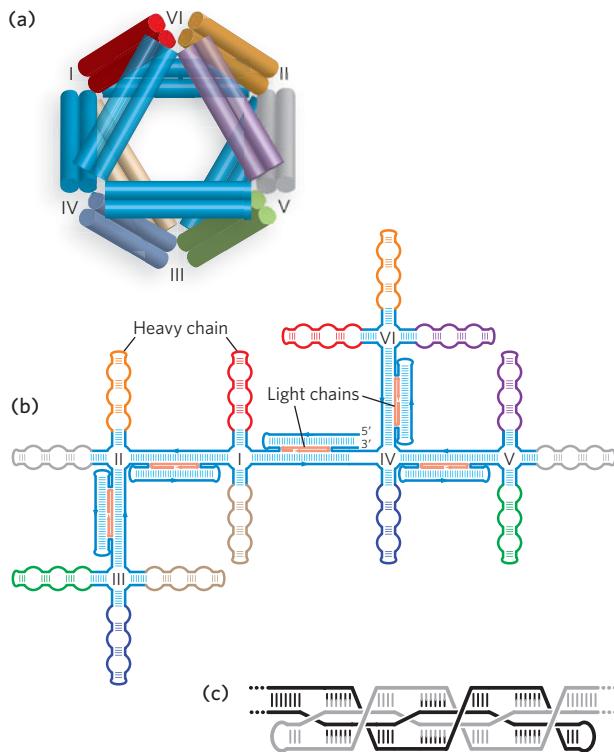


FIGURE 1 (a) The designed three-dimensional DNA octahedral structure has 12 struts—the octahedron edges (double cylinders)—connected by six flexible joints (labeled I–VI). The joints are four-way junctions that connect the core-layer double helices of each strut. Colors correspond to the colored segments in (b), which shows the secondary structure of the branched-tree folding intermediate. This consists of a single 1,669-nucleotide DNA chain (heavy chain) and five unique 40-nucleotide light chains. (c) A close-up of the base-pairing scheme for each strut. Black and gray colors indicate two separate parts of the DNA strand that interact to form a strut. The cross-over base pairing gives the strut its structure and strength. [Source: Adapted from W. M. Shih, J. D. Quispe, and G. F. Joyce, *Nature* 427:618–621, 2004, Fig. 1.]

- Repetitive sequences, such as tracts of A or T residues separated by G- or C-rich segments, cause bends in the DNA molecule. Bending can help facilitate DNA-protein binding.
- DNA strands with inverted repeat sequences can form hairpin or cruciform structures that play roles in recombination and regulation of gene expression. Triplex or tetraplex forms of DNA can occur, though rarely, and may function in DNA metabolism.

6.3 RNA Structure

The discovery that RNA molecules play key roles in converting the genetic information contained in DNA into the proteins that perform structural and catalytic functions in cells motivated the quest to determine the molecular structures of RNA. In the early 1970s, Alexander Rich, Aaron Klug, and Sung-Hou Kim independently solved the structures of transfer RNAs, revealing how

accessible, and then cooled to induce base-pair formation. The changes in DNA structure resulting from base-pair formation were monitored by observing changes in DNA mobility through a gel matrix and visualizing assembled DNA octahedra by electron microscopy (**Figure 2**). The microscopic analysis showed that, as expected from the sequence design, the DNA strands folded with 12 struts or edges joined

at six four-way junctions to form hollow octahedra approximately 22 nm in diameter.

In theory, such DNA structures could be chemically modified to provide binding sites for small molecules or proteins inside the enclosed space. If this idea can be shown to work, simple DNA base pairing will have been harnessed as a practical tool for drug delivery.

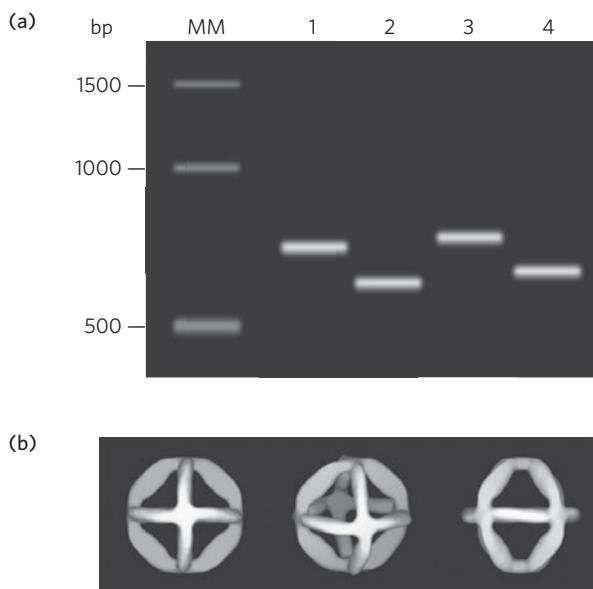


FIGURE 2 Gel electrophoresis and electron microscopy revealed the DNA octahedron assembly through base pairing. (a) Agarose gel electrophoresis of octahedron-forming DNA under different conditions. Lane MM: marker lane with DNA size standards (number of base pairs is indicated on the left). Lane 1: 1,669-nucleotide strand folded in the absence of Mg^{2+} . Lane 2: 1,669-nucleotide strand folded in the presence of Mg^{2+} . Lane 3: 1,669-nucleotide strand and 40-nucleotide light chains folded in the absence of Mg^{2+} . Lane 4: 1,669-nucleotide strand and 40-nucleotide light chains folded in the presence of Mg^{2+} . The Mg^{2+}

shields the negative charges of the backbone phosphates and also promotes base pairing, creating a more compact structure that migrates more quickly through the gel. The resolution of the gel is not sufficient to detect differences in mobility of the DNA \pm the 40-nucleotide light chains.

(b) Three views of the three-dimensional images of assembled DNA octahedra generated computationally, using electron micrographs as the starting point. [Source: W. M. Shih, J. D. Quispe, and G. F. Joyce, *Nature* 427:618–621, 2004, (a) adapted from Fig. 2; (b) Fig. 3b.]

tRNAs carry the amino acids that are used in protein synthesis on the ribosomes. The field of RNA structural biology then languished for almost 20 years, due in part to the technical difficulty of preparing RNA samples for study in the laboratory and an unawareness of the wide variety of biological functions of RNA.

The situation changed in 1990, driven by the discoveries of many new kinds of RNA molecules in

cells and viruses. In fact, many viral genomes, including those of HIV, HCV, and the influenza viruses, are made entirely of RNA, which can be either single-stranded or double-stranded. Questions about the stability of RNA genomes and how their structures might differ from those of genomic DNA increased the urgency of discovering how RNA molecules are structured.



Clockwise from top left:
Alexander Rich
[Source: Courtesy of Donna Coveney/MIT News.]
Aaron Klug [Source:
© James King-Holmes.]
Sung-Hou Kim [Source:
Courtesy of Sung-Hou Kim.]

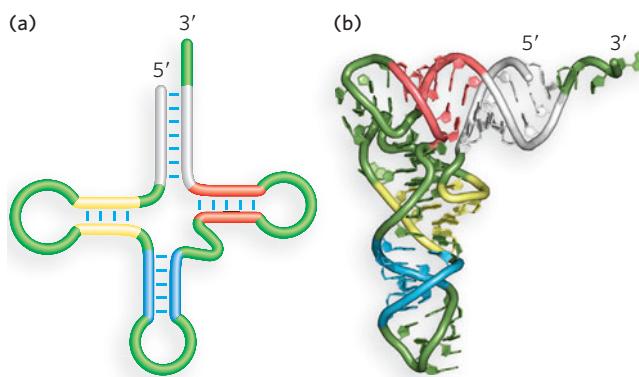


FIGURE 6-22 The three-dimensional structure of tRNA.

(a) The secondary structure of tRNA forms a cloverleaf containing four helices that meet at a central junction. The structure contains several non-Watson-Crick base pairs and loops. (b) The three-dimensional structure of tRNA^{Phe} (a tRNA specific for phenylalanine in protein synthesis), determined by x-ray crystallography, is shown as a ribbon diagram. [Source: (b) PDB ID 1EHZ.]

RNAs Have Helical Secondary Structures

The wide-ranging functions of RNA reflect a structural diversity much richer than that observed in DNA molecules. The propensity of RNA to form compact folded shapes was first revealed in the 1970s with the determination of the molecular structures of several tRNA molecules (Figure 6-22). These structures showed that a single strand of RNA folds back on itself to form short base-paired or partially base-paired segments connected by unpaired regions (Figure 6-23). This property, called **RNA secondary structure**, enables RNA molecules to fold into many different shapes that lend themselves to many different biological functions.

As in DNA, the paired strands in RNA are antiparallel and tend to assume a right-handed helical conformation dominated by base-stacking interactions. Unlike DNA, however, the base-paired segments of RNA are interspersed with a variety of other, non-Watson-Crick, base pairings such as A-A and G-U (Figure 6-24) (see How We Know). In addition, RNA secondary structures include regions of unpaired nucleotides, which can interact with noncontiguous sequences to stabilize the three-dimensional folding. Such interactions produce compact shapes containing surfaces or crevices that bind other molecules or form sites capable of catalyzing chemical reactions, much like protein enzymes.

The greater structural variety in RNA relative to DNA reflects the three main chemical differences between the two polynucleotides: the pentose

(2'-deoxyribose in DNA vs. ribose in RNA), the base composition (thymidine vs. uridine), and the sugar pucker of the pentose (C-2' endo vs. C-3' endo). The presence of the 2'-hydroxyl group on the sugar of RNA

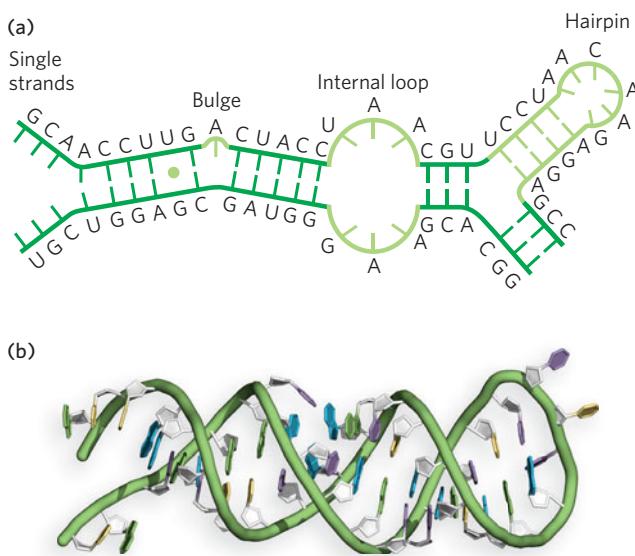


FIGURE 6-23 Double-helical characteristics of RNA.

(a) Some of the diversity in the secondary structure of RNA is shown in these examples of G-U base pairing, bulges, internal loops, and hairpin loop structures. (b) An RNA helix in the form of a hairpin structure similar to the one shown in (a); notice how the unpaired bases at the hairpin loop are incorporated into the structure while maintaining the overall helical geometry of an A-form duplex.

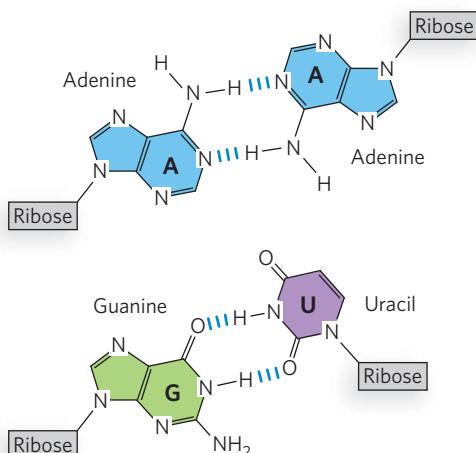


FIGURE 6-24 A-A and G-U base pairs in RNA.

nucleotides provides an extra site for hydrogen bonding, potentially stabilizing three-dimensional folding. This hydroxyl group also influences the sugar pucker, leading to more closely spaced phosphates on the 5' and 3' sides of each sugar and hence to a more compact, A-form helical structure (see Figure 6-17).

As in the DNA double helix, RNA base stacking is made energetically favorable by the resulting burial of hydrophobic surfaces away from the hydrophilic surroundings. In tRNA, as well as in the more recently discovered structures of catalytic RNA molecules and ribosomes, virtually all of the bases are stacked, even when they are not part of Watson-Crick base pairings. In each case, structural motifs, such as base-triple interactions (see Figure 6-21a, b) and helix-helix packing, allow stable three-dimensional folding.

Double-stranded RNAs do exist in nature, such as those that form the genomes of some viruses. These RNAs exist as long helical structures analogous to the DNA double helix. In addition, some RNAs do not seem to form stable three-dimensional structures from local base-pairing interactions. For example, mRNAs seem to perform their function as transient carriers of genetic information without adopting any specific three-dimensional structure. These RNAs may fold into three-dimensional structures only in the presence of bound proteins, forming complexes called **ribonucleoproteins (RNPs)**.

RNAs Form Various Stable Three-Dimensional Structures

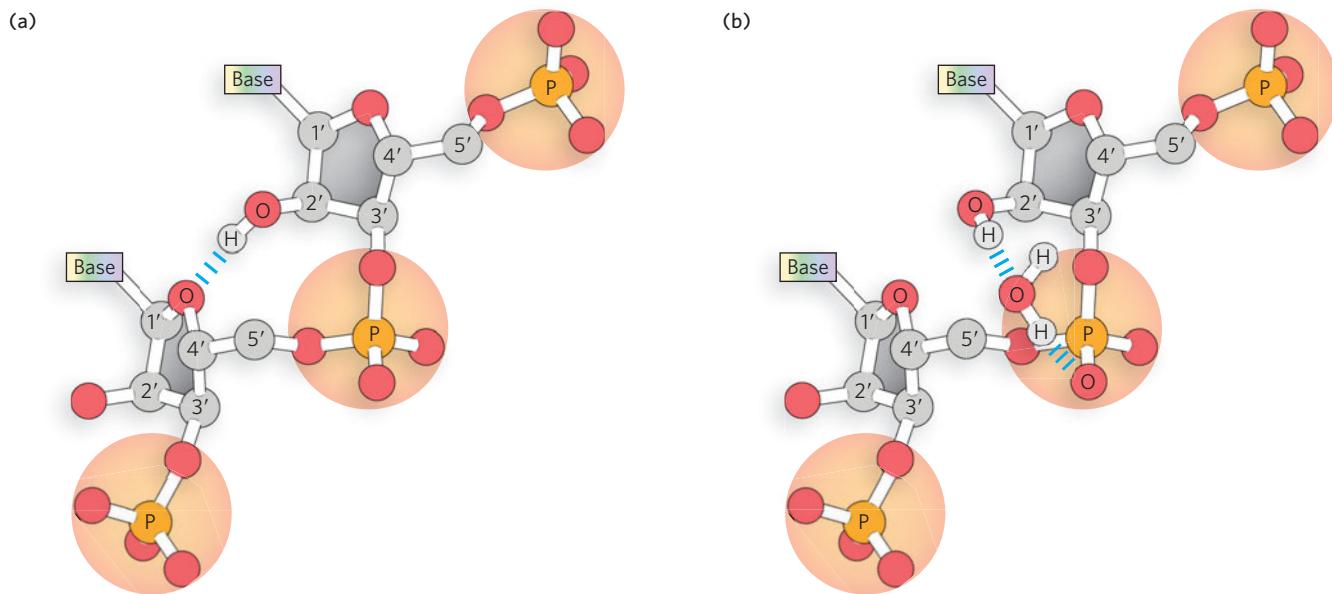
Most of the highly structured RNAs contain noncanonical base pairs and backbone conformations not observed in DNA. In many cases, the 2'-hydroxyl group

on ribose, a chemical feature that distinguishes RNA from DNA, seems to be directly or indirectly responsible for these unique structural properties. Recall that the presence of this 2' hydroxyl gives ribose its C-3' endo geometry (its sugar pucker), as distinct from the C-2' endo characteristic of deoxyribose. This seemingly small difference in chemical conformation leads to a distinct helical geometry for RNA relative to DNA. Furthermore, it enables direct or water-mediated hydrogen bonding between the 2'-hydroxyl of one RNA nucleotide and the adjacent ribose or phosphate (Figure 6-25). As a result, RNA helices are more thermodynamically stable than are DNA helices of the same length and sequence.

Where complementary sequences are present in an RNA molecule, the predominant double-stranded structure in the RNA is an A-form, right-handed double helix (see Figure 6-17). The A-RNA helix has a wider, shallower minor groove and a narrower, deeper major groove compared with the B-form helix observed for most DNA. The A-form geometry has a shorter distance between adjacent phosphates in the sugar-phosphate backbone than in B-DNA, a consequence of the C-3' endo sugar pucker of ribose. A B-form RNA helix has not been observed in nature. Z-RNA helices have been induced to form in the laboratory under high-salt or high-temperature conditions, but are not known to occur in cells.

As mentioned previously, mismatched or unmatched bases are common in base-paired segments of RNA, locally disrupting the regular A-form helix and resulting in bulges or internal loops (see Figure 6-23). The potential for base pairing within a single strand of RNA frequently produces thermodynamically stable secondary structures that consist of hairpinlike conformations capped by connecting loops. Such hairpins are the most common type of RNA secondary structure, often containing specific short sequences at their ends (such as UUCG or GAAA) that form particularly energetically favorable loops. The nucleotides in the loops are arranged to maximize hydrogen bonding and base stacking, thereby enhancing thermodynamic stability. Important additional structural contributions are made by hydrogen bonds that are not part of canonical Watson-Crick base pairs. These properties, evident in the structures of tRNAs and catalytic RNAs, were also evident in the ribosome structures solved by Jamie Cate and others (see Moment of Discovery). The functions of these highly structured RNAs, like those of proteins, depend on their three-dimensional properties.

Weak interactions, especially base-stacking interactions, play a major role in stabilizing RNA structures,

**FIGURE 6-25** Stabilization of RNA secondary structure.

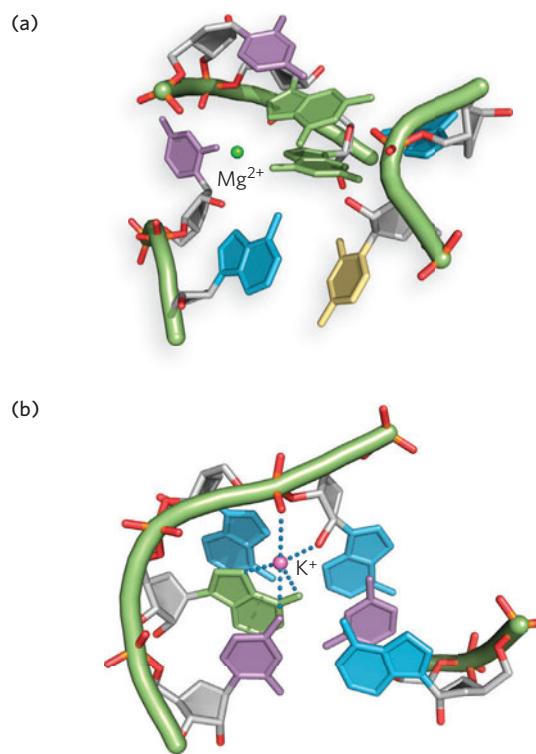
The 2'-hydroxyl group on ribose causes RNA to favor the C-3' endo sugar pucker, due to (a) direct hydrogen bonding between a 2' hydroxyl on one ribose and the oxygen on an adjacent ribose, and (b) hydrogen bonding, through a water molecule, between a 2' hydroxyl and a phosphate oxygen.

[Source: Adapted from W. Saenger, *Principles of Nucleic Acid Structure*, Springer-Verlag, 1984, Fig. 4-11.]

just as they do in DNA. The 2' hydroxyl is often involved in hydrogen bonds and van der Waals interactions that stabilize alternative helical shapes and conformations such as loops and kinks that require the close approach of the two phosphodiester backbones. Divalent and monovalent metal ions—such as Mg^{2+} , Ca^{2+} , K^+ , and Na^+ —bind to specific sites in RNA and help shield the negative charge of the backbone, allowing parts of the molecule to pack more tightly together (Figure 6-26).

The analysis of RNA structure, along with the relationship between structure and function, is an emerging field of inquiry with many of the same complexities as the analysis of protein structure. The importance of understanding RNA structure has grown as we have become increasingly aware of the large number of functional roles for RNA molecules. For example, we now

know that the extensive structural features found in the genomic RNA of HIV control viral gene expression (Highlight 6-3).

**FIGURE 6-26** Mg^{2+} and K^+ binding in RNA structure.

(a) Divalent magnesium ions (Mg^{2+}) coordinate to phosphate groups and stabilize the close approach of phosphate backbones in the folded structure of the P4-P6 domain of the *Tetrahymena* group I ribozyme (a catalytic RNA).
 (b) Monovalent potassium ions (K^+) bind to specific sites in the P4-P6 domain, where they favor interactions between backbone and bases. [Source: (a) and (b) PDB ID 1GID.]

HIGHLIGHT 6-3 MEDICINE

RNA Structure Governing HIV Gene Expression

RNA structures can be investigated by testing for reactivity to various chemical reagents. This is possible because bases that are paired or are folded inside the RNA, or those that are conformationally rigid, are protected against modification or cleavage by reactive chemicals. By analyzing which sites in a folded RNA molecule are resistant or sensitive to chemical reagents, researchers can obtain information about its structural features.

Kevin Weeks and his colleagues at the University of North Carolina used this approach to analyze the role of RNA structure in regulating gene expression in HIV-1 (a strain of the human immunodeficiency virus). Viral RNA was extracted from virus particles (virions) under gentle conditions, such that the native structure was preserved. The RNA was then treated with a chemical reagent, 1-methyl-7-nitroisatoic anhydride (1M7), which preferentially acylates conformationally flexible nucleotides at the 2'-OH of ribose. After the RNA was allowed to react with 1M7, acylated sites were detected by reverse transcription. The purified enzyme reverse transcriptase was used to make a DNA copy of the viral RNA; acylation blocked the progression of polymerization, generating a truncated DNA fragment extending up to the site of acylation in the sequence.

By fractionating the resulting DNAs, the sites of acylation were detected and mapped onto the HIV-1 genome sequence.

Analysis of these data revealed that in addition to encoding viral proteins in its nucleotide sequence, the viral genomic RNA is also three-dimensionally structured



Kevin Weeks [Source:
Courtesy of Kevin Weeks.]

to optimize protein production (Figure 1). Highly structured regions in the RNA—and hence a dearth of experimentally acylated sites—occur within sequences that encode interprotein linkers or loops between protein domains. This finding implies that RNA structure slows down the rate at which ribosomes move through those parts of the protein-coding sequence, providing time for the newly synthesized viral proteins to fold into their active structures. The results of this study underscore the idea that the HIV genome, and perhaps mRNAs, contain structured regions that regulate expression of the proteins they encode. RNA secondary and higher-order structure may therefore constitute an important component of the genetic code.

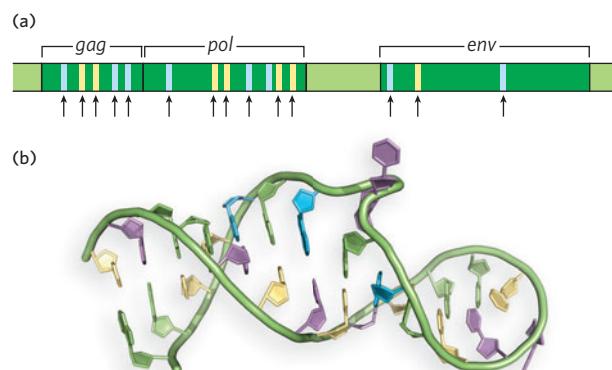


FIGURE 1 The organization and structure of the HIV-1 genomic RNA influences viral protein expression.
 (a) This portion of the HIV-1 genome shows three open reading frames, *gag*, *pol*, and *env*, that are translated into long polyprotein precursors; proteolytic cleavage generates multiple proteins from each reading frame. Blue bars represent interprotein linkers, and yellow bars indicate sequence encoding loops between protein domains. Black arrows mark regions of the RNA genome that are highly structured; note that these areas coincide with areas that link proteins or protein domains.
 (b) An example of RNA secondary structure in the HIV-1 genome. [Sources: (a) Adapted from J. M. Watts et al., *Nature* 460:711–716, 2009, Fig. 1; (b) PDB ID 2KMJ.]

SECTION 6.3 SUMMARY

- RNA is chemically better suited than DNA to forming stable three-dimensional folds, due to the 2'-hydroxyl group of ribose in the RNA backbone. The presence of the 2' hydroxyl makes RNA vulnerable to

hydrolysis, but it also allows for additional hydrogen bonding between segments of an RNA molecule.

- RNA molecules can exist as long double-helical structures, typical of viral genomic RNA, or, more commonly, as single strands that fold up into short

helical regions connected by loops and unpaired segments.

- Base-paired segments of RNA generally adopt the compact geometry of A-form helices. This structure arises because RNA favors a different sugar pucker from that found in DNA (C-3' endo in RNA vs. C-2' endo in DNA), which causes the phosphates in the backbone to become more closely spaced than in B-DNA. In some large RNAs, short helices interact through RNA-RNA and RNA-metal ion contacts to form complex three-dimensional structures.
- Base pairs other than canonical A=U and C≡G pairs are common in RNAs, including A-A and G-U. In all cases, base pairs or single bases are most stable when stacked on top of one another in a helix.
- Divalent and monovalent metal ions (Mg^{2+} , Ca^{2+} , K^+ , and Na^+) bind to specific sites in RNA and help shield the negative charge of the backbone, allowing parts of the molecule to pack more tightly together.

of pH or to temperatures above 80°C, its viscosity decreases sharply, indicating that the DNA has undergone a physical change. This change is due to **denaturation**, or **melting**, of the double-helical DNA, and can also occur with RNA. Disruption of both the hydrogen bonding between paired bases and the base stacking causes the double helix to unwind, forming two single strands that are completely separate from each other along the entire (or partial) length of the molecule. No covalent bonds in the nucleic acid are broken during denaturation (Figure 6-27).

Renaturation of a DNA or RNA molecule is a rapid, one-step process, as long as a double-helical segment of at least a dozen residues still unites the two strands. When the temperature or pH is returned to the range in which most organisms live, the unwound segments of the two strands spontaneously rewind to yield the intact duplex. This process, called **annealing**, involves re-formation of all the base pairs in the double helix. If the strands were completely separated, renaturation occurs in two steps. In the first step, which is relatively slow, complementary sequences in the two

6.4 Chemical and Thermodynamic Properties of Nucleic Acids

To understand how nucleic acids function, we must understand their chemical properties as well as their structures. The role of DNA as a repository of genetic information depends in part on its inherent stability. The chemical transformations that do happen are generally very slow in the absence of an enzyme catalyst. The long-term storage of information without alteration is so important to a cell, however, that even very slow changes in DNA structure can be physiologically significant. Other, nondestructive alterations of DNA do occur and are essential to function, such as the strand separation that must precede replication or transcription. For RNAs, chemical modifications can play significant roles in ensuring correct structure and function.

In addition to providing insights about physiological processes, our understanding of nucleic acid chemistry gives us a powerful array of technologies that have applications in molecular biology, medicine, and forensic science. We examine here the chemical properties of DNA and RNA, as well as some of these technologies. Many more techniques and applications are discussed in Chapter 7.

Double-Helical DNA and RNA Can Be Denatured

Solutions of carefully isolated, double-stranded DNA are highly viscous at pH 7.0 and room temperature (25°C). When such a solution is subjected to extremes

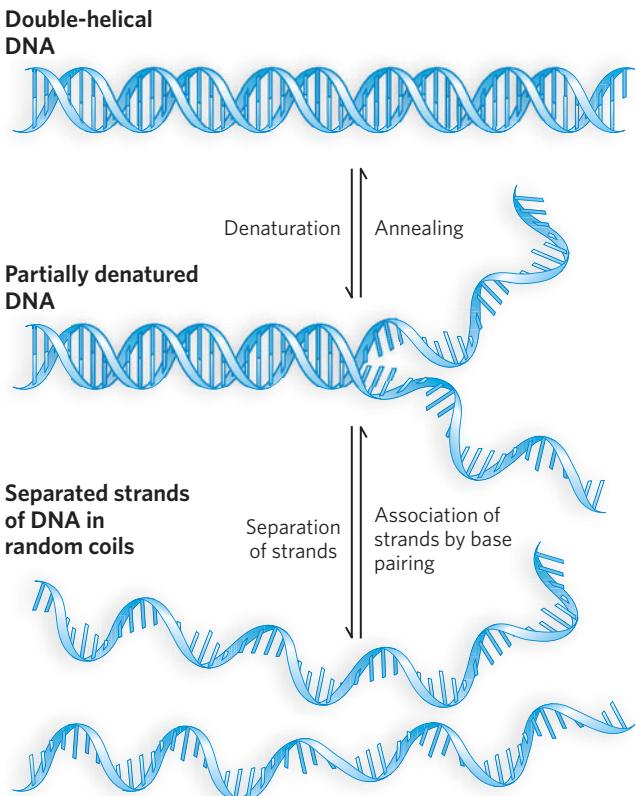


FIGURE 6-27 Reversible denaturation of DNA. DNA is shown here, but RNA is also capable of denaturation and reannealing.

strands “find” each other by random collisions and form a short segment of double helix. The second step is much faster: the remaining unpaired bases successively come into register as base pairs, and the two strands “zipper” themselves together to form the double helix.

The close interaction of stacked bases in a nucleic acid has the effect of decreasing its absorption of UV light relative to that of a solution with the same concentration of free nucleotides, and the absorption is further decreased by the pairing of two complementary strands. Hydrogen bonding and base stacking in the double helix limit the resonance of the aromatic rings of the bases, thereby decreasing UV light absorption. This is known as the **hypochromic effect** (Figure 6-28). Once the DNA is denatured, the base pairs are disrupted and the two strands separate into randomly coiled chains. The resonance of the bases in each strand is no longer constrained. As a result, the UV light absorption of single-stranded DNA is approximately 40% higher than that of double-stranded DNA at the same concentration. This increase in absorption is called the **hyperchromic effect**. The transition from double-stranded DNA to the single-stranded, denatured form can thus be detected by monitoring the absorption of UV light.

DNA molecules in solution denature when they are heated slowly. Each species of DNA has a characteristic denaturation temperature, or **melting point**

(T_m), defined as the temperature at which half the DNA is denatured. In general, the higher the content of G≡C base pairs in the DNA, the higher its melting point. This is because G≡C base pairs, with three hydrogen bonds, require more heat energy to dissociate than do A=T base pairs. Careful determination of the melting point of a DNA specimen, under fixed conditions of pH and ionic strength, can yield an estimate of its base composition. If denaturation conditions are carefully controlled, regions that are rich in A=T base pairs will specifically denature while most of the DNA remains double-stranded. Such denatured regions, or bubbles, can be visualized with electron microscopy (Figure 6-29). Strand separation of DNA must occur *in vivo* during processes such as DNA replication and transcription. As we’ll see in Chapter 11, the DNA sites where these processes are initiated are often rich in A=T base pairs.

RNA duplexes or RNA-DNA hybrid duplexes can also be denatured. Notably, RNA duplexes are more stable than DNA duplexes. At neutral pH, the denaturation of a double-helical RNA often requires temperatures 20°C or more higher than those required to denature a DNA molecule with a comparable sequence. The stability of an RNA-DNA hybrid is generally intermediate between that of double-stranded RNA and DNA.

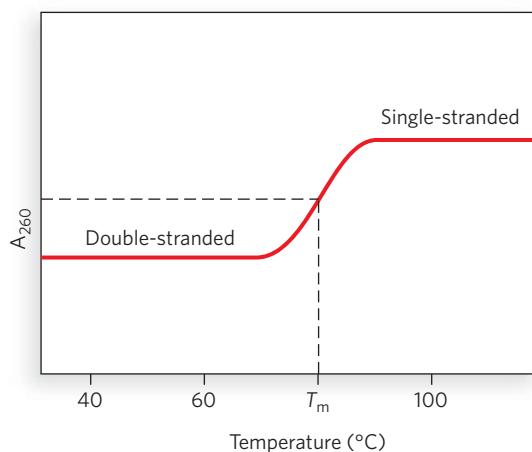


FIGURE 6-28 UV light absorption by DNA. The transition from double-stranded DNA to the single-stranded, denatured form can be detected by monitoring UV light absorption by the sample (shown here as A_{260} , absorbance at 260 nm). The melting point (T_m) is the temperature at which half the DNA in the sample is denatured.

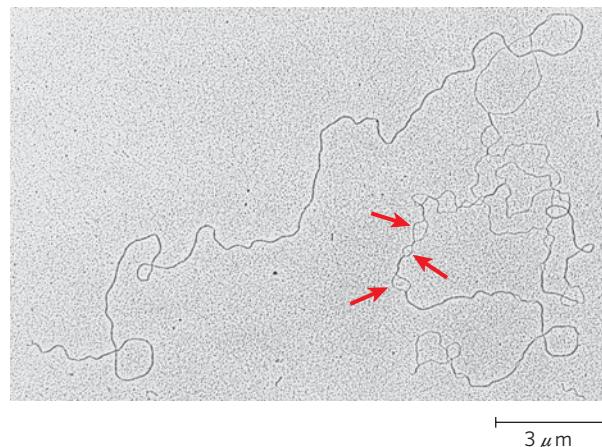


FIGURE 6-29 Partially denatured DNA. The DNA shown in this electron micrograph was partially denatured, then fixed to block renaturation during sample preparation. The arrows point to some single-stranded bubbles where denaturation has occurred. The regions that denature are reproducible and are rich in A=T base pairs. [Source: Ross B. Inman, University of Wisconsin-Madison, Department of Molecular Biology.]

Nucleic Acids from Different Species Can Form Hybrids

The ability of two complementary DNA strands to pair with each other can be used to detect similar DNA sequences in different species or within the genome of a single species. If double-stranded DNAs isolated from human cells and from mouse cells are completely denatured by heating, then mixed and kept at 25°C for many hours, much of the DNA will anneal. Most of the mouse DNA strands anneal with complementary mouse DNA strands to form mouse duplex DNA; similarly, most human DNA strands anneal with complementary human DNA strands. However, some strands of the mouse DNA will associate with human DNA to yield **hybrid duplexes**, in which segments of a mouse DNA strand form base-paired regions with segments of a human DNA strand (Figure 6-30). The ability to hybridize DNA from different species is a valuable laboratory tool for exploring evolutionary relationships. Different species have proteins and RNAs with similar functions—and often similar structures. In many cases, the DNAs encoding these proteins and RNAs have similar sequences. The closer the evolutionary relationship between two species, the more extensively their DNAs will hybridize. For example, human DNA hybridizes much more extensively with mouse DNA than with DNA from yeast.

The hybridization of DNA strands from different sources forms the basis for a powerful set of techniques essential to the practice of modern molecular genetics. A specific gene's DNA or RNA sequence can be detected in the presence of many other sequences by hybridization with a **probe**, a carefully chosen nucleic acid sequence complementary to the gene of interest. To be visualized in the laboratory, the probe must be labeled in some way, usually radioactively or with a fluorophore (a compound carrying a fluorescent group). The probe that is selected depends on what is known about the gene under investigation. Sometimes, a gene from another species that has sequence similarity to the gene of interest makes a suitable probe. Or, if the protein product of a gene has been purified, probes can be designed and synthesized by working backward from the amino acid sequence, deducing the DNA sequence that would code for it. Researchers can obtain the necessary DNA sequence information from sequence databases that detail the structure of millions of genes from a wide range of organisms. Because base-pairing stability is sensitive to pH and temperature, these parameters can be adjusted experimentally to detect nucleic acid

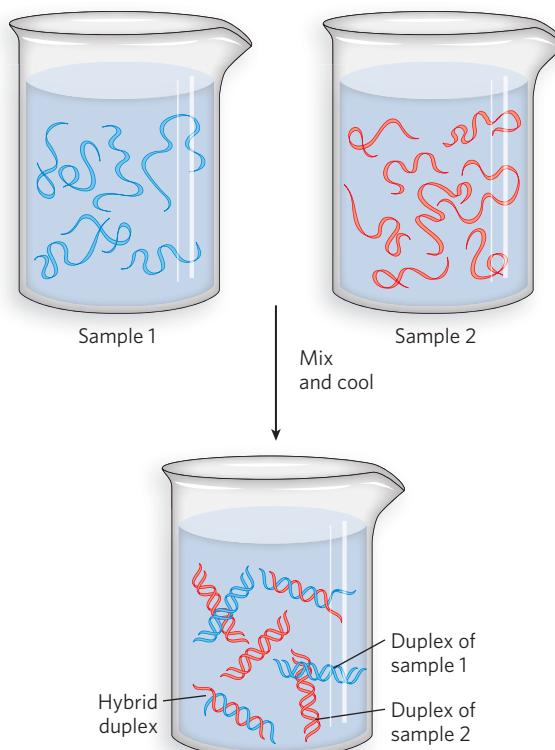


FIGURE 6-30 Cross-species DNA hybridization. Two DNA samples can be compared by heating to denature the strands and then cooling the mixture to allow complementary strands to form duplexes. The greater the similarity between the two samples, the more hybrid duplexes will form, with one strand derived from the first species and the other from the second.

sequences with varying degrees of complementarity to the probe. The technique is sensitive enough to reveal sequences that differ by a single base pair. This can be critically important in medical and forensic applications.

Hybridization techniques for detecting specific DNAs or RNAs are diverse, and the selected method depends on how the starting sample is prepared. One classic approach, called colony blot hybridization, is shown in Figure 6-31. This technique identifies specific DNA sequences from a collection of sequences that have been inserted into bacterial cells. (Such a preparation is known as a DNA library; see Chapter 7.) In this method, nitrocellulose paper is pressed onto an agar plate containing many individual bacterial colonies, each containing a different inserted DNA sequence. Some cells from each colony adhere to the paper, forming a replica of the plate. The paper is treated with alkali to disrupt the cell membranes and denature the released DNA, which remains bound to the region of

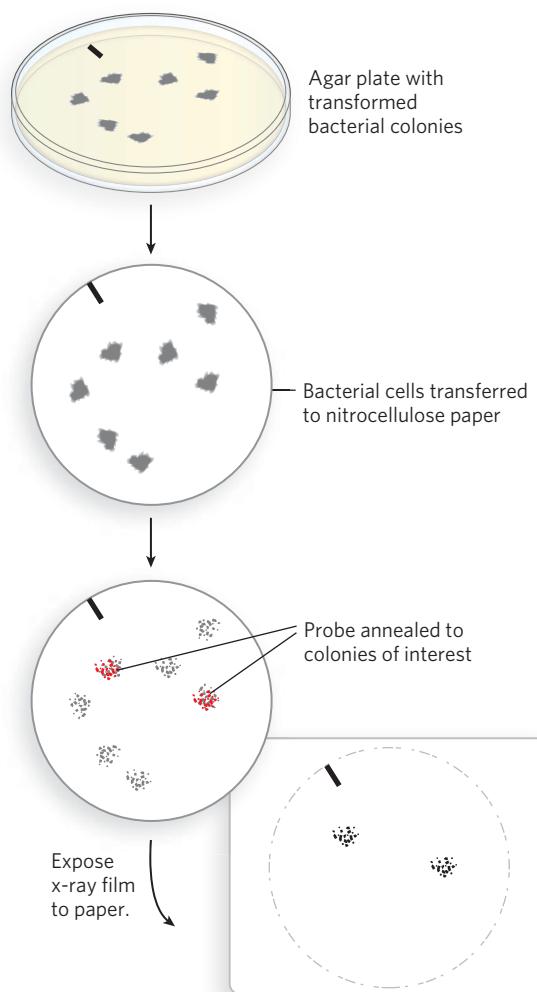


FIGURE 6-31 Colony blot hybridization. See text for details.

the paper around the original colony. An added radioactive DNA probe anneals only to its complementary DNA. After any unannealed probe DNA is washed away, the hybridized DNA can be detected by exposing the probe-containing paper to photographic film (a technique called autoradiography). More commonly, the radioactive probe is detected by exposing the paper to a phosphor storage screen, which is then scanned in a phosphoimager.

In other common hybridization methods, **gel electrophoresis** is used to separate DNA or RNA molecules by size (Figure 6-32a). A variation of gel electrophoresis, used to detect proteins under denaturing conditions, was discussed in Chapter 4 (see Highlight 4-1). Here, the gel matrix is not denaturing but instead is made of agarose, a kelp-derived material that does not disrupt nucleic acid base pairing.

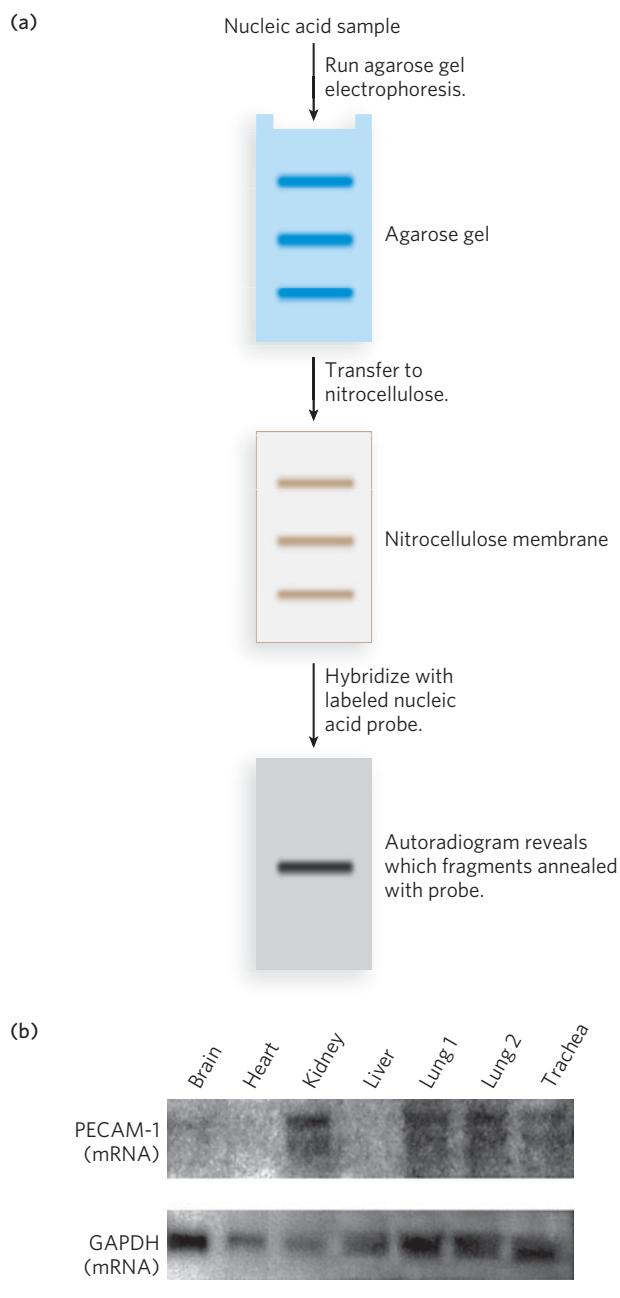
The starting nucleic acid sample is in solution in a test tube and is applied to a slot at one end of the gel, and a voltage is applied. Because DNA and RNA molecules are negatively charged, they migrate toward the positive end of a gel matrix in an electric field. Larger molecules tend to move more slowly than smaller ones, so this provides a means of separating nucleic acids by size. Following electrophoresis, the gel is placed on a nitrocellulose membrane, and this assembly is sandwiched between two pieces of filter paper. The stack is put into an alkaline solution, and capillary action transfers the nucleic acid from the gel matrix to the nitrocellulose membrane. Once on the membrane, the nucleic acid is immobilized and can then be hybridized with a DNA or RNA probe, labeled so that it can be detected by measuring its radioactivity or fluorescence.

When used to detect DNA, this method is known as **Southern blotting**, named for Edwin Southern, who invented the technique at the University of Edinburgh. When used for RNA detection, the technique is called **Northern blotting**, because of its similarity to the Southern method. Applications of these techniques include identifying a person on the basis of a single hair left at the scene of a crime, or predicting the onset of a disease decades before symptoms appear. Northern blotting can also be used to detect the levels of a particular type of RNA in different body tissues (Figure 6-32b), providing fascinating insight into how cells regulate the expression of genes.

Nucleotides and Nucleic Acids Undergo Uncatalyzed Chemical Transformations

Purines and pyrimidines, and the nucleotides of which they are a part, can undergo spontaneous alterations in their covalent structure. The rate of these reactions is generally very slow, but as noted earlier, they are physiologically significant because of the cell's low tolerance for changes in its genetic information. Alterations in DNA structure that produce permanent genetic changes are known as **mutations**. Extensive evidence suggests an intimate link between the accumulation of mutations in an individual organism and the processes of aging and carcinogenesis.

Several nucleotide bases undergo **deamination**, a spontaneous loss of their exocyclic amino groups. For example, under typical cellular conditions, deamination of cytosine (in DNA) to uracil occurs in about 1 of every 10^7 C residues in 24 hours (Figure 6-33). This corresponds to about 100 spontaneous events per day, on average, in a mammalian

**FIGURE 6-32** Southern and Northern blotting techniques.

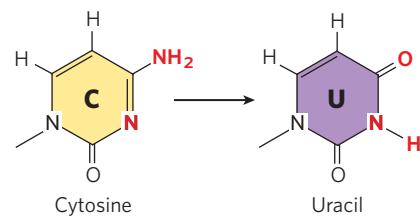
(a) Gel electrophoresis is used to size-fractionate a DNA (Southern blotting) or RNA (Northern blotting) mixture. The samples are then transferred to a nitrocellulose membrane by capillary action. Once bound to the membrane, the nucleic acids can be detected by using short radiolabeled oligonucleotide probes that base-pair to the samples on the membrane. (b) Northern blot analysis of RNA isolated from various human tissues. For each sample, approximately 10 µg of total RNA was separated on a 1.2% agarose-formaldehyde gel, transferred to a membrane, and hybridized to a ^{32}P -labeled probe, an mRNA for human platelet endothelial cell adhesion molecule (PECAM-1). The same blot was also probed with a cDNA (complementary DNA, a DNA copy of an mRNA sequence; see Chapter 7) for glyceraldehyde 3-phosphate dehydrogenase (GAPDH) to control for the amount of material in each lane. Note the differences in PECAM-1 RNA levels detected in the different tissues; two bands are observed for PECAM-1 in each lane because there are two distinct forms of the mRNA for this gene. [Source: (b) Courtesy of Yongji Wang.]

during replication. Establishing thymine as one of the four bases in DNA may well have been a crucial turning point in evolution, making the long-term storage of genetic information possible.

Cytosine deamination also provides innate cellular defense against viral infection. A family of human proteins called APOBECs catalyze cytosine deamination in the viral genome during the initial round of replication by HIV. This hypermutation results in many nonviable viral particles, eventually destroying the coding capacity of the virus. In HIV and related viruses, the protein Vif binds to APOBECs and triggers their degradation. Vif has therefore become an important antiviral target, because viruses lacking this protein are much less capable of establishing chronic infection in human cells.

cell. Deamination of adenine and guanine occurs at about 1/100th this rate.

The slow cytosine deamination reaction seems innocuous enough, but it is almost certainly the reason that DNA contains thymine rather than uracil. Uracil is the product of cytosine deamination, and it is readily recognized as foreign in DNA and is removed by a repair system (see Chapter 12). If DNA normally contained uracil, the recognition of U residues resulting from cytosine deamination would be more difficult, and unrepaired uracils would lead to permanent sequence changes as they were paired with adenines

**FIGURE 6-33** Cytosine deamination to uracil. Only the base is shown.

Base Methylation in DNA Plays an Important Role in Regulating Gene Expression

Certain nucleotide bases in DNA molecules are enzymatically methylated, usually after DNA synthesis is complete. Adenine and cytosine are methylated more frequently than guanine (Figure 6-34a). Methylation is generally confined to certain sequences or regions of a DNA molecule. For example, more than half of all CpG sequences in mammalian genomes are methylated on the C residue (see below). Methylation tends to inhibit gene expression, because the methylated DNA is not efficiently

copied into RNA. In many cancers, gene regulatory regions in DNA become abnormally hypermethylated. This can result in the silencing of genes that would otherwise control cell growth. DNA methylation may affect gene transcription by physically blocking the binding of proteins that facilitate transcription. Other proteins, however, can specifically bind to methylated DNA and recruit additional proteins that help form compact, inactive regions of chromosomal DNA.

All known DNA methylases (methyltransferases) use S-adenosylmethionine as a methyl group donor (see Figure 6-13). *E. coli* has two prominent methylation systems. One serves as part of a defense mechanism that helps the cell distinguish its DNA from foreign DNA by marking its own DNA with methyl groups, and destroying foreign, nonmethylated DNA, a process known as restriction modification. The other system methylates A residues in the sequence 5'-GATC-3' to *N*⁶-methyldeoxyadenosine. Methylation in this case is mediated by the Dam (DNA adenine methylation) methylase, a component of a system that repairs the mismatched base pairs that occasionally form during DNA replication (see Chapter 12).

In eukaryotic cells, about 5% of C residues in DNA are methylated to 5-methyldeoxycytidine (see Figure 6-34a). Methylation is most common at CpG sequences, producing methyl-CpG symmetrically on both strands of the DNA. The extent of methylation of CpG sequences varies by molecular region in large eukaryotic DNA molecules. Methylation suppresses the migration of segments of DNA called transposons (see Chapter 14). These methylations of C residues also have structural significance. The presence of 5-methyldeoxycytidine in an alternating CpG sequence markedly increases the tendency for that segment of DNA to assume the Z form.

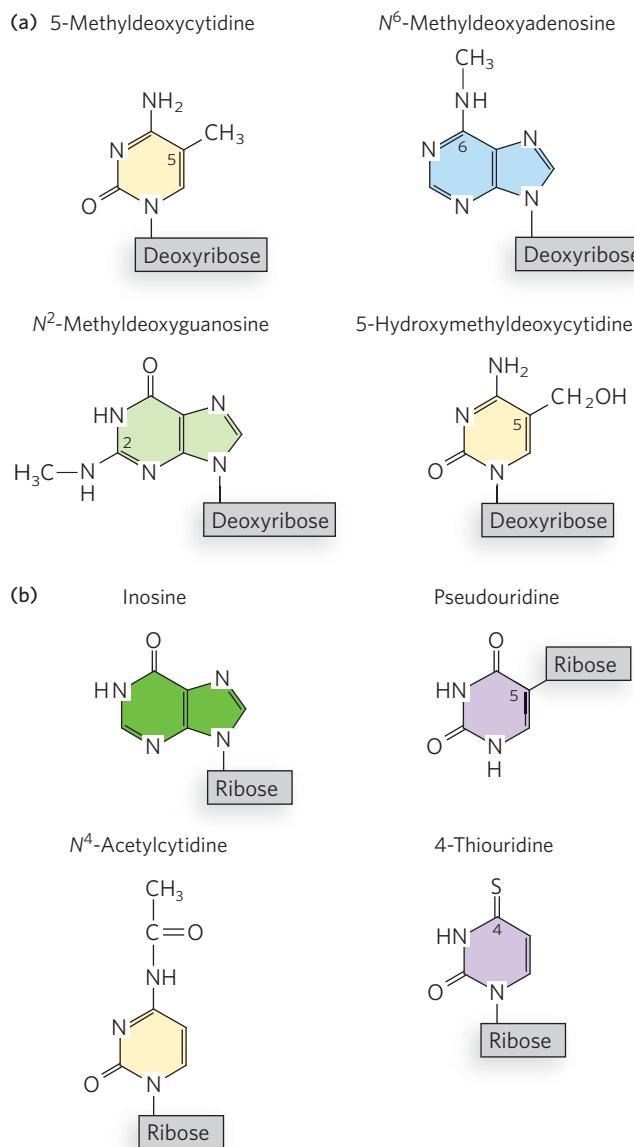


FIGURE 6-34 Chemical modifications in DNA and RNA.

(a) Modified nucleotides in DNA. The most common postsynthetic modification to DNA is base methylation. 5-Methyldeoxycytidine occurs in the DNA of animals and higher plants; the other methylated bases shown here can also be produced by specific enzymes. (b) Modified nucleotides in RNA. Enzyme-catalyzed RNA base modifications are common in tRNA and rRNA, although the function of such alterations is not always clear. The presence of *N*⁴-acetylcytidine in bacterial tRNAs may enhance protein synthesis.

KEY CONVENTION

When a chemical group attached to an atom in the purine or pyrimidine ring is altered, the ring position of the substituent is indicated by the number of that atom—for example, 5-methylcytosine, 7-methylguanine, and 5-hydroxymethylcytosine; the element to which the substituent is attached (N, C, O) is not identified. When a chemical group is altered on an exocyclic atom, the type of atom is identified and the ring position to which it is attached is denoted with a superscript. For example, the amino nitrogen attached to C-6 of adenine is N^6 ; the carbonyl oxygen and amino nitrogen at C-6 and C-2 of guanine are O^6 and N^2 , respectively.

RNA Molecules Are Often Site-Specifically Modified In Vivo

Like DNA, many functional RNAs are posttranscriptionally modified at specific nucleotides (see [Figure 6-34b](#)). Some of the first examples were discovered in ribosomal and transfer RNAs. In some cases, modifications involve the addition of a functional group to an existing nucleotide in the sequence. For example, a methyl group can be added to the 2' hydroxyl of ribose, thereby blocking its ability to form a hydrogen bond. In bacteria, some tRNAs are modified with N^4 -acetylcytidine in a process thought to contribute to the accuracy of protein synthesis. In other cases, the base itself is changed, or its linkage to the ribose—the glycosidic bond—is altered. For instance, inosine, 4-thiouridine, and pseudouridine are relatively common in tRNAs and rRNAs.

Many of the enzymes that catalyze these chemical modifications of RNA are known. They are often evolutionarily conserved, indicating that RNA modification has been occurring in biological systems for a long time. More difficult to figure out is the function of these chemical changes in RNA. Molecular biologists can produce unmodified versions of RNAs in the laboratory and compare their functions to those of the chemically altered counterparts isolated from cells. This approach has only rarely discerned much of an effect of a modified base. However, genetic experiments in which an RNA-modifying enzyme is mutated or deleted from an organism suggest that these enzymes give cells a subtle but important selective advantage over organisms that don't modify their RNA. Some evidence supports the hypothesis that RNA modifications stabilize RNA structures and help RNAs interact with proteins in the cell.

The Chemical Synthesis of DNA and RNA Has Been Automated

Knowledge of DNA and RNA chemistry provided the basis for devising methods to synthesize nucleic acids in the laboratory. This technology has paved the way for many biochemical advances that depend on the ability to synthesize oligonucleotides with any chosen sequence. The chemical methods for synthesizing nucleic acids were developed primarily by H. Gobind Khorana and his colleagues in the 1970s. Refinement and automation of these methods have made possible the rapid and accurate synthesis of DNA strands.

DNA (or RNA) synthesis is carried out with the growing strand attached to a solid support ([Figure 6-35](#)). First, a nucleotide is attached to the support, a glass or polystyrene bead, through its 3'-hydroxyl group, and polynucleotide synthesis proceeds in the $3' \rightarrow 5'$ direction. This is the opposite of the direction of biological polynucleotide synthesis by polymerase enzymes, which occurs in the $5' \rightarrow 3'$ direction. Functional groups on the bases and phosphates, including hydroxyls and amines, are transiently protected with chemical groups that are readily removed after the synthesis is complete. The 5'-hydroxyl group is temporarily protected by a dimethoxytrityl (DMT) group; the DMT group is removed from the end of the growing polymer chain at the beginning of each cycle (step 1) to permit extension of the chain by another nucleotide (step 2). Oxidation of the phosphite linkage between the nucleotides completes the cycle (step 3). When chain synthesis is complete, protecting groups are removed from the bases and phosphates, and the oligonucleotide chain is cleaved from its solid support (steps 4, 5, 6). The efficiency of each addition step is very high, allowing the routine laboratory synthesis of polymers containing 70 to 80 nucleotides and, in some laboratories, much longer strands.

Oligonucleotide synthesis is very useful for techniques such as Southern and Northern blotting and for the polymerase chain reaction (PCR) and DNA sequencing, which are discussed in Chapter 7. In addition, chemical synthesis makes it possible to incorporate chemical modifications in the polymer product, such as biotin groups, extra phosphates, sulfhydryl groups, and methyl groups. These functional groups are useful for such applications as specific labeling of a DNA strand or stabilization of an RNA oligonucleotide against enzymatic degradation in cells.

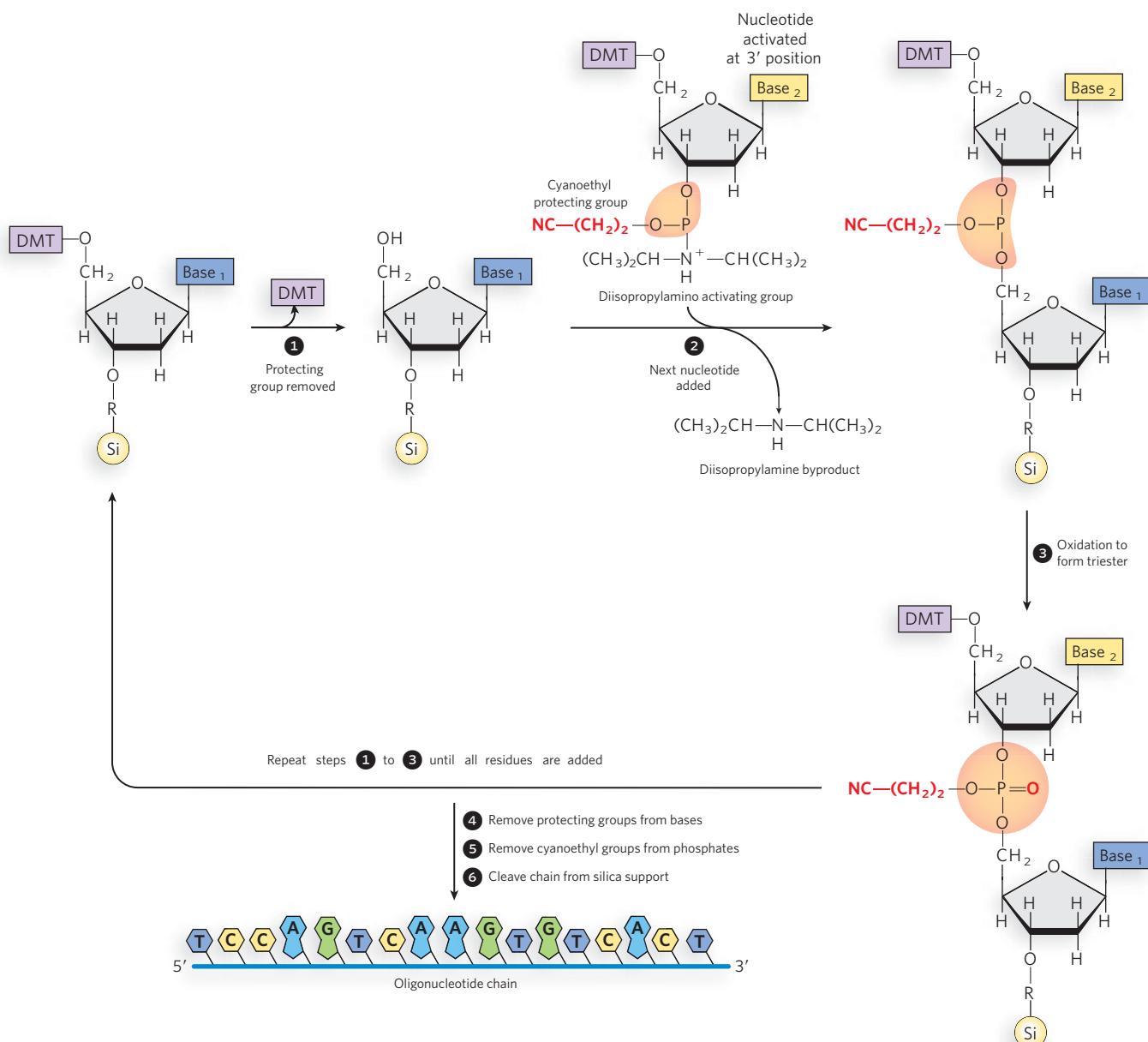


FIGURE 6-35 Solid-phase synthesis of nucleic acids. The oligonucleotide is synthesized in the 3' → 5' direction, starting with a single nucleotide that is covalently attached to a solid support, such as a glass bead. In a repeated series

of chemical reactions, nucleotides that have been protected from aberrant reaction with suitable functional groups are reacted to produce a new phosphodiester linkage. See text for details.

SECTION 6.4 SUMMARY

- Native DNA undergoes reversible unwinding and separation of the strands (melting) on heating or at extremes of pH. DNAs rich in G≡C base pairs have higher melting points than DNAs rich in A=T base pairs.
- Hybridization, or base pairing between two strands of nucleic acid from different sources, is the basis for important techniques used to study and isolate specific genes and RNAs.
- Southern blotting is a method by which a specific DNA sequence can be identified in a mixture,

following size-based fractionation of the DNA sample by agarose gel electrophoresis. A probe complementary to the DNA of interest is labeled with a radioactive or fluorescent functional group. After transferring the size-fractionated DNA from the gel to a membrane, the probe is hybridized to the sample to visualize the sequence of interest.

- Northern blotting, analogous to Southern blotting, is used for detecting specific RNA sequences.
- Mutations are alterations in DNA structure that produce permanent changes in the encoded genetic information. Deamination of cytosine is a common chemical mutation in DNA that can damage the genetic code if not corrected by the cell. Deamination of viral nucleic acid can be used to defend against viral infection.
- Select A and C residues in DNA are often enzymatically methylated after DNA synthesis. *E. coli* uses methylation to distinguish between host and foreign DNA and to facilitate the repair of mismatched base pairs that arise from replication errors. In eukaryotes, DNA methylation often inhibits gene expression.
- RNA can be chemically modified by enzymes that introduce methyl or acetyl groups at specific sites or alter a nucleotide base in other ways. These modifications may stabilize RNA structures and can also influence RNA recognition by proteins.
- DNA and RNA polymers of any sequence can be synthesized with simple, automated procedures involving chemical and enzymatic methods. Solid-phase synthesis of DNA and RNA occurs in the 3' → 5' direction, using chemically protected nucleotides that are selectively deprotected and coupled to the growing polynucleotide chain in successive cycles.

Unanswered Questions

Although many details of nucleic acid structure are well understood, future challenges involve linking the chemistry of these molecules to their behavior in biological systems. Here are several interesting questions in the field.

1. **What are the functions of noncanonical DNA structures in cells?** We don't yet know whether non-B-DNA functions in specific cellular processes. For example, some evidence suggests that with its left-handed twist, Z-DNA relieves some of the torsional strain that would otherwise build up during DNA transcription. Perhaps for this reason, the potential to form Z-DNA structures correlates with genomic regions of active transcription. But definitive proof of these ideas has been elusive. Whether three-stranded or four-stranded structures are biologically relevant is also a topic that remains ripe for experimentation.
2. **Do mRNAs have stable three-dimensional structures?** Although mRNAs were once thought to be spaghetti-like molecules, increasing evidence hints that they may have stable structures that contribute to biological function. For example, many mRNAs include long sequences that extend beyond the coding region of the gene and are critical for proper gene regulation. Specific proteins bind to these regions and probably recognize structures within them.
3. **How widespread is chemical modification of RNA?** Modified nucleotides in tRNA and rRNA have been recognized for a long time, but we do not know whether other RNAs in cells contain such chemical changes. This is an important question, because modifications could influence the function of RNAs that play various roles in controlling gene expression and therefore might be relevant to understanding disease pathways.

How We Know

DNA Is a Double Helix

Sayre, A. 1975. *Rosalind Franklin and DNA*. New York: W.W. Norton & Co.

Watson, J.D. 1968. *The Double Helix: A Personal Account of the Discovery of the Structure of DNA*. New York: Atheneum.

Watson, J.D., and F.H. Crick. 1953. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature* 171:737–738.

By the early 1950s, DNA had been confirmed as the genetic material in cells, but its structure was unknown. Given that structural information would be key to understanding heredity, the race was on to solve the mystery of DNA structure. Researchers including Rosalind Franklin, Raymond Gosling, and Maurice Wilkins were measuring x-ray diffraction of DNA fibers generated by drawing them from solution using a glass rod. The diffraction patterns reflected the symmetry of the DNA molecules in the fibers, thereby providing an important clue to their overall arrangement. By examining the diffraction pattern from fibers oriented perpendicular to the x-ray beam, investigators could deduce the helical symmetry of the molecules inside. These data were interpreted in the light of Chargaff's rules, which state that in a given DNA sample, the fraction of A equals the fraction of T, and the fraction of C equals the fraction of G. Possible models of the DNA in the fiber were produced, and their calculated diffraction patterns were compared with the experimentally derived patterns.

Rosalind Franklin's famous Photograph 51 revealed a particularly well-resolved x-ray diffraction pattern of a DNA fiber that was interpreted to determine the 3.4 Å distance between base pairs and the 34 Å periodicity of the helix (characteristic of B-form DNA; see Section 6.2) (Figure 1). Watson and Crick made extensive use of this

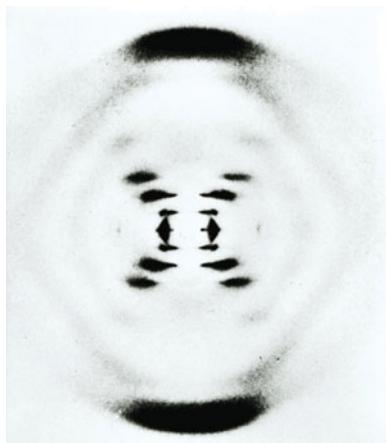


FIGURE 1 Franklin's "Photograph 51" provided the information necessary to solve the double-helical structure of DNA. [Source: Omikron/Photo Researchers.]

image, along with related diffraction data, to develop a model of the three-dimensional structure of DNA that proved to be correct (Figure 2). This was done at a time before sophisticated computer modeling was possible: Watson and Crick presented their work as a physical model of the double helix constructed on a wire support! Unlike other competing models of DNA, the Watson-Crick structure had the sugar-phosphate backbone winding around the outside of the helix, with the bases pointing to the interior, where they formed base-pairing interactions between the two strands.

Watson and Crick's work was published in a letter to the British journal *Nature* in 1953. In the same issue, several other papers provided experimental support for the Watson-Crick model. The double-helical structure immediately suggested a mechanism by which DNA strands could be faithfully copied from one generation to the next. In a famously understated final sentence of their paper, Watson and Crick wrote: "It has not escaped our notice" that the specificity of base pairing could ensure accurate DNA replication.

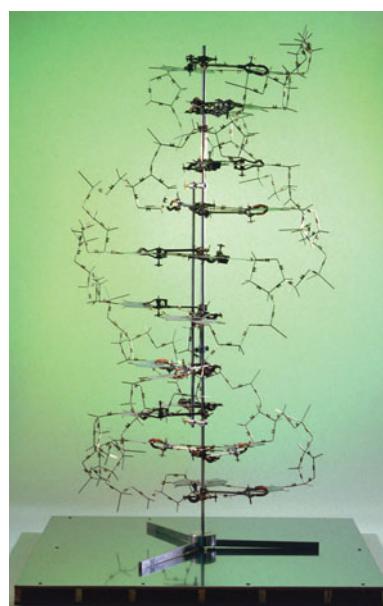


FIGURE 2 A replica of the DNA model built by Watson and Crick. The original model is on display in the Science Museum in London. [Source: Science Museum/SSPL.]

DNA Helices Have Unique Geometries That Depend on Their Sequence

Wing, R., H. Drew, T. Takano, C. Broka, S. Tanaka, K. Itakura, and R. E. Dickerson. 1980.

Crystal structure analysis of a complete turn of B-DNA. *Nature* 287:755–758.

The discovery of the DNA double helix marked the dawn of molecular biology. However, it was not until 1980 that the first single crystal of a DNA molecule was obtained. This was an important landmark in its own right, because for the first time it became possible to determine the exact helical parameters of a defined DNA sequence. Why did it take almost 30 years after the work of Watson, Crick, Franklin, Wilkins, and Gosling for specific DNA sequences to be crystallized?

The answer is technology. Until the late 1970s, it wasn't possible to synthesize DNA molecules in the laboratory, so investigators could not produce enough of a specific sequence to make growth of single crystals feasible. Once the methodology was available to synthesize DNA oligonucleotides on solid supports, short DNA molecules of specific length and sequence could be produced in milligram quantities. This material could be purified, and it crystallized readily when concentrated slowly in the presence of suitable buffers. Single crystals of DNA offered some distinct advantages over the DNA fibers analyzed by Franklin, Gosling, and Wilkins. DNA fibers can readily form from a mixture of DNAs of different lengths and sequences, but the structures obtained by analyzing the fiber diffraction patterns produce an "averaged" structure of all the molecules in the fiber. In contrast, single crystals, by definition, are formed by arrays of identical molecules.

Richard Dickerson and his colleagues recognized the wealth of information to be gained by solving a molecular structure of single DNA crystals. They used a self-complementary dodecamer sequence, CGC GAATTCGCG, to solve the first single-crystal structure of DNA. The overall double-helical structure agreed well with that determined by Watson and Crick, but

many new details about the geometry of the helix were revealed (Figure 3). This structure, known as the Dickerson dodecamer, ushered in an era of high-resolution structural determinations of DNA, and eventually to crystallographically determined structures of specific DNAs bound to protein partners. The study of individual DNA sequences also led to extensive studies of DNA-small molecule interactions and to research on the effects of DNA mutations on helical geometry. This work guided the development of certain anticancer drugs, such as *cis*-platin, that bind and distort DNA and thereby disrupt its replication in rapidly growing cells.

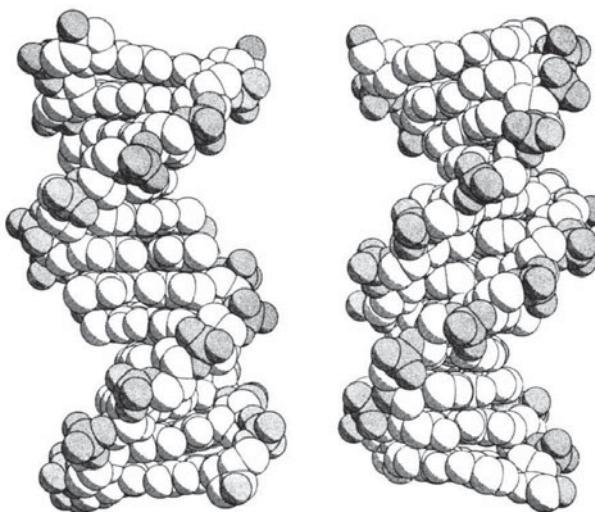


FIGURE 3 The Dickerson dodecamer structure revealed, for the first time, the details of helical geometry for a specific DNA sequence. The drawings are oriented to show the major groove (left) and minor groove (right) in the B-DNA helix. [Source: R. Wing et al., *Nature* 287:755–758, 1980, Fig. 4.]

Ribosomal RNA Sequence Comparisons Provided the First Hints of the Structural Richness of RNA

Gutell, R.R., N. Larsen, and C.R. Woese. 1994. Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiol. Rev.* 58:10–26.

Ribosomes have been around for a long time, and the sequences of the RNAs they contain have been constrained over the course of evolution by the requirements of making functional ribosomes to catalyze protein synthesis. For this reason, Carl Woese recognized in the 1960s that comparing ribosomal RNA sequences would provide valuable information about the evolutionary relationships among different organisms. Working over many years, he and his colleague Harry Noller assembled careful alignments of the 16S and 23S rRNAs from a large number of microbes. This work led Woese to propose the three-domain theory of life—Eubacteria (now classified simply as Bacteria), Archaeabacteria (now Archaea), and Eukarya (eukaryotes).

The comparative analysis approach begun by Woese and Noller was continued by Robin Gutell, who expanded the comparison to include 16S and 23S rRNA sequences from multicellular organisms, including humans. Gutell's critical analysis provided the first hints that these RNAs formed specific three-dimensional structures important to their function. One of the key insights from comparative rRNA sequence analysis was the discovery of noncanonical (i.e., non-Watson-Crick) base pairings. Although these had already been observed in tRNA structures, the much larger sizes of rRNA sequences provided vastly more data. For both 16S and 23S rRNAs, much of the sequence could be folded up into base-paired segments. Comparisons between species showed that the base pairings were much more conserved than were the actual nucleotide

sequences. This was because a change in the identity of a nucleotide on one side of a base-paired stretch was typically matched by a mutation in its base-pair partner such that base pairing was maintained. Woese and Gutell also noticed that in many cases, such compensatory base changes occurred for base “pairs” not previously thought to form, such as G-U, A-A, and G-A (see Figure 6-24). In this way, long before high-resolution structures of large RNAs became available, it was clear that RNA molecules are much more tolerant of non-Watson-Crick base pairings than is DNA (Figure 4).

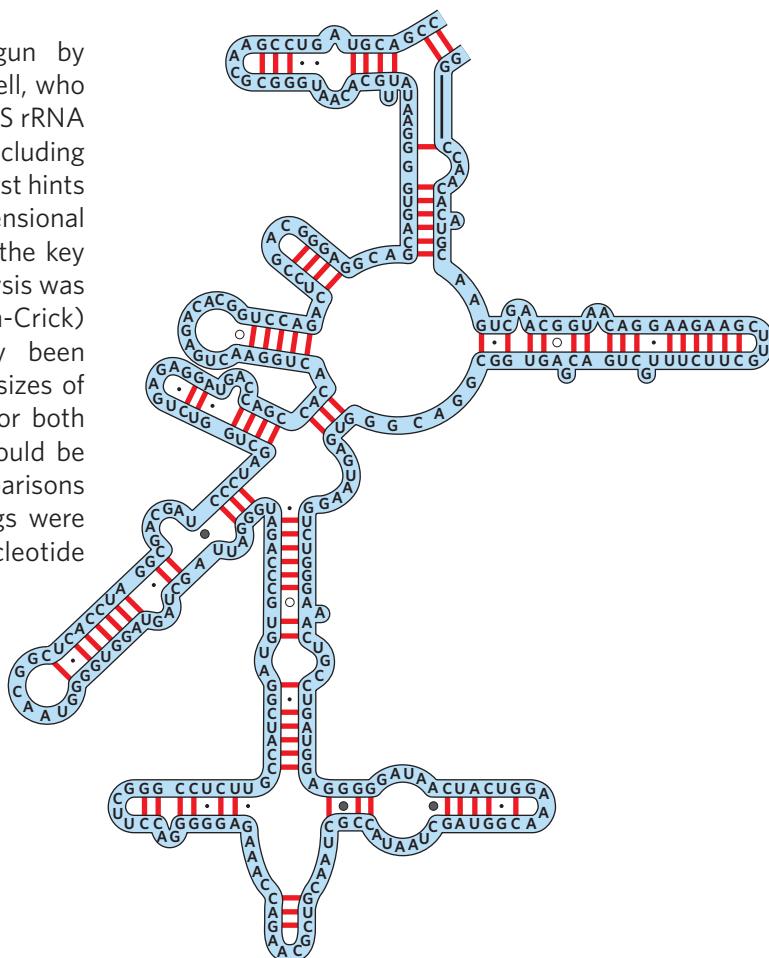


FIGURE 4 This model of secondary structure for *E. coli* 16S rRNA shows canonical base pairs connected by lines, G-U pairs connected by dots, A-G pairs connected by open circles, and other noncanonical pairings connected by solid circles. [Source: J. J. Cannone et al., The Comparative RNA Web (CRW) Site: An Online Database of Comparative Sequence, 2002.]

Key Terms

nucleotide, p. 177	base pairing, p. 181	annealing, p. 200
adenine (A), p. 178	Chargaff's rules, p. 185	hypochromic effect, p. 201
guanine (G), p. 178	antiparallel, p. 186	hyperchromic effect, p. 201
cytosine (C), p. 178	B-form DNA (B-DNA), p. 188	melting point (T_m), p. 201
thymine (T), p. 178	A-form DNA (A-DNA), p. 188	hybrid duplex, p. 202
uracil (U), p. 178	Z-form DNA (Z-DNA), p. 188	probe, p. 202
deoxyribonucleotide, p. 178	palindrome, p. 190	gel electrophoresis, p. 203
ribonucleotide, p. 178	hairpin, p. 191	Southern blotting, p. 203
phosphodiester bond, p. 178	RNA secondary structure, p. 196	Northern blotting, p. 203
oligonucleotide, p. 180	denaturation, p. 200	
polynucleotide, p. 180	melting, p. 200	

Problems

- In the 1980s, Tom Cech and Sidney Altman discovered that RNA could function as an enzyme. List two properties of enzymes that must hold true for this characterization to be correct.
- What positions in the ring of a purine nucleotide in DNA have the potential to form hydrogen bonds but are not involved in Watson-Crick base pairing?
- During his studies of the base composition of DNAs from various species, Erwin Chargaff obtained the following data for several human samples. The data show ratios of moles of each base to moles of phosphate in samples from various tissue types. Note that the error in the molar ratios is about ± 0.03 .

	Sperm 1	Sperm 2	Thymus	Liver Carcinoma
Adenine	0.29	0.27	0.28	0.27
Guanine	0.18	0.17	0.19	0.18
Cytosine	0.18	0.18	0.16	0.15
Thymine	0.31	0.30	0.28	0.27
Recovery	0.96	0.92	0.91	0.87

What can you conclude from these data?

- A part of one strand of a double-helical DNA molecule has the sequence: 5'-GATTACAGCCTTAGTTAAATTC TAAGGCTGGTA-3'.
- (a) Write out the sequence of the complementary strand of DNA.
(b) Does this strand have the potential to form any kind of alternative DNA structure? Does the double-stranded DNA of which it is a part have the potential to form an alternative structure? If so, what structure or structures might form?
- A double-stranded DNA oligonucleotide in which one of the strands has the sequence 5'-TAATACGACTCAC

TATAGGG-3' has a melting temperature (T_m) of 59°C. If a double-stranded RNA oligonucleotide of identical sequence (substituting U for T) is constructed, will its T_m be higher or lower?

- If the DNA and RNA oligonucleotides of Problem 5 are both present in an aqueous solution near neutral pH, how will their structures differ (apart from the presence of U vs. T in RNA vs. DNA)?
- Why does RNA contain uracil rather than thymine?
- Part of a chromosome has the sequence (on one strand): 5'-ATTGCATCCCGCGGTGCGCGCGATCCCCTTACTTTCCG-3'. Underline the part of this sequence that is most likely to take up the Z conformation.
- Why does DNA form a double-helical structure?
- Why are DNA and RNA considered acids?
- The cells of many eukaryotic organisms have highly specialized systems that repair G-T mismatches in DNA. The mismatch is always converted to a G≡C, never to an A=T base pair. This G-T mismatch repair system occurs in addition to a more general repair system that fixes virtually all types of mismatches. Suggest why cells might require a specialized G-T mismatch repair system, and why cells would specifically convert the G-T to G≡C.
- If a tRNA sequence were synthesized as DNA instead of RNA, could it function in protein synthesis? Why or why not?
- What dictates the strength of association between two DNA strands?
- What sequence characteristics would you expect for regions of a chromosome that encode highly structured RNA molecules?
- Why were single crystals critical to determining the molecular structure of DNA? (Hint: See How We Know.)

Data Analysis Problem

Hershey, A.D., and M. Chase. 1952. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J. Gen. Physiol.* 36:39–56.

- 16.** The stage occupied by Watson and Crick in 1953 was set by many other scientists, primarily Alfred Hershey and Martha Chase (see Chapter 2, How We Know). By 1952, a number of experiments had pointed to DNA as the genetic material, but much controversy remained, and protein was still regarded as a prime candidate. The Hershey and Chase experiments published in 1952 are credited with eliminating any remaining doubt in the scientific community that DNA alone was the genetic material.

The experiments made use of *E. coli* and one of its viruses, bacteriophage T2. When the experiments began, it was known that T2 consisted of a DNA core surrounded by a protein coat. T2 attached itself to the bacterial host and injected its core material into the host cell. New copies of T2 were made within the host, and the bacterial cell was lytically destroyed when the new T2 copies were released. Although it was clear that the T2 protein coat remained outside the bacterial cell, many workers thought that both DNA and some protein were introduced into the host when the genetic material was injected by T2, or that the attached protein shells played a role in the production of the T2 progeny. In either case, it was possible that protein was the genetic material.

Prior to their famous blender experiment (described shortly), Hershey and Chase carried out a series of experiments to better establish what parts of the T2 bacteriophage were introduced into the bacterial cells. They knew that T2 could be inactivated by osmotic shock, leaving behind T2 “ghosts” consisting of the viral coats bereft of their internal contents (which were released into the medium). Two different batches of T2 bacteriophage were grown, one labeled with ^{35}S and the other with ^{32}P . Half of each preparation was subjected to osmotic shock by incubation in 3 M NaCl followed by rapid dilution into distilled water. The authors described preparations thus treated as “plasmolyzed.” Each of the four resulting preparations was further subjected to four additional treatments.

(1) Addition of acid, followed by centrifugation; the supernatant was monitored for acid-soluble radioactivity. (2) Treatment with DNase (an enzyme that reduces DNA to nucleotides), followed by the addition of acid and centrifugation; acid-soluble radioactivity was again determined. (3) Addition of bacterial host cells, followed by centrifugation and determination of radioactivity in the bacterial cell pellet. And (4) treatment with T2 immune sera (antibodies), followed by centrifugation of the immune-precipitated material. The results are shown in Table 1.

- What macromolecules are labeled by ^{35}S ? What macromolecules are labeled by ^{32}P ?
- Why didn't the researchers use other common radioactive labels that were available at the time (^{14}C and ^{3}H)?
- Little or no radioactivity appears in the first acid-soluble supernatant in any of the preparations. Explain.
- Most of the ^{32}P label, but little of the ^{35}S , appears in the acid-soluble supernatant after DNase treatment. Explain.
- In the samples not treated with osmotic shock, most of both labels was found adsorbed to bacteria (in the cell pellet), but only the ^{35}S label appeared in the cell pellet in the plasmolyzed samples. Explain.
- In the control samples, both labels were precipitated by the T2 antibodies, but only the ^{35}S was precipitated in the plasmolyzed samples. Explain.
- What general conclusions can you draw from these experiments?

Several additional experiments, combined with those above, established that the bulk of the phage DNA was introduced into the cells during infection and thus might contribute to the production of progeny. About 30% of the ^{32}P label ended up in the progeny phage. But what happened to the phage protein? This consideration led to the blender experiment. Electron micrographs had previously shown that T2 attaches to the outside of bacterial cells and is attached to them by a long tail. The researchers reasoned that this “precarious attachment” could be eliminated by shearing.

Table 1 Composition (%) of T2 Bacteriophage Ghosts and Solution of Plasmolyzed Phage

Preparation	Whole Phage Labeled with:		Plasmolyzed Phage Labeled with:	
	$^{32}\text{P}(\%)$	$^{35}\text{S}(\%)$	$^{32}\text{P}(\%)$	$^{35}\text{S}(\%)$
1. Acid-soluble	—	—	1	—
2. Acid-soluble after treatment with DNase	1	1	80	1
3. Adsorbed to sensitive bacteria	85	90	2	90
4. Precipitated by T2 antibodies	90	99	5	97

Table 2 Effect of Multiplicity of Infection on Elution of Phage Membranes from Infected Bacteria

Running Time in Blender (min)	Multiplicity of Infection	32P-labeled Phage		35S-labeled Phage	
		Isotope Eluted (%)	Infected Bacteria Surviving (%)	Isotope Eluted (%)	Infected Bacteria Surviving (%)
0	0.6	10	120	16	101
2.5	0.6	21	82	81	78
0	6.0	13	89	46	90
2.5	6.0	24	86	82	85

Bacteria were infected with either ^{32}P - or ^{35}S -labeled phage. Two different multiplicities of infection (number of phage added per cell) were used, 0.6 and 6.0. After a few minutes, the cells were centrifuged and resuspended in fresh media. Some were subjected to a 2.5-minute treatment in a blender. Samples were then taken. For one sample, the cells were centrifuged and the supernatant measured to determine how much isotope had been removed from the cells. Another sample was left to produce phage progeny and titrated to determine what fraction of the infected cells were producing phage. The results are shown in Table 2.

- (h) Why did the investigators centrifuge and then resuspend the phage after infection?
- (i) How much of the ^{35}S label is stripped from the cells by the blender?
- (j) What general conclusions can you draw from this experiment?

A careful subsequent examination of the progeny phage indicated that less than 1% of the ^{35}S label ended up in the progeny. Combining this with the other data, the researchers could make a clear case that DNA was the genetic material guiding the production of phage progeny.

Additional Reading

The Structure and Properties of Nucleotides

Frank-Kamenetskii, M.D., and S.M. Mirkin. 1995. Triplex DNA structures. *Annu. Rev. Biochem.* 64:65–95.

Friedberg, E.C., G.C. Walker, and W. Siede. 1995. *DNA Repair and Mutagenesis*. New York: W.H. Freeman and Company. A good source for more information on the chemistry of nucleotides and nucleic acids.

Keniry, M.A. 2000. Quadruplex structures in nucleic acids. *Biopolymers* 56:123–146. A good summary of the structural properties of quadruplexes.

Rich, A., and S. Zhang. 2003. Timeline: Z-DNA—The long road to biological function. *Nat. Rev. Genet.* 4:566–572.

Shafer, R.H. 1998. Stability and structure of model DNA triplexes and quadruplexes and their interactions with small ligands. *Prog. Nucleic Acid Res. Mol. Biol.* 59:55–94.

Wells, R.D. 1988. Unusual DNA structures. *J. Biol. Chem.* 263:1095–1098. Minireview, presenting a concise summary.

DNA Structure

Collins, A.R. 1999. Oxidative DNA damage, antioxidants, and cancer. *Bioessays* 21:238–246.

Kornberg, A., and T.A. Baker. 2005. *DNA Replication*, 2nd ed. New York: W.H. Freeman and Company. The best place to start to learn more about DNA structure.

Marnett, L.J., and J.P. Plastaras. 2001. Endogenous DNA damage and mutation. *Trends Genet.* 17:214–221.

Olby, R.C. 1994. *The Path to the Double Helix: The Discovery of DNA*. New York: Dover Publications.

Sayre, A. 1978. *Rosalind Franklin and DNA*. New York: W.W. Norton & Co.

Sinden, R.R. 1994. *DNA Structure and Function*. San Diego: Academic Press. A fine discussion of many topics covered in this chapter.

Watson, J.D. 1968. *The Double Helix: A Personal Account of the Discovery of the Structure of DNA*. New York: Atheneum (paperback edition, New York: Simon & Schuster/Touchstone Books, 2001).

RNA Structure

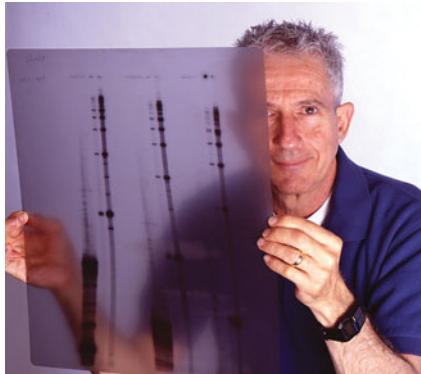
Chang, K.Y., and G. Varani. 1997. Nucleic acids structure and recognition. *Nat. Struct. Biol.* 4(Suppl.):854–858. A description of the application of NMR (nuclear magnetic resonance) to the determination of nucleic acid structure.

Hecht, S.M., ed. 1996. *Bioorganic Chemistry: Nucleic Acids*. Oxford: Oxford University Press. A very useful text.

Moore, P.B. 1999. Structural motifs in RNA. *Annu. Rev. Biochem.* 68:287–300.

Saenger, W. 1984. *Principles of Nucleic Acid Structure*. New York: Springer-Verlag. A somewhat dated but classic text, containing one of the best in-depth compilations of nucleic acid structural data available.

Studying Genes



Norman Arnheim [Source: Courtesy of University of Southern California.]

My students borrowed some lambda phage DNA size markers from my colleague Ken Marcu to use in their Southern blotting experiments with the human and mouse DNA. Soon my students began telling me that a particular probe sequence from the nontranscribed region of a mouse ribosomal gene was lighting up a specific lambda phage DNA fragment. I can still see myself standing in the lab, looking at my students' Southern blots showing the lambda fragment—but knowing that it couldn't be lambda DNA that was complementary to the mouse sequence! So what was going on?

Ken told me he had modified the size marker for his work by adding a cloned sequence fragment from a mouse immunoglobulin gene, which immediately set us off in an interesting and unexpected new direction. Using a method called *in situ* hybridization, in which DNA probes are hybridized to intact chromosomes, we soon discovered that the mouse ribosomal DNA fragment was complementary to sequences on all the mouse chromosomes! So whatever this sequence was, it was interspersed throughout mouse genomic DNA. We ultimately found that this sequence was everywhere in human, bird, and amphibian DNA. We sequenced one of the complementary fragments and discovered that the sequence contained 17 tandem CA repeats. Although the investigation of CA repeat function is still ongoing, the repeats were used extensively for years in the genetic mapping of human diseases and for forensic purposes. I love the serendipity of science!

—**Norman Arnheim**, on the 1980 discovery of interspersed CA repeats in genomic DNA

Moment of Discovery

My lab at SUNY Stony Brook studied genetic variation in mouse ribosomal genes in the days before genomic sequencing or even the ability to readily synthesize DNA oligonucleotides. Using Ed Southern's method for hybridizing short DNA fragments to complementary sequences in genomic DNA samples, we made restriction enzyme maps of mouse and human ribosomal DNA.

7.1 Isolating Genes for Study (Cloning) 216

7.2 Working with Genes and Their Products 226

7.3 Understanding the Functions of Genes and Their Products 242

The set of methods encompassed by biotechnology is fundamental to the advancement of modern biology, defining present and future frontiers and illustrating many important principles of the life sciences. With the ability to elucidate the laws governing enzyme catalysis, macromolecular structure, cellular metabolism, and information pathways, researchers can focus on the mechanisms of increasingly complex processes. Cell division, immunity, embryogenesis, sensory perception, oncogenesis, cognition—all are orchestrated in an elaborate symphony of molecular and macromolecular interactions that we are beginning to understand with increasing clarity. The real implications of the scientific journey begun in the nineteenth century are found in the ever-increasing power to analyze and alter living systems.

A major source of molecular insights about complex biological processes is the cell's own information archive: its DNA. The complement of genetic information in a cell—one complete copy of the information required to specify that organism—is known as the organism's **genome**. A molecular biologist is often interested in the function of one or a few genes in a genome. The sheer size of the genome, however, presents an enormous challenge: how does one find and study a particular gene among the tens of thousands of genes nested in the billions of base pairs of a mammalian genome? Solutions began to emerge in the 1970s.

Decades of advances by scientists working worldwide in genetics, biochemistry, cell biology, and physical chemistry came together in the laboratories of Paul Berg, Herbert Boyer, and Stanley Cohen to yield techniques for locating, isolating, preparing, and studying small segments of DNA derived from much larger chromosomes. Techniques for DNA cloning paved the way to the modern fields of genomics, transcriptomics, and proteomics: the study of genes, mRNA transcripts, and proteins on the scale of whole cells and organisms (see Chapter 8). These new approaches are transforming basic research, agriculture, medicine, ecology, forensics, and many other fields, while occasionally presenting society with difficult choices and ethical dilemmas.

Every student and instructor, when considering the topics we present in this text, encounters two conflicting realities. First, the methods we describe were made possible by advances in molecular biology. Hence, one must understand the fundamental concepts of DNA replication, RNA transcription, protein synthesis, and gene regulation to appreciate how these methods work. But, the alternative reality recognizes that modern molecular biology relies on these same methods to such an extent that a current treatment of the subject becomes impossible without a proper introduction to the technology. By presenting these methods early in the book,



Clockwise from top left:

Paul Berg [Source: Courtesy of the National Library of Medicine.]

Herbert Boyer [Source: Courtesy of The Regent of University of California, San Francisco.]

Stanley Cohen [Source: Ted Streshinsky/Time Life Pictures/Getty Images.]

we acknowledge that they are inextricably interwoven with both the advances that gave rise to them and the newer discoveries they now make possible. The background we necessarily provide makes the discussion here not just an introduction to technology but also a preview of many of the fundamentals of molecular biology encountered in later chapters.

This chapter is not designed to consider every method that is relevant to molecular biology. Indeed, we'll be discussing many additional techniques in later chapters where they are uniquely relevant to a specific area of study. However, a particular set of methods—a set often associated with biotechnology—has become a critical engine of discovery driving the advances discussed throughout this text. Those methods now become our focus. We begin by outlining the principles of the now-classic discipline of DNA cloning, then illustrate the range of applications and the potential of many newer technologies that support and accelerate the advance of molecular biology. Finally, we take a look at some of the technologies that allow us to elucidate gene function.

7.1 Isolating Genes for Study (Cloning)

A **clone** is an identical copy. The term originally was applied to cells produced when a cell of a single type was isolated and allowed to reproduce to create a

population of identical cells. **DNA cloning** involves separating a specific gene or DNA segment from a larger chromosome, attaching it to a small molecule of carrier DNA, introducing this modified DNA into a host cell, then replicating the DNA by increasing both the cell number and the copy number of the cloned DNA in each cell. The result is selective amplification of a particular gene or DNA segment.

The cloning of DNA from any organism entails five general steps:

- 1. Cutting DNA at precise locations.** Enzymes called **restriction endonucleases** act as molecular scissors, recognizing specific sequences in DNA and cleaving the DNA into smaller fragments.
- 2. Selecting a small molecule of DNA capable of self-replication.** These small DNAs are called **cloning vectors** (a vector is a carrier or delivery agent). Most cloning vectors used in the laboratory are modified versions of naturally occurring small DNA molecules found in bacteria, or small viral DNAs.
- 3. Joining two DNA fragments covalently.** The enzyme **DNA ligase** links the cloning vector to the DNA fragment to be cloned. Composite DNA molecules of this type, comprising covalently linked segments from two or more sources, are called **recombinant DNAs**.
- 4. Moving recombinant DNA from the test tube to a host organism.** The host organism provides the enzymatic machinery for DNA replication.
- 5. Selecting or identifying host cells that contain recombinant DNA.** The cloning vector generally has

features that allow the host cells to survive in an environment where cells lacking the vector would die. Cells containing the vector are thus “selectable” in that environment.

The methods used for accomplishing these and related tasks are collectively referred to as **recombinant DNA technology** or, more informally, **genetic engineering**.

Much of our initial discussion focuses on DNA cloning in the bacterium *Escherichia coli*, the first organism used for recombinant DNA work and still the most common host cell. *E. coli* has many advantages. Its DNA metabolism (like many of its other biochemical processes) is well understood, many naturally occurring cloning vectors associated with *E. coli* are well characterized, and techniques are available for easily moving DNA from one bacterial cell to another. The principles discussed here are also broadly applicable to DNA cloning in other organisms, as we'll see later in the chapter.

Genes Are Cloned by Splicing Them into Cloning Vectors

DNA can be cloned from any cellular or viral source. Although the approaches are determined partly by the DNA source and what is known about it, all cloning efforts have a few enzymes and procedures in common. Recombinant DNA technology relies on a set of enzymes made available through decades of research on nucleic acid metabolism (Table 7-1). Two classes of enzymes are particularly important (Figure 7-1). First, restriction endonucleases recognize DNA at specific **recognition sequences** (or restriction sites) and cleave

Table 7-1 Some Enzymes Used in Recombinant DNA Technology

Enzyme	Function
Type II restriction endonuclease	Cleaves DNA at specific base sequences
DNA ligase	Joins two DNA molecules or fragments
DNA polymerase I (<i>E. coli</i>)	Fills single-stranded gaps in duplex DNA by stepwise addition of nucleotides to 3' ends
Reverse transcriptase	Makes a DNA copy of an RNA molecule
Polynucleotide kinase	Adds a phosphate to the 5'-OH end of a polynucleotide, to label it or permit ligation
Terminal transferase	Adds homopolymer tails to the 3'-OH ends of a linear duplex
Exonuclease III	Removes nucleotide residues from the 3' ends of a DNA strand
Bacteriophage λ exonuclease	Removes nucleotides from the 5' ends of a duplex to expose single-stranded 3' ends
Alkaline phosphatase	Removes terminal phosphates from the 5' end, the 3' end, or both

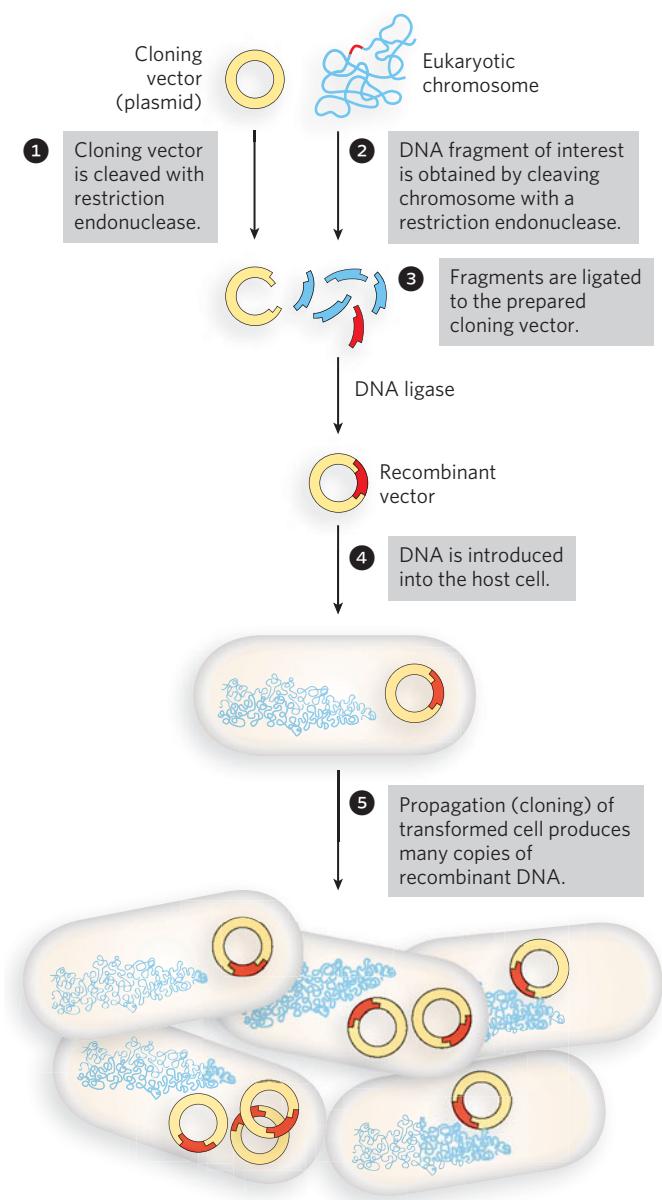


FIGURE 7-1 DNA cloning. The process involves cutting two DNAs with restriction enzymes, joining (ligating) the fragments together with DNA ligase, and using the recombinant DNA products to transform a suitable host cell. (This drawing is not to scale; the size of the *E. coli* chromosome relative to that of a typical cloning vector (such as a plasmid) is much greater than depicted here.)

it to generate a set of smaller fragments. Second, a DNA fragment of interest can be joined to the DNA of a suitable cloning vector by DNA ligase. The recombinant vector is then introduced into a host cell, which amplifies the DNA fragment in the course of many generations of cell division.

Restriction endonucleases are found in a wide range of bacterial species. Werner Arber discovered in the early 1960s that the biological function of these enzymes is to recognize and cleave foreign DNA (the DNA of an infecting virus, for example); such DNA is said to be *restricted*. Acting in a system with other enzymes that protect the host DNA, restriction endonucleases participate in a kind of immune system in bacteria. There are three types of restriction endonucleases, differing in their complexity and the typical distance between recognition sequence and cleavage site. **Type II restriction endonucleases**, first reported by Hamilton Smith in 1970, are the simplest, require no ATP for their activity, and cleave the DNA within the recognition sequence. Daniel Nathans quickly put this group of restriction endonucleases to use, demonstrating their extraordinary utility by developing novel methods for mapping and analyzing genes and genomes.

Thousands of restriction endonucleases have been discovered in different bacterial species, and more than 100 different DNA sequences are recognized by one or more of these enzymes. The recognition sequences are usually 4 to 8 base pairs (bp) long and palindromic (the recognition sequence, read in the 5'→3' direction, is the same on both strands of DNA). However, a few of them fall slightly outside this norm. Table 7-2 lists the sequences recognized by a few Type II restriction endonucleases.

Some restriction endonucleases make staggered cuts across the two DNA strands, leaving 2 to 4 nucleotides of one strand unpaired at each resulting end. Depending on which restriction enzyme is used, cleavage might occur such that the extended strand has either a 5' or a 3' end (called a 5' or 3' overhang). These unpaired strands are referred to as **sticky ends**, because they can base-pair with each other or with the complementary sticky ends of any other DNA fragments (Figure 7-2a). Other restriction endonucleases cleave both strands of DNA straight across, at the opposing phosphodiester bonds, leaving no unpaired bases on the ends, and thus produce what are often called **blunt ends** (Figure 7-2b).

The average size of the DNA fragments produced by cleaving genomic DNA with a restriction endonuclease depends on the frequency with which a particular recognition sequence occurs in the DNA molecule; this in turn depends largely on the length of the sequence. In a DNA molecule with a random sequence in which all four nucleotides are equally abundant, a 6 bp sequence recognized by a restriction endonuclease would occur, on average, once every 4^6 (4,096) bp. A 4 bp recognition sequence would occur much more often, about once every 4^4 (256) bp. In laboratory experiments, the fragment size can be increased by

Table 7-2 Recognition Sequences for Some Type II Restriction Endonucleases

BamHI	5'-GGATCC-3' CCTAGG * ↑	HindIII	5'-AAGCTT-3' TTCGAA ↑
ClaI	5'-ATCGAT-3' TAGCTA * ↑	NotI	5'-GC [*] GCCGC-3' CGCCGGCG ↑
EcoRI	5'-GAATTC-3' CTTAAG * ↑	PstI	5'-CTGCAG-3' GACGTC ↑ *
EcoRV	5'-GATATC-3' CTATAG ↑ ↓*	PvuII	5'-CAGCTG-3' GT [*] CGAC ↑
HaeIII	5'-GGCC-3' CCGG *↑	Tth111I	5'-GACNNNGTC-3' CTGNNNCAG ↑

Note: Arrows denote phosphodiester bonds cleaved by each restriction endonuclease. Asterisks mark bases that are methylated by the corresponding methyltransferase (where known). N denotes any base. Each enzyme name consists of a three-letter abbreviation of the bacterial species from which it is derived, sometimes followed by a strain designation and a Roman numeral to distinguish restriction endonucleases isolated from the same bacterial species or strain. Thus, BamHI is the first (I) restriction endonuclease characterized from *Bacillus amyloliquefaciens*, strain H.

terminating the reaction before completion—that is, before the enzyme molecules have had a chance to cleave every recognition sequence in the DNA sample. The result is a partial digest. Fragment size can also be increased by using a special class of endonucleases called homing endonucleases (Figure 13-24). These recognize and cleave much longer recognition sequences (12 to 40 bp).

Once a DNA molecule has been cleaved into fragments, a particular fragment of known size can be separated by agarose or acrylamide gel electrophoresis (see Chapter 6). For a typical mammalian genome, however, cleavage by a restriction endonuclease usually yields too many different DNA fragments to permit easy isolation of a particular fragment. As described later in this chapter, the amplification of a gene or DNA fragment by the polymerase chain reaction (PCR) provides an important alternative approach. PCR can be used to amplify a particular gene or genomic segment of interest in a form that makes its cloning and isolation simpler.

After the target DNA fragment is isolated, DNA ligase can be used to join it to a cloning vector. The ligation reaction is greatly facilitated if the ends to

be joined (ligated) have complementary sticky ends, as was apparent in the earliest recombinant DNA experiments (see How We Know). This is normally accomplished by cleaving the vector DNA with the same restriction enzyme used to prepare the target DNA fragments. DNA ligase catalyzes the formation of a phosphodiester bond between a 3' hydroxyl at the end of one DNA strand and a 5' phosphate at the end of another strand (see Figure 5-12).

Researchers can create new DNA sequences by inserting synthetic DNA fragments, called **linkers**, between the ends that are being ligated. Inserted DNA fragments with multiple recognition sequences for restriction endonucleases (often useful later in the experiment as points for inserting additional DNA by cleavage and ligation) are known as **polylinkers** (Figure 7-3).

Cloning Vectors Allow Amplification of Inserted DNA Segments

Genes or genomic segments are cloned for many different reasons. This is reflected in the use of a large variety of cloning vectors. The principles that govern the

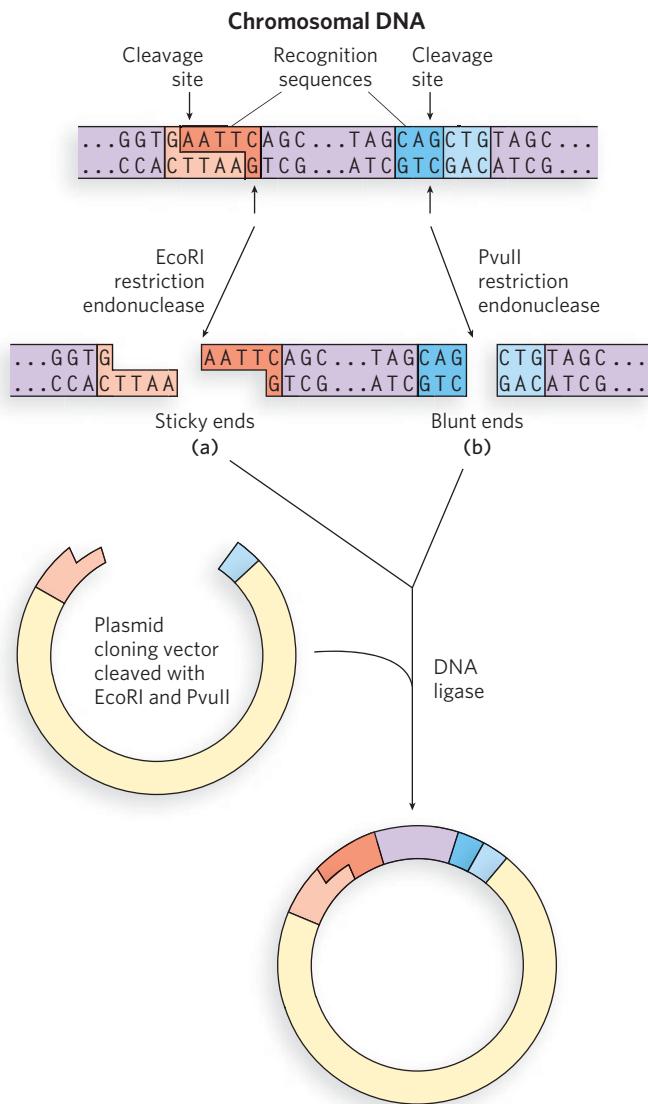


FIGURE 7-2 Cleavage of DNA molecules by restriction endonucleases. When Type II restriction endonucleases cleave DNA, they leave either (a) sticky ends (with protruding single strands) or (b) blunt ends. The restriction fragments can be ligated to other DNAs, such as the plasmid cloning vector shown here. Ligation is facilitated by the annealing of complementary sticky ends, and it is less efficient for DNA fragments with blunt ends than for those with complementary sticky ends. DNA fragments with noncomplementary sticky ends (i.e., those created by different restriction enzymes) generally are not ligated.

delivery of recombinant DNA in clonable form to a host cell, and its subsequent amplification in the host, are well illustrated by considering some popular cloning vectors used in experiments with *E. coli* and yeast—plasmids, bacterial artificial chromosomes, and yeast

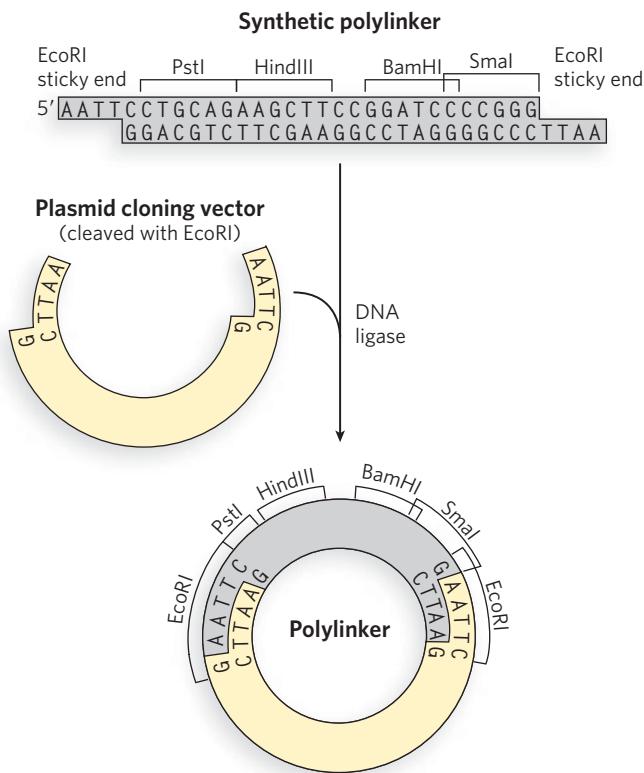


FIGURE 7-3 DNA polylinkers. A synthetic DNA fragment with recognition sequences for several restriction endonucleases—a fragment known as a polylinker—can be inserted into a plasmid that has been cleaved by a restriction endonuclease.

artificial chromosomes. Modern cloning vectors provide an array of options, allowing an investigator to tailor the cloning exercise to a particular goal: DNA sequencing, gene expression for protein purification, study of the effects of mutations, or creation of many kinds of gene alterations.

Plasmids A **plasmid** is a circular DNA molecule that replicates separately from the host chromosome. The wide variety of naturally occurring bacterial plasmids range in size from 5,000 to 400,000 bp. Many of the plasmids found in bacterial populations are little more than molecular parasites, similar to viruses but with a more limited capacity to transfer from one cell to another. To survive in the host cell, plasmids incorporate several specialized sequences that enable them to make use of the cell's resources for their own replication and gene expression.

Naturally occurring plasmids usually have a symbiotic role in the cell. They may provide genes that confer resistance to antibiotics or that perform new functions for

the cell. For example, the Ti plasmid of *Agrobacterium tumefaciens* allows the host bacterium to colonize the cells of plants and make use of the plant's resources. The same properties that enable plasmids to grow and survive in a bacterial or eukaryotic host are useful to molecular biologists who want to engineer a vector for cloning a specific DNA segment. The classic *E. coli* plasmid pBR322, constructed in 1977, is a good example of a plasmid with features useful in almost all cloning vectors (Figure 7-4):

1. The plasmid pBR322 has an **origin of replication**, or **ori**, a sequence where replication is initiated by cellular enzymes (see Chapter 11). This sequence is required to propagate the plasmid. An associated regulatory system is present that limits replication to maintain pBR322 at a level of 10 to 20 copies per cell.
2. The plasmid contains genes that confer resistance to the antibiotics tetracycline (Tet^R) and ampicillin (Amp^R), allowing the selection of cells that contain the intact plasmid or a recombinant version of the plasmid (discussed below).
3. Several unique recognition sequences in pBR322 are targets for restriction endonucleases (PstI , EcoRI , BamHI , Sall , and PvuII), providing sites where the plasmid can be cut to insert foreign DNA.

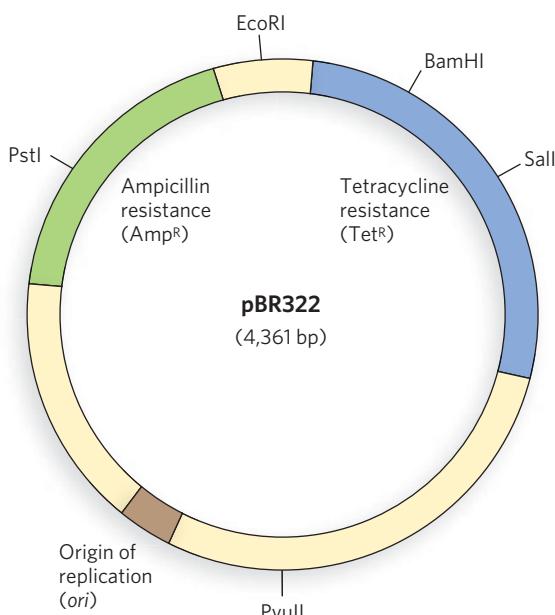


FIGURE 7-4 The constructed *E. coli* plasmid pBR322. This plasmid, one of the first to be constructed, was designed expressly for cloning in *E. coli*. See text for details.

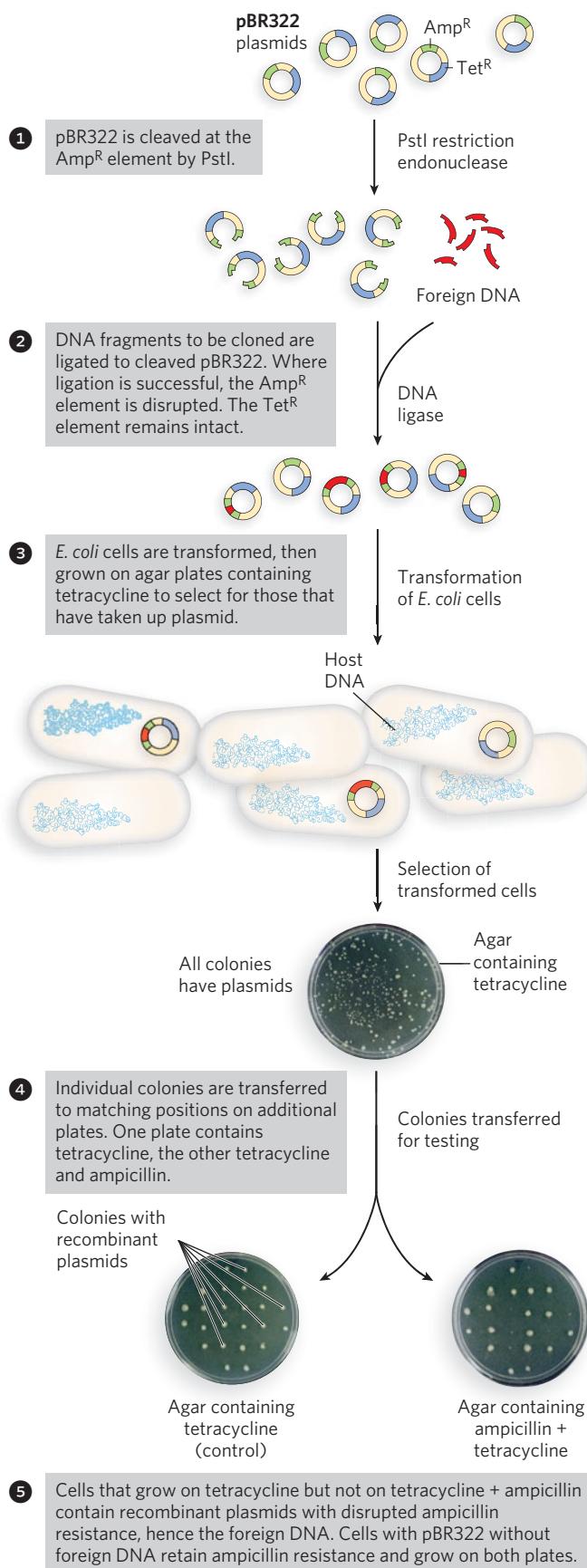
4. The small size of the plasmid (4,361 bp) facilitates its entry into cells and the biochemical manipulation of the DNA. This small size was generated simply by trimming away many DNA segments from a larger, parent plasmid—sequences that the molecular biologist does not need.

The replication origins inserted in common plasmid vectors were originally derived from naturally occurring plasmids. As in pBR322, each of these origins is regulated to maintain a particular number of plasmid copies in a cell (the plasmid copy number). Depending on the origin used, the plasmid copy number can vary from one to hundreds or thousands per cell, providing many options for investigators. Two different plasmids cannot function in the same cell if they use the same origin of replication, because the regulation of one will interfere with the replication of the other. Such plasmids are said to be incompatible. When a researcher wants to introduce two or more different plasmids into a bacterial cell, each plasmid must have a different replication origin.

In the laboratory, small plasmids can be introduced into bacterial cells by a process called **transformation**. The cells (often *E. coli*, but other bacterial species are also used) and plasmid DNA are incubated together at 0°C in a calcium chloride solution, then subjected to heat shock by rapidly shifting the temperature to between 37°C and 43°C. For reasons not well understood, some of the cells treated in this way take up the plasmid DNA. Some species of bacteria, such as *Acinetobacter baylyi*, are naturally competent for DNA uptake and do not require the calcium chloride–heat shock treatment. In an alternative method, cells incubated with the plasmid DNA are subjected to a high-voltage pulse. This approach, called **electroporation**, transiently renders the bacterial membrane permeable to large molecules.

Regardless of the approach, relatively few cells take up the plasmid DNA, so a method is needed to identify those that do. The usual strategy is to utilize one of two types of genes in the plasmid, referred to as **selectable** and **screenable** markers. **Selectable markers** either permit the growth of a cell (positive selection) or kill the cell (negative selection) under a defined set of conditions. The plasmid pBR322 provides examples of both positive and negative selection (Figure 7-5). A **screenable marker** is a gene encoding a protein that causes the cell to produce a colored or fluorescent molecule. Cells are not harmed when the gene is present, and the cells that carry the plasmid are easily identified by the colored or fluorescent colonies they produce.

Transformation of typical bacterial cells with purified DNA (never a very efficient process) becomes less



successful as plasmid size increases, and it is difficult to clone DNA segments longer than about 15,000 bp when plasmids are used as the vector.

To illustrate the use of a plasmid as a cloning vector, consider a typical bacterial gene, that encoding a recombinase called the RecA protein (see Chapter 14). In most bacteria, the gene encoding RecA is one of thousands of other genes on a chromosome millions of base pairs long. The *recA* gene is just over 1,000 bp long. A plasmid would be a good choice for cloning a gene of this size. As described later, the cloned gene can be altered in a variety of ways, and the gene variants can be expressed at high levels to enable purification of the encoded proteins.

Bacterial Artificial Chromosomes Large genome sequencing projects often require the cloning of much longer DNA segments than can typically be incorporated into standard plasmid cloning vectors such as pBR322. To meet this need, plasmid vectors have been developed with special features that allow the cloning of very long segments (typically 100,000 to 300,000 bp) of DNA. Once such large segments of cloned DNA have been added, these are large enough to be thought of as chromosomes, and are known as **bacterial artificial chromosomes**, or **BACs** (Figure 7-6).

A BAC vector is a relatively simple plasmid, generally not much larger than other plasmid vectors. To accommodate very long segments of cloned DNA, BAC vectors have stable origins of replication that maintain the plasmid at one or two copies per cell. The low copy number is useful in cloning large segments of DNA, because it limits the opportunities for unwanted recombination reactions that can unpredictably alter large cloned DNAs over time. BACs also include *par* genes, which encode proteins that direct the reliable distribution of the recombinant chromosomes to daughter cells at cell division, thereby increasing the likelihood of each daughter cell carrying one copy, even when few copies are present. The BAC vector includes both selectable and screenable markers. The BAC vector shown in Figure 7-6 contains a gene for resistance to the antibiotic chloramphenicol (Cm^R). Positive selection for vector-containing cells occurs on agar plates containing this antibiotic. A *lacZ* gene, required for production of the enzyme β -galactosidase, is a screenable marker that

FIGURE 7-5 Use of pBR322 to clone foreign DNA. The entire procedure is illustrated, including both positive and negative selection. [Photos courtesy of Elizabeth A. Wood, Department of Biochemistry, University of Wisconsin-Madison.]

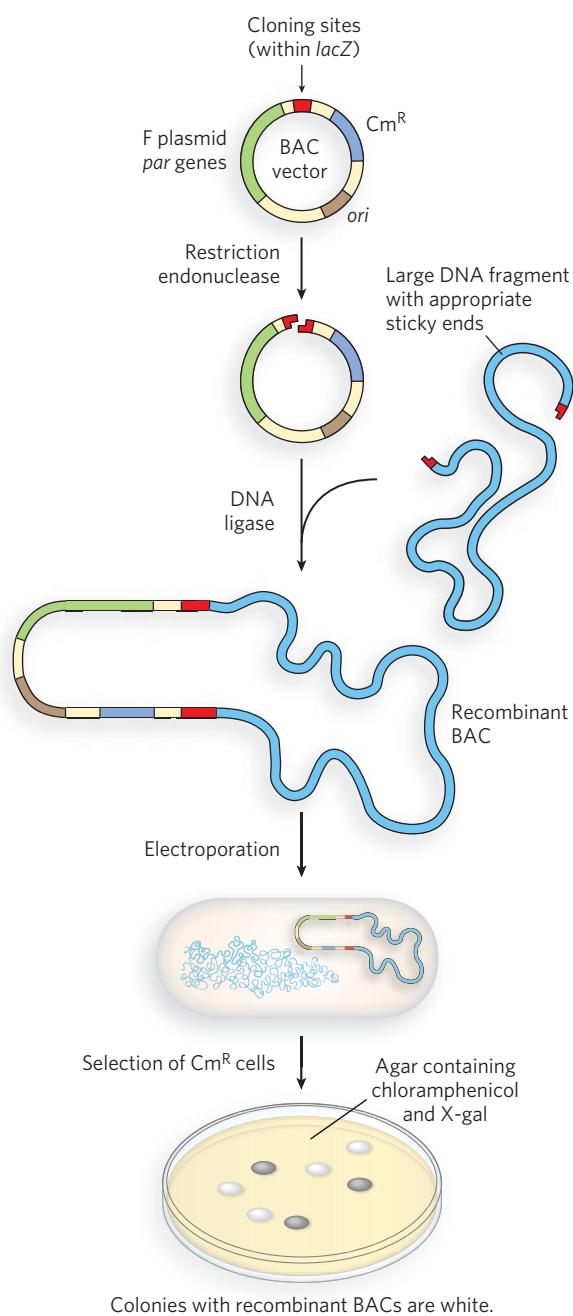


FIGURE 7-6 Bacterial artificial chromosomes (BACs) as cloning vectors. After treatment with an appropriate restriction endonuclease, a BAC and a long fragment of DNA are ligated. The recombinant BAC is transferred into *E. coli* by electroporation, and colonies with recombinant BACs are selected by growth on media containing both the antibiotic chloramphenicol and X-gal, the substrate for β -galactosidase that produces a colored product. See text for details.

can reveal which cells contain plasmids—now chromosomes—that incorporate the cloned DNA segments. The β -galactosidase catalyzes the conversion of the

colorless molecule 5-bromo-4-chloro-3-indolyl- β -D-galactopyranoside (X-gal) to a blue product. If the gene is intact and expressed, the colony containing it will be blue. If gene expression is disrupted by the introduction of a cloned DNA segment, the colony will be white.

Yeast Artificial Chromosomes As with *E. coli*, yeast genetics is a well-developed discipline. The genome of *Saccharomyces cerevisiae* contains only 14×10^6 bp (less than four times the size of the *E. coli* chromosome), and its entire sequence is known. Yeast is also very easy to maintain and grow on a large scale in the laboratory. Plasmid vectors have been constructed for yeast, employing the same principles that govern the use of *E. coli* vectors. Convenient methods are now available for moving DNA into and out of yeast cells, thus permitting the study of many aspects of eukaryotic cell biochemistry. Some recombinant plasmids incorporate multiple replication origins and other elements that allow them to be used in more than one species (e.g., in yeast and *E. coli*). Plasmids that can be propagated in cells of two or more species are called **shuttle vectors**.

Research on large genomes and the associated need for high-capacity cloning vectors led to the development of **yeast artificial chromosomes**, or **YACs** (Figure 7-7). YAC vectors contain all the elements needed to maintain a eukaryotic chromosome in the yeast nucleus: a yeast origin of replication, two selectable markers, and specialized sequences (derived from the telomeres and centromere) needed for stability and proper segregation of the chromosomes at cell division (see Chapter 9). In preparation for its use in cloning, the vector is propagated as a circular bacterial plasmid. Cleavage with a restriction endonuclease (BamHI in Figure 7-7) removes a length of DNA between two telomere sequences (TEL), leaving the telomeres at the ends of the linearized DNA. Cleavage at another internal site (by EcoRI in Figure 7-7) divides the vector into two DNA segments, referred to as vector arms, each with a different selectable marker.

The genomic DNA to be cloned is prepared by partial digestion with restriction endonucleases to obtain a suitable fragment size. Genomic fragments are then separated by **pulsed field gel electrophoresis**, a variation of gel electrophoresis that segregates very large DNA segments. DNA fragments of appropriate size (up to about 2×10^6 bp) are mixed with the prepared vector arms and ligated. The ligation mixture is then used to transform yeast cells (pretreated to partially degrade their cell walls) with these very large DNA molecules—which now have the structure and size to be considered yeast chromosomes. Culture on a medium that requires

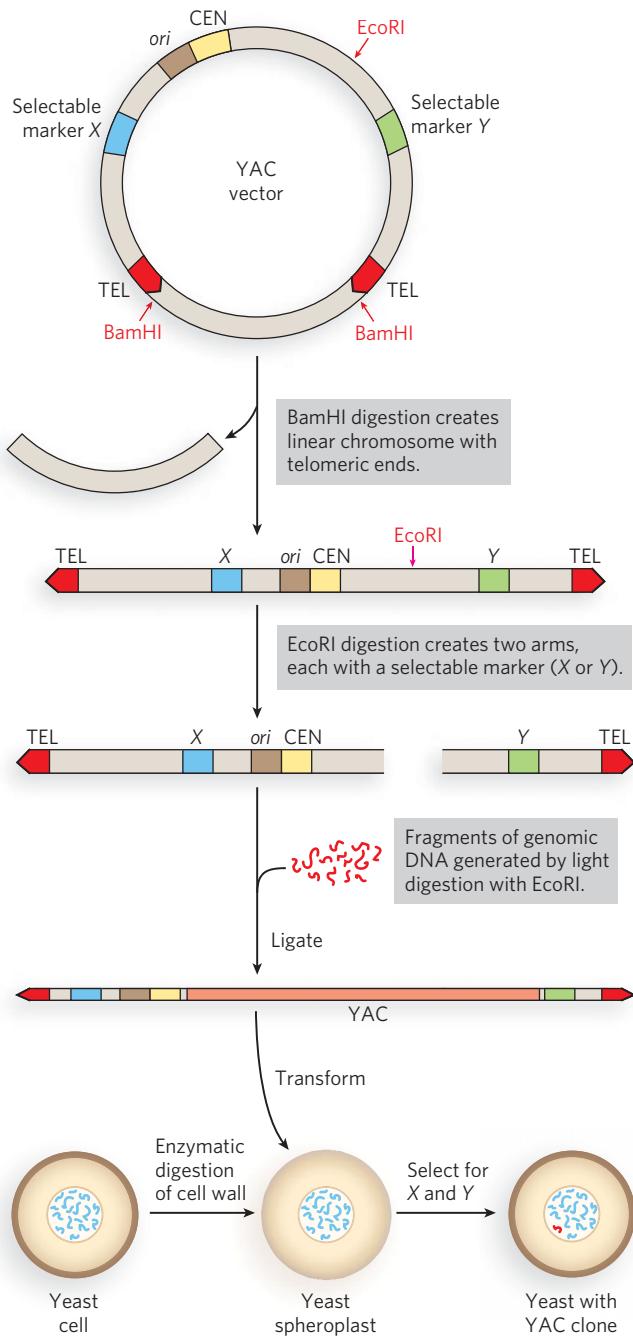


FIGURE 7-7 Construction of a yeast artificial chromosome (YAC).

A YAC vector includes an origin of replication (*ori*), a centromere (CEN), two telomeres (TEL), and selectable markers (here designated X and Y). Two separate DNA arms are generated by digestion with BamHI and EcoRI, each arm having a telomeric end and one selectable marker. A large DNA fragment, produced by EcoRI digestion, is ligated to the two arms, creating a YAC. The YAC is transferred into yeast cells (which have been prepared by removing the cell wall to form spheroplasts). The transformed cells are selected for X and Y, and the surviving cells propagate the DNA insert.

the presence of both selectable marker genes ensures the growth of only those yeast cells that contain an artificial chromosome with a large insert sandwiched between the two vector arms. The stability of YAC clones increases with the length of the cloned DNA segment (up to a point). Those with inserts of more than 150,000 bp are nearly as stable as normal cellular chromosomes, whereas those with inserts less than 100,000 bp long are gradually lost during mitosis (so, generally, there are no yeast cell clones carrying only the two vector ends ligated together or vectors with only short inserts). YACs that lack a telomere at either end are rapidly degraded.

As with BACs, YAC vectors can be used to clone very long segments of DNA. In addition, the DNA cloned in a YAC can be altered to study the function of specialized sequences in chromosome metabolism, mechanisms of gene regulation and expression, and many other problems in eukaryotic molecular biology.

DNA Libraries Provide Specialized Catalogs of Genetic Information

A **DNA library** is a collection of DNA clones, gathered together for purposes of genome sequencing, gene discovery, or determination of gene function. The library can take a variety of forms, depending on the source of the DNA and the ultimate purpose of the library.

One of the largest is a **genomic library**, produced when the complete genome of an organism is cleaved into thousands of fragments and *all* the fragments are cloned by insertion into a cloning vector. Building such a library traditionally has been a prelude to large sequencing projects. The first step is *partial* digestion of the DNA by restriction endonucleases, such that any given sequence will appear in fragments of a range of sizes—a range compatible with the cloning vector, ensuring that virtually all sequences are represented among the clones in the library. Fragments that are too large or too small for cloning are removed by centrifugation or electrophoresis. The cloning vector, such as a BAC or YAC, is cleaved with the same restriction endonuclease used to digest the DNA and ligated to the genomic DNA fragments. The ligated DNA mixture is then used to transform bacteria or yeast cells to produce a library of cells, each cell harboring a different recombinant DNA molecule. Ideally, all of the DNA in the genome under study is represented in the library. Each transformed bacterium or yeast cell grows into a colony, or clone, of identical cells, each cell bearing the same recombinant plasmid, one of

many represented in the overall library. In some of the newer sequencing technologies, the step of introducing the library DNA into cells is skipped and the genomic DNA fragments are sequenced directly (as described later in the chapter).

With the increasing availability of genome sequences, the utility of genomic libraries is diminishing, and investigators are building more specialized libraries for studying gene function. An example is a library that includes only those sequences of DNA that are *expressed*—that is, transcribed into RNA—in a given organism, or even just in certain cells or tissues. Such a library lacks the noncoding DNA that makes up a large portion of many eukaryotic genomes. The researcher first extracts mRNA from an organism, or from specific cells of an organism, and then prepares the **complementary DNAs (cDNAs)**. This multistep reaction, shown in **Figure 7-8**, relies on the enzyme reverse transcriptase, which synthesizes DNA from a template RNA (the enzyme is derived from a class of RNA viruses called retroviruses; see Chapter 14). The resulting double-stranded DNA fragments are inserted into a suitable vector and cloned, creating a population of clones called a **cDNA library**.

The search for a particular gene is made easier by focusing on a cDNA library generated from the mRNAs of a cell known to express that gene. For example, if we wished to clone globin genes, we could first generate a cDNA library from erythrocyte precursor cells, in which about half the mRNAs code for globins. A particular gene or gene segment in a library can be detected by the hybridization techniques introduced in Chapter 6. If a researcher knows something about the sequence of the DNA being sought, a short nucleic acid complementary to that sequence can be synthesized, labeled, and used to identify cells carrying a recombinant plasmid that incorporates that particular sequence.

SECTION 7.1 SUMMARY

- Genes are isolated for study by cloning them into vectors that permit their selection and amplification. A gene or genomic segment is cut out of a chromosome with a restriction enzyme and ligated into a vector. The recombinant vector is transferred into a host cell and is amplified in this transformed cell.
- Gene cloning relies on an arsenal of enzymes made available by advances in molecular biology, including restriction endonucleases, DNA ligase, DNA polymerase, and reverse transcriptase.

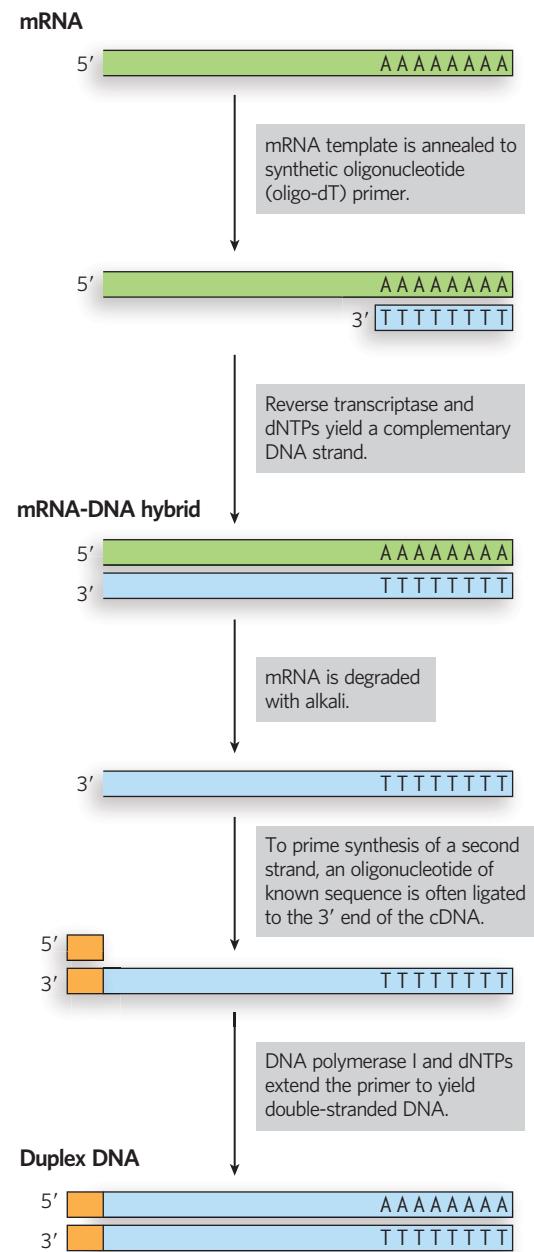


FIGURE 7-8 Building a cDNA library from mRNA. A cell's total mRNA content includes transcripts from thousands of genes, and the cDNAs generated from this mRNA are correspondingly heterogeneous. Reverse transcriptase can synthesize DNA on an RNA or DNA template. To prime the synthesis of a second DNA strand, oligonucleotides of known sequence are ligated to the 3' end of the first strand, and the double-stranded cDNA so produced is cloned into a plasmid.

- Important cloning vectors include plasmids, bacterial artificial chromosomes, and yeast artificial chromosomes. BACs and YACs allow the cloning of very long DNA segments.

- DNA libraries are specialized archives available for use in gene sequencing, gene discovery, or the functional characterization of proteins.

7.2 Working with Genes and Their Products

The isolation of a gene, or any segment of genomic DNA, generally has one of two purposes. One is to examine the DNA itself, determine its sequence, study its structure and/or function, and compare it with other DNA segments. Researchers in physical biochemistry, for example, might be interested in the structure of an unusual repeated sequence. Evolutionary biologists and forensic scientists might be interested in comparing the sequence of the DNA segment with the same segment taken from other individuals in a population, or with related DNA segments from other species. The other possible purpose is to work with the protein or RNA product of the isolated gene. These gene products are at the heart of every biological process. Genetic engineering provides tools not only for the isolation and study of proteins and RNA, but also for their alteration for myriad purposes.

The isolation and examination of DNA segments has been greatly facilitated by PCR technology, and we discuss this first. We then explore modern DNA-sequencing methods and a variety of techniques for expressing and altering gene products—primarily proteins—to understand their function and harness them for new purposes.

Gene Sequences Can Be Amplified with the Polymerase Chain Reaction

Genome projects are now proceeding worldwide, creating rapidly growing international databases containing the complete genome sequences of hundreds of organisms. Such programs are providing unprecedented access to gene sequence information. This effort, in turn, is simplifying the process of cloning individual genes for more detailed analysis. If we know the sequence of at least the end portions of a DNA segment we are interested in, we can hugely amplify the number of copies of that DNA segment with the **polymerase chain reaction (PCR)**, a process conceived by Kary Mullis in 1983 (see How We Know). The amplified DNA can then be cloned by the methods described earlier or used in a variety of analytical procedures.

The PCR procedure has an elegant simplicity, and relies on enzymes called DNA polymerases. DNA polymerases synthesize DNA strands from deoxyribonucleotides, using a DNA template. Further, DNA polymerases

do not synthesize DNA de novo, but instead must add nucleotides to preexisting strands, referred to as primers, as described in Chapter 11. Two synthetic oligonucleotides are prepared, complementary to sequences on opposite strands of the target DNA at positions defining the ends of the segment to be amplified. The oligonucleotides serve as replication primers that can be extended by a DNA polymerase. The 3' ends of the hybridized primers are oriented toward each other and positioned to prime DNA synthesis across the DNA segment (Figure 7-9a). Basic PCR requires four components: a DNA sample containing the segment to be amplified, the pair of synthetic oligonucleotide primers, deoxynucleoside triphosphates (dNTPs), and DNA polymerase. The reaction mixture is heated briefly to denature the DNA, separating the two strands. The mixture is cooled so that the primers can anneal to the DNA. The high concentration of primers increases the likelihood that they will anneal to each strand of the denatured DNA before the two DNA strands (present at a much lower concentration) can reanneal to each other. The primed segment is then replicated selectively by the DNA polymerase, using the pool of dNTPs. The cycle of heating, cooling, and replication is repeated 25 to 30 times over a few hours in an automated process, amplifying the DNA segment between the primers until it can be readily analyzed or cloned. Each cycle increases the amount of the DNA segment by a factor of 2, so the concentration of this DNA grows exponentially. After 20 cycles, the DNA segment has been amplified more than a millionfold (2^{20}); after 30 cycles, more than a billionfold. All other DNA in the sample remains unamplified. PCR uses a heat-stable DNA polymerase, such as the *Taq* polymerase, which remains active after every heating step and does not have to be replenished.

By careful design of the primers used for PCR, the amplified segment can be altered by the inclusion, at each end, of additional DNA not present in the chromosome that is being targeted. For example, restriction endonuclease cleavage sites can be included to facilitate the subsequent cloning of the amplified DNA (Figure 7-9b).

This technology is highly sensitive: PCR can detect and amplify as little as one DNA molecule in almost any type of sample—including some quite ancient ones. The double-helical structure of DNA makes it a highly stable molecule (see Chapter 6), but DNA does degrade slowly over time (through reactions described in Chapter 12). However, PCR has allowed the successful cloning of rare, undegraded DNA segments from samples more than 40,000 years old. Investigators have used the technique to clone DNA fragments from the mummified remains of humans and extinct animals, such as the woolly mammoth, creating the new fields of molecular

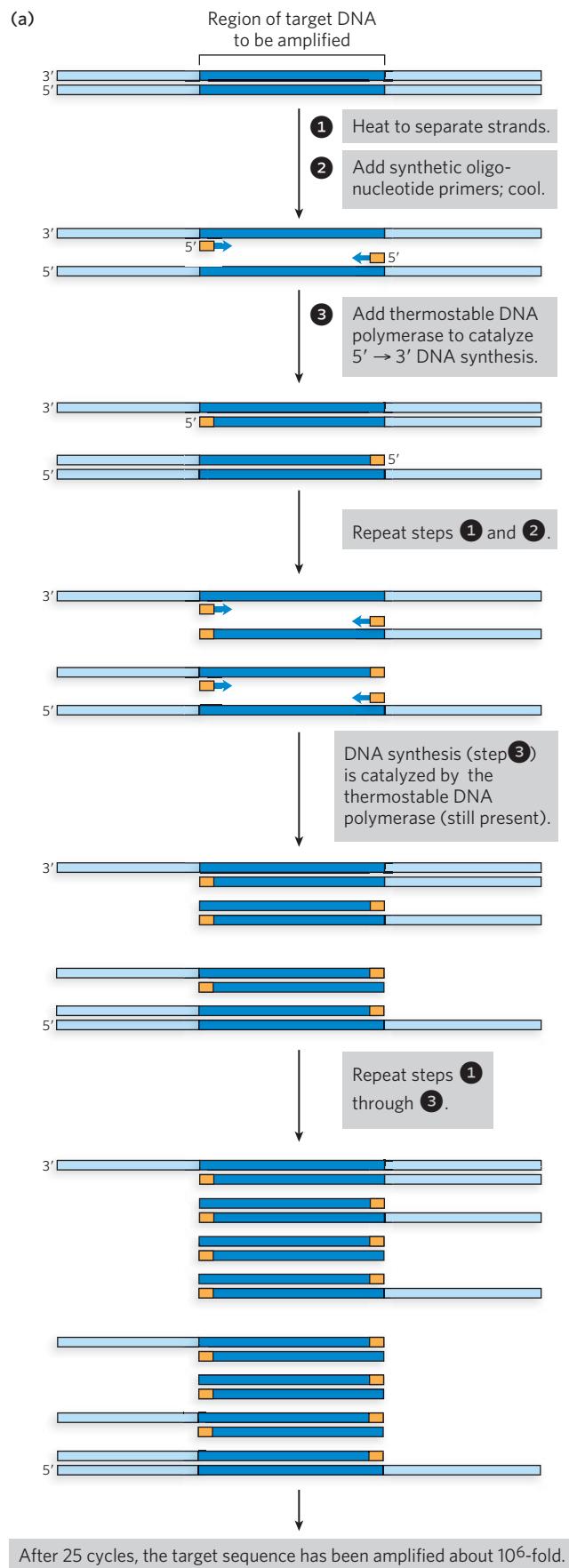
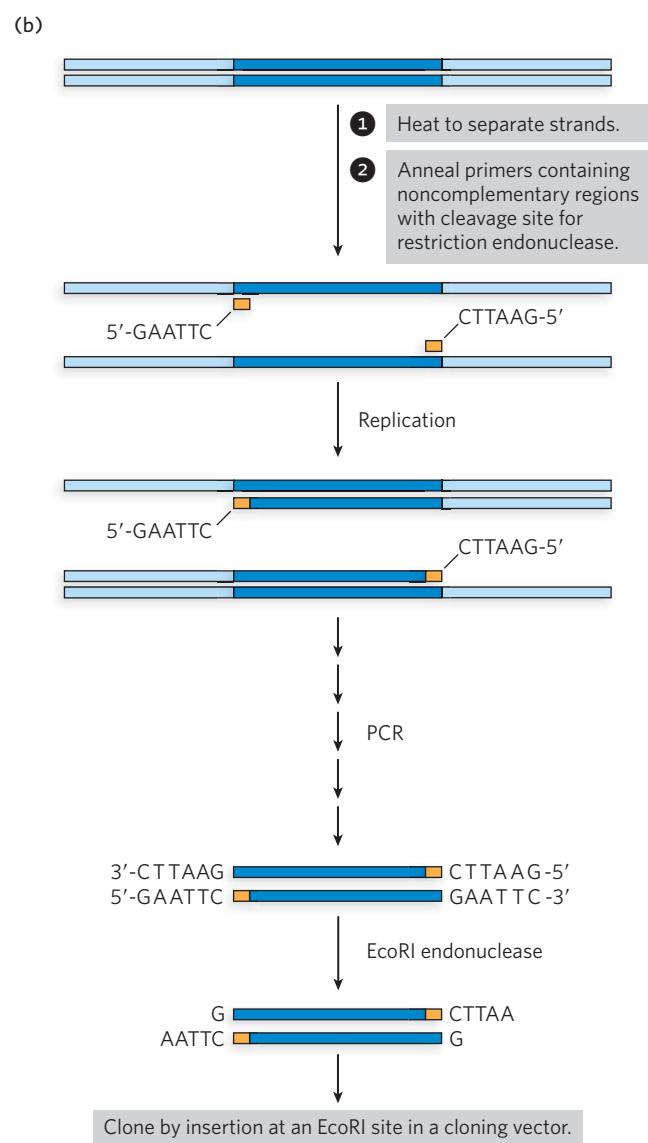


FIGURE 7-9 Amplification of a DNA segment by the polymerase chain reaction (PCR). PCR leads to specific amplification of DNA in a segment defined by the two designed DNA primers.



archaeology and molecular paleontology. DNA from burial sites has been amplified by PCR and used to trace ancient human migrations. Epidemiologists can use PCR-enhanced DNA samples from human remains to trace the evolution of human pathogenic viruses. Thus, in addition to its usefulness for cloning DNA, PCR is a potent tool in forensic medicine (Highlight 7-1). It is also being used for detecting viral infections before they cause symptoms and for the prenatal diagnosis of a wide array of genetic diseases.

Given the extreme sensitivity of PCR methods, contamination of samples is a serious issue. In many applications, including forensic and ancient DNA tests, controls must be run to make sure the amplified DNA is not derived from the researcher or from contaminating bacteria.

Many specialized adaptations of PCR have increased the utility of the method. For example, sequences in RNA can be amplified if reverse transcriptase is used in the first PCR cycle (see Figure 7-8). After the DNA strand is made, using the RNA as a template, the remaining cycles can be carried out with DNA polymerases by normal PCR protocols. This **reverse transcriptase PCR (RT-PCR)** can be used, for example, to detect sequences derived from living cells (which are transcribing their DNA into RNA) as opposed to dead tissues.

PCR protocols can also be made quantitative for estimating the relative copy numbers of particular sequences in a sample. The approach is called **real-time PCR** (or quantitative PCR, or qPCR). If a DNA sequence is present in higher than usual amounts in a sample—for example, certain genes may be amplified so that they are present in many copies in the cells that make up a cancerous tumor—real-time PCR can reveal the increased representation of that sequence. In brief, the PCR is carried out in the presence of a probe that emits a fluorescent signal when the PCR product is present (Figure 7-10). If the sequence of interest is present at higher levels than other sequences in the sample, the PCR signal will reach a predetermined threshold faster. Reverse transcriptase PCR and real-time PCR can be combined to determine the relative transcription levels of genes in a cell under different environmental conditions, or to study the regulation of transcription of one or more genes.

The Sanger Method Identifies Nucleotide Sequences in Cloned Genes

In its capacity as a repository of information, a DNA molecule's most important property is its nucleotide sequence. Until the late 1970s, determining the sequence of a nucleic acid containing even 5 or 10 nucle-

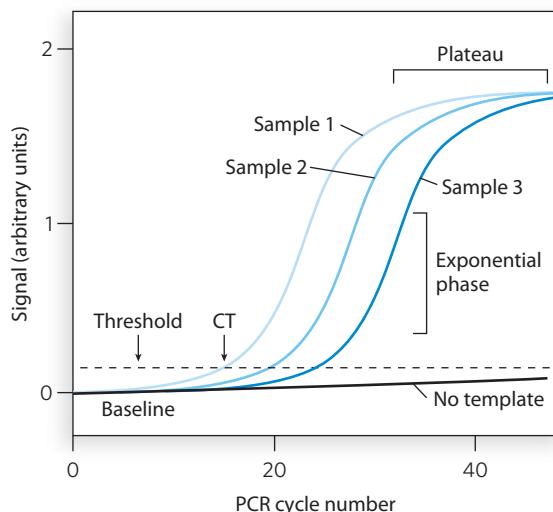
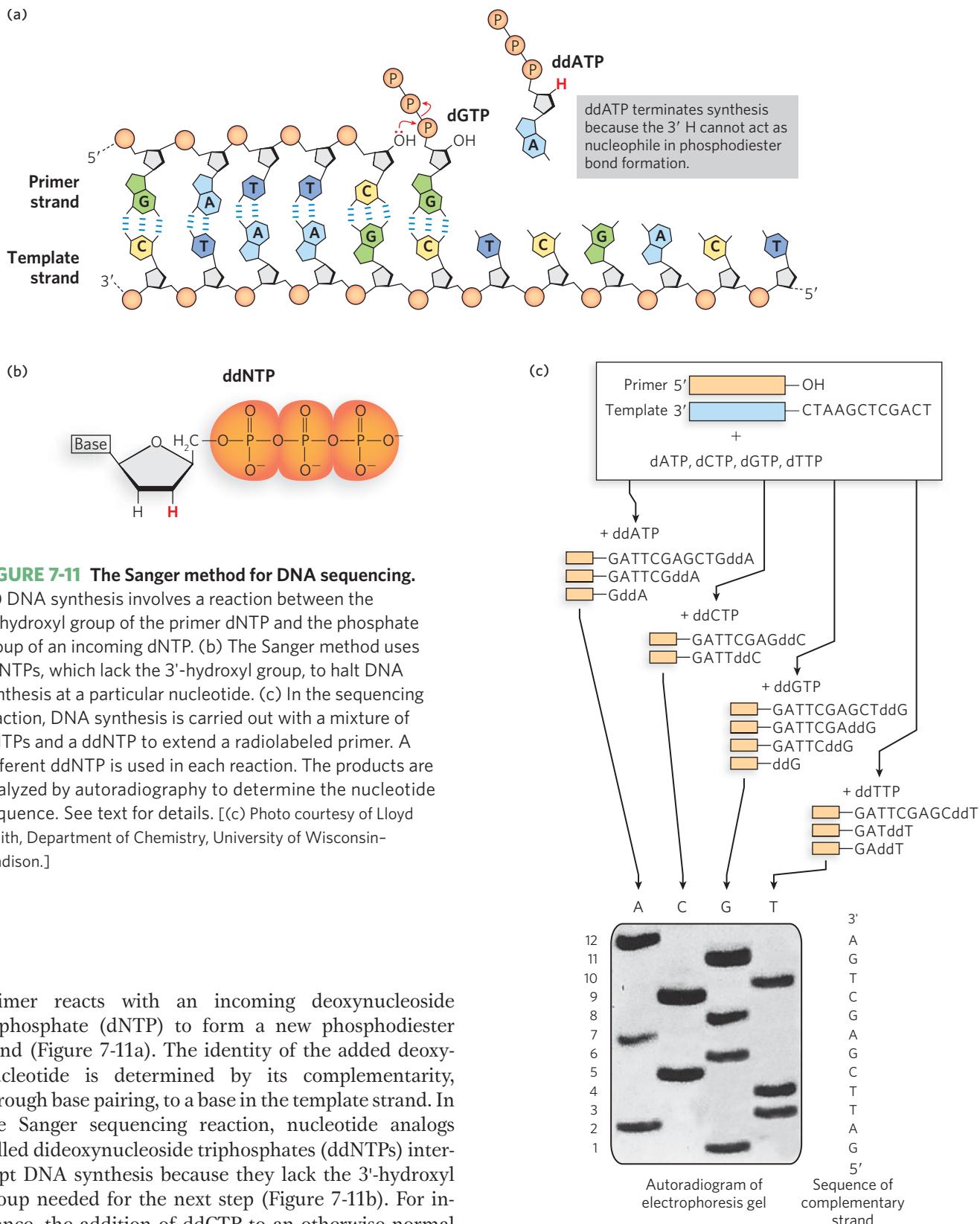


FIGURE 7-10 Real-time PCR. PCR can be used quantitatively, by carefully monitoring the progress of a PCR amplification and determining when a DNA segment has been amplified to a specific threshold level. The amount is determined by measuring the level of a fluorescent probe attached to the DNA. When a segment is present in greater amounts in one sample than another, its amplification reaches the threshold level earlier. CT is the cycle number at which the threshold is first surpassed.

otides was very laborious. The development of two techniques in 1977, one by Allan Maxam and Walter Gilbert and the other by Frederick Sanger, made possible the sequencing of larger DNA molecules with an ease unimagined just a few years before. The techniques depend on an improved understanding of nucleotide chemistry and DNA metabolism, and on improved electrophoretic methods for separating DNA strands that differ in size by only one nucleotide (see Figure 6-32 for a description of gel electrophoresis). In work with short DNA oligonucleotides (up to a few hundred nucleotides), polyacrylamide is often used instead of agarose as the gel matrix, because it enables researchers to detect small size differences between DNA fragments.

Although the two methods are similar in approach, the **Sanger method**, also known as the dideoxy chain-termination method, has proved to be technically easier and is in more widespread use (Figure 7-11). This method makes use of the mechanism of DNA synthesis by DNA polymerases (see Chapter 11). It requires the enzymatic synthesis of a DNA strand complementary to the strand under analysis, using a radioactively labeled primer and dideoxynucleotides. In the reaction catalyzed by DNA polymerase, the 3'-hydroxyl group of the



HIGHLIGHT 7-1 TECHNOLOGY

A Potent Weapon in Forensic Medicine

One of the most accurate methods for placing an individual at the scene of a crime is a fingerprint. The advent of recombinant DNA technology has made a much more powerful tool available: **DNA genotyping** (also called DNA fingerprinting or DNA profiling). As first described by English geneticist Alec Jeffreys in 1985, the method is based on **sequence polymorphisms**, slight sequence differences among individuals—1 in every 1,000 bp, on average. Each difference from the prototype human genome sequence (the first one obtained) occurs in some fraction of the human population; every person has some differences from this prototype.

Forensic work focuses on differences in the lengths of **short tandem repeat (STR)** sequences. An STR locus is a short DNA sequence, repeated many times in tandem at a specific location in a chromosome; usually, the repeated sequence is 4 bp long. The loci most often used in STR genotyping are short—4 to 50 repeats (16 to 200 bp for tetranucleotide repeats) and have multiple length variants in the human population. More than 20,000 tetranucleotide STR loci have been characterized in the human genome. And more than a million STRs of all types may be present in the human genome, accounting for about 3% of all human DNA.

The length of a particular STR in a given individual can be determined with the aid of the polymerase chain reaction (see Figure 7-9). The use of PCR also makes the procedure sensitive enough to be applied to the very small samples often collected at crime scenes. The DNA sequences flanking STRs are unique to each type of STR and are identical (except for very rare mutations) in all humans. PCR primers are targeted to this flanking DNA and are designed to amplify the DNA across the STR (Figure 1a). The length of the PCR product then reflects the length of the STR in that sample. Because each human inherits one chromosome of each chromosome pair from each parent, the STR lengths on the two chromosomes are often different, generating two different STR lengths from one individual. The PCR products are subjected to electrophoresis on a very thin polyacrylamide gel in a capillary tube. The resulting bands are converted into a set of peaks that accurately reveal the size of each PCR fragment and thus the length of the STR in the corresponding allele. Analysis of multiple STR loci can yield a profile that is unique to an individual (Figure 1b). This is typically done with a commercially available kit that includes PCR primers unique to each locus, linked to colored dyes to help distinguish the different PCR products. PCR amplification

Table 1 Properties of the Loci Used for the CODIS Database

Locus Name	Chromosome	Repeat Motif*	Repeat Length (range) [†]	Number of Alleles Seen [‡]
CSF1PO	5	TAGA	5–16	20
FGA	4	CTTT	12.2–51.2	80
TH01	11	TCAT	3–14	20
TPOX	2	GAAT	4–16	15
VWA	12	[TCTG][TCTA]	10–25	28
D3S1358	3	[TCTG][TCTA]	8–21	24
D5S818	5	AGAT	7–18	15
D7S820	7	GATA	5–16	30
D8S1179	8	[TCTA][TCTG]	7–20	17
D13S317	13	TATC	5–16	17
D16S539	16	GATA	5–16	19
D18S51	18	AGAA	7–39.2	51
D21S11	21	[TCTA][TCTG]	12–41.2	82
Amelogenin [§]	X, Y	Not applicable		

Source: Adapted from J. M. Butler, *Forensic DNA Typing*, 2nd ed., Academic Press, 2006, p. 96.

*Brackets indicate alternating repeats.

[†]Repeat lengths observed in the human population. Partial or imperfect repeats are seen in some alleles.

[‡]Number of different alleles observed in the human population. Careful analysis of the same locus in many individuals is a prerequisite to its use in forensic DNA typing.

[§]Amelogenin is a gene, of slightly different size on the X and Y chromosomes, that is used to establish gender.

enables investigators to obtain STR genotypes from less than 1 ng of partially degraded DNA, an amount that can be obtained from a single hair follicle, a drop of blood, a small semen sample, or samples that might be months or even many years old. When good STR genotypes are obtained,¹⁸ the chance of misidentification is less than 1 in 10^{18} (a quintillion).

The successful forensic use of STR analysis required standardization, first attempted in the United Kingdom in 1995. The U.S. standard, called the Combined DNA Index System (CODIS), established in 1998, is based on 13 well-studied STR loci, which must be present in any DNA-typing experiment carried out in the United States (Table 1). The amelogenin gene is also used as a marker in the analyses. Present on the human sex chromosomes, this gene

has a slightly different length on the X and Y chromosomes. PCR amplification across this gene thus generates different-sized products that can reveal the sex of the DNA donor. By the beginning of 2010, the CODIS database contained more than 7 million STR genotypes and had assisted more than 100,000 forensic investigations.

DNA genotyping has been used to both convict and acquit suspects, and to establish paternity with an extraordinary degree of certainty. The impact of these procedures on court cases will continue to grow as standards are refined and as international STR genotyping databases grow. Even very old mysteries can be solved. In 1996, STR genotyping helped confirm the identification of the bones of the last Russian czar and his family, who were assassinated in 1918.

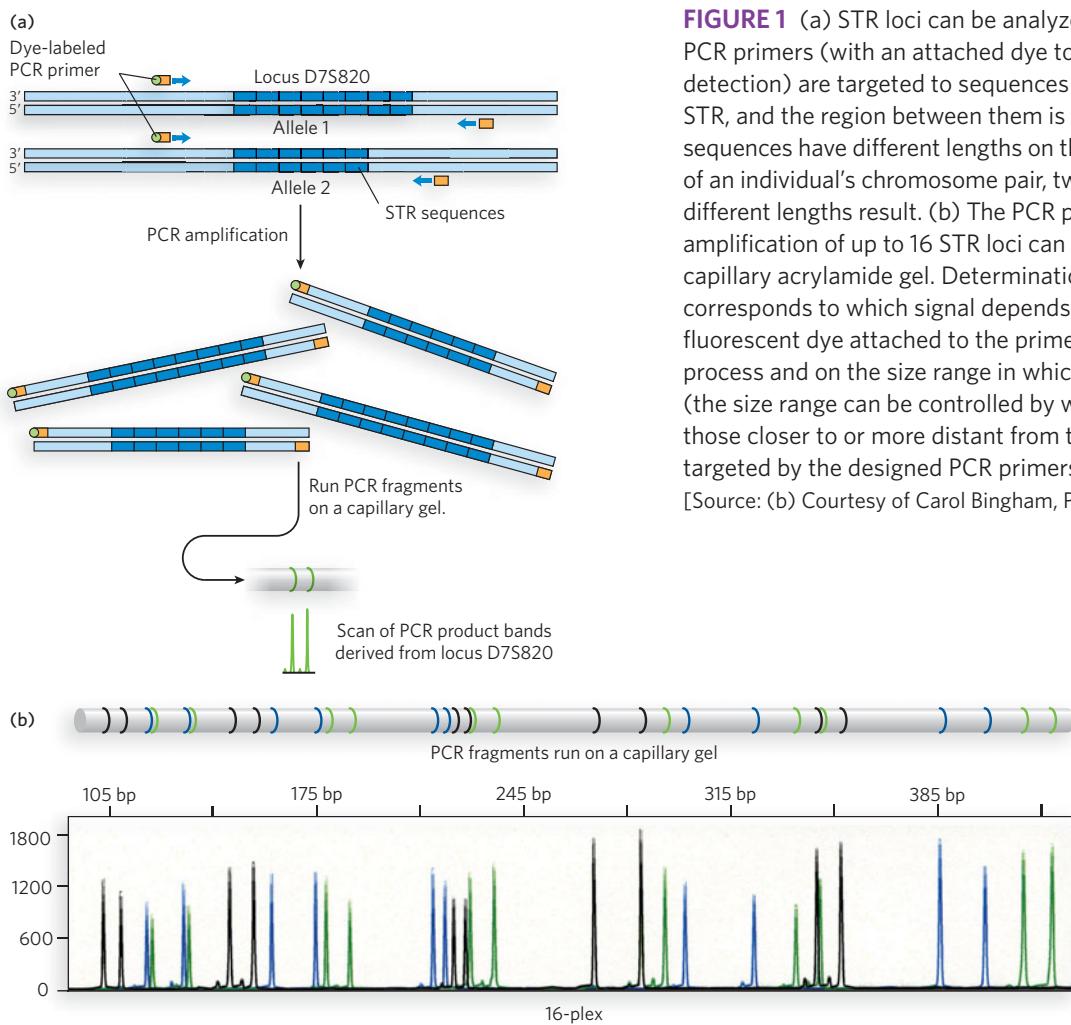


FIGURE 1 (a) STR loci can be analyzed by PCR. Suitable PCR primers (with an attached dye to aid in subsequent detection) are targeted to sequences on each side of the STR, and the region between them is amplified. If the STR sequences have different lengths on the two chromosomes of an individual's chromosome pair, two PCR products of different lengths result. (b) The PCR products from amplification of up to 16 STR loci can be run on a single capillary acrylamide gel. Determination of which locus corresponds to which signal depends on the color of the fluorescent dye attached to the primers used in the process and on the size range in which the signal appears (the size range can be controlled by which sequences—those closer to or more distant from the STR—are targeted by the designed PCR primers).

[Source: (b) Courtesy of Carol Bingham, Promega Corporation.]

Given the excess of dCTP over ddCTP, the chance that the analog will be incorporated instead of dC is small. But ddCTP is present in sufficient amounts to ensure that each new strand has a high probability of acquiring at least one ddC at some point during synthesis. The result is a solution containing a mixture of labeled fragments, each ending with a C residue. Each G residue in the template generates a set of C-terminated fragments of a particular length, such that the different-sized fragments, separated by electrophoresis, reveal the location of C residues in the synthesized DNA strand. This procedure is repeated separately for each of the four ddNTPs, and the sequence of the DNA strand can be read directly from an autoradiogram of the gel (Figure 7-11c). Because shorter DNA fragments migrate faster, the fragments near the bottom of the gel represent the nucleotide positions closest to the primer (the 5' end), and the sequence is read (in the 5' → 3' direction) from bottom to top. Note that the sequence obtained is that of the strand *complementary* to the template strand being analyzed.

DNA sequencing was first automated by a variation of the Sanger method, in which each of the four dideoxynucleotides used for a reaction was labeled with a differently colored fluorescent tag (Figure 7-12). With this technology, researchers could sequence DNA molecules containing thousands of nucleotides in a few hours, and the entire genomes of hundreds of organisms were sequenced in this way. For example, in the Human Genome Project, researchers sequenced all 3.2×10^9 bp of the DNA in a human cell (see Chapter 8).

In this era of genome sequencing, newer methods are evolving rapidly, dramatically speeding the process and lowering the cost of large sequencing projects. Some of the methods make use of chemistry that subtly alters the Sanger method (Highlight 7-2). These technologies have

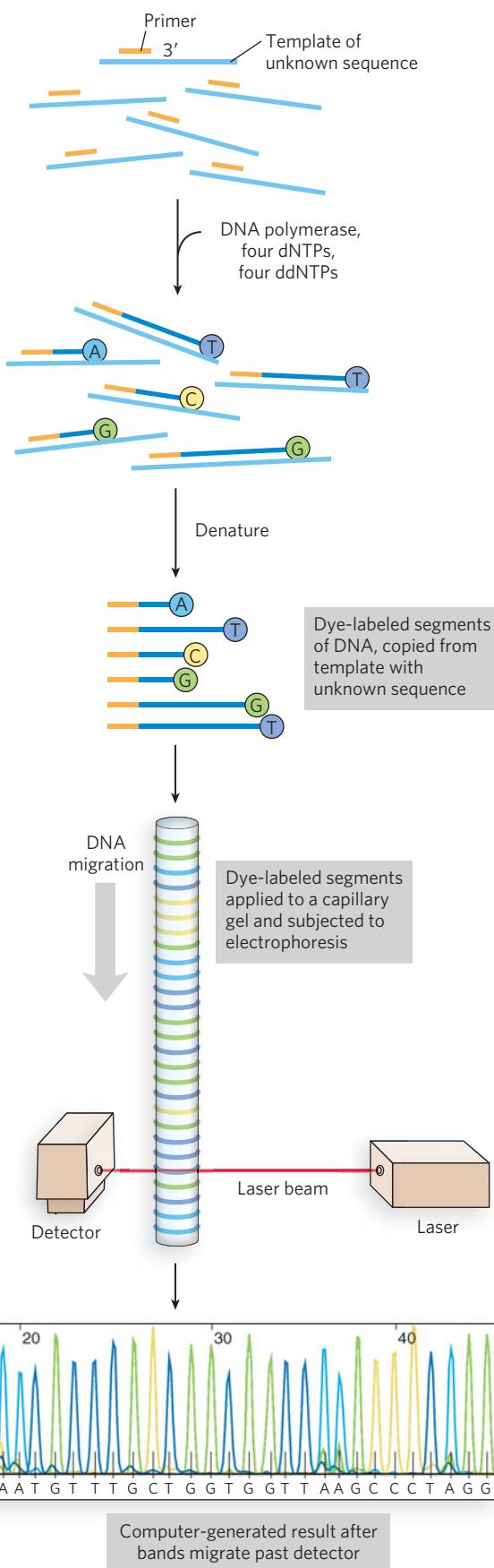


FIGURE 7-12 Automation of DNA-sequencing reactions.

In the Sanger method, each ddNTP can be linked to a fluorescent (dye) molecule that gives the same color to all the fragments terminating in that nucleotide, a different color for each nucleotide. All four labeled ddNTPs are added together. The resulting colored DNA fragments are separated by size in an electrophoretic gel in a capillary tube (a refinement of gel electrophoresis that allows for faster separations). All fragments of a given length migrate through the capillary gel together in a single band, and the color associated with each band is detected with a laser beam. The DNA sequence is read by identifying the color sequences in the bands as they pass the detector. This information is fed directly to a computer, and the sequence is calculated. The amount of fluorescence in each band is represented as a peak in the computer output. [Source: Data provided by Lloyd Smith, Department of Chemistry, University of Wisconsin-Madison.]

been automated and can generate millions of base pairs of DNA sequence information on one machine in minutes. Computers store the information and splice it together by finding sequence overlaps among the clusters. The major limitation of these methods is the length of the individual DNA sequences generated, often only 20 to 100 nucleotides. These lengths are insufficient to accurately map DNA segments containing repeated DNA sequences. However, nonrepetitive parts of the genome can be faithfully reconstructed. In addition, there are many instances where it is useful to resequence an already sequenced genome. For example, the DNA sequencing of several or many individuals of a single species may detect sequence diversity among individuals of a population or sequence changes that result from mutagenesis or evolution. The new generation of sequencing techniques is effective for these genome resequencing experiments. They also provide just a taste of the revolution to come in the burgeoning field of genomics (see Chapter 8). The eventual goal is to reduce the cost so that every human can have his or her genome sequenced—resulting, ideally, in the delivery of highly personalized medicine. The obstacles to realizing this vision are not just technical. Ethical issues, such as the potential for restricted access to insurance or job discrimination based on genetic conditions revealed in the genomic sequence, must also be addressed.

Cloned Genes Can Be Expressed to Amplify Protein Production

Frequently, it is the product of a cloned gene, rather than the gene itself, that is of primary interest—particularly when the protein has commercial, therapeutic, or research value. Molecular biologists use purified proteins to elucidate protein function, study reaction mechanisms, generate antibodies, reconstitute complex cellular activities in the test tube with purified components, examine protein binding partners, and many other purposes. With an increased understanding of the fundamentals of DNA, RNA, and protein metabolism and their regulation in *E. coli*, investigators can now manipulate cells to express cloned genes in order to study their protein products. The general goal is to alter the sequences around a cloned gene so as to trick the host organism into producing the protein product of the gene, often at very high levels. This overexpression of a protein can make its subsequent purification a lot easier.

We'll use the expression of a eukaryotic protein in a bacterium as an example. Most cloned eukaryotic genes lack the DNA sequence elements required for their controlled expression in bacterial cells—promoters (sequences that instruct RNA polymerase where to bind to initiate mRNA synthesis), ribosome-binding sites

(sequences that allow translation of the mRNA to protein), and additional regulatory sequences (see Chapter 15). Therefore, appropriate bacterial regulatory sequences for transcription and translation must be inserted at the correct positions, relative to the eukaryotic gene, in the vector DNA. In some cases, cloned genes are so efficiently expressed that their protein product represents 10% or more of the cellular protein. At these concentrations, some foreign proteins can kill the host cell (usually *E. coli*), so the cloned gene expression must be limited to the few hours before the planned harvesting of the cells.

Cloning vectors with the transcription and translation signals needed for the regulated expression of a cloned gene are called **expression vectors**. The rate of expression of the cloned gene is controlled by replacing the gene's own promoter and regulatory sequences with more efficient and convenient versions supplied by the vector. Generally, a well-characterized promoter and its regulatory elements are positioned near several unique restriction sites for cloning, so that genes inserted at the restriction sites will be expressed from the regulated promoter elements (Figure 7-13). Some of these vectors incorporate other features, such as a

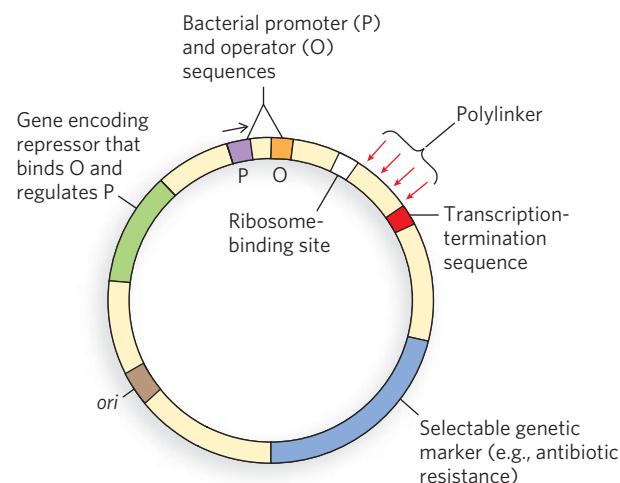


FIGURE 7-13 DNA sequences in a typical *E. coli* expression vector. The gene to be expressed is inserted into one of the restriction sites in the polylinker, near the promoter (P), with the end of the gene encoding the N-terminus of the protein positioned closest to the promoter. The promoter allows efficient transcription of the inserted gene, and the transcription-termination sequence sometimes improves the amount and stability of the mRNA produced. The operator (O) permits regulation by a repressor that binds to it. The ribosome-binding site provides sequence signals for the efficient translation of the mRNA derived from the gene. The selectable marker allows the selection of cells containing the recombinant DNA.

HIGHLIGHT 7-2 TECHNOLOGY

DNA Sequencing: Ever Faster and Cheaper

As modern sequencing techniques have pushed the limits of the Sanger method, alternatives have begun to appear. In these new approaches, thousands or even millions of DNA sequences are generated in parallel. The sequencing strategy is similar to that used in the Sanger method, but multiple innovations have allowed a miniaturization of the procedure.

First, individual genomic DNA segments are replicated so that all the resulting copies stay together as a dense cluster of identical DNA molecules. The clusters are distributed on a solid surface, separated spatially from one another. Millions of clusters may be present on a silica surface just centimeters wide, and each cluster represents a different segment of DNA from the genome. Second, as the sequencing reactions proceed, an image of the surface is captured and the identity of the nucleotide added to each cluster is recorded as either a color or a flash of light. The clusters amplify this sequencing signal sufficiently to allow detection. The image for each sequence cycle records each added nucleotide for all the clusters in the image, and all clusters are thus sequenced in parallel, greatly increasing the speed of sequencing.

Two widely utilized next-generation sequencers use different strategies to accomplish the sequencing reactions. One of these, 454 Sequencing, uses a strategy called pyrosequencing in which the added nucleotides are detected with flashes of light (Figure 1). The four trinucleoside triphosphates (unaltered) are added one at a time, in a repeating sequence. The pulse of nucleotide added is retained on the surface just long enough for DNA polymerase to add that nucleotide to any cluster where it is complementary to the next template base in the sequence. Excess nucleotide is destroyed quickly with the enzyme apyrase before the next nucleotide pulse. When a specific nucleotide is successfully added to the strands of a cluster, pyrophosphate is released as a by-product. Another enzyme in the

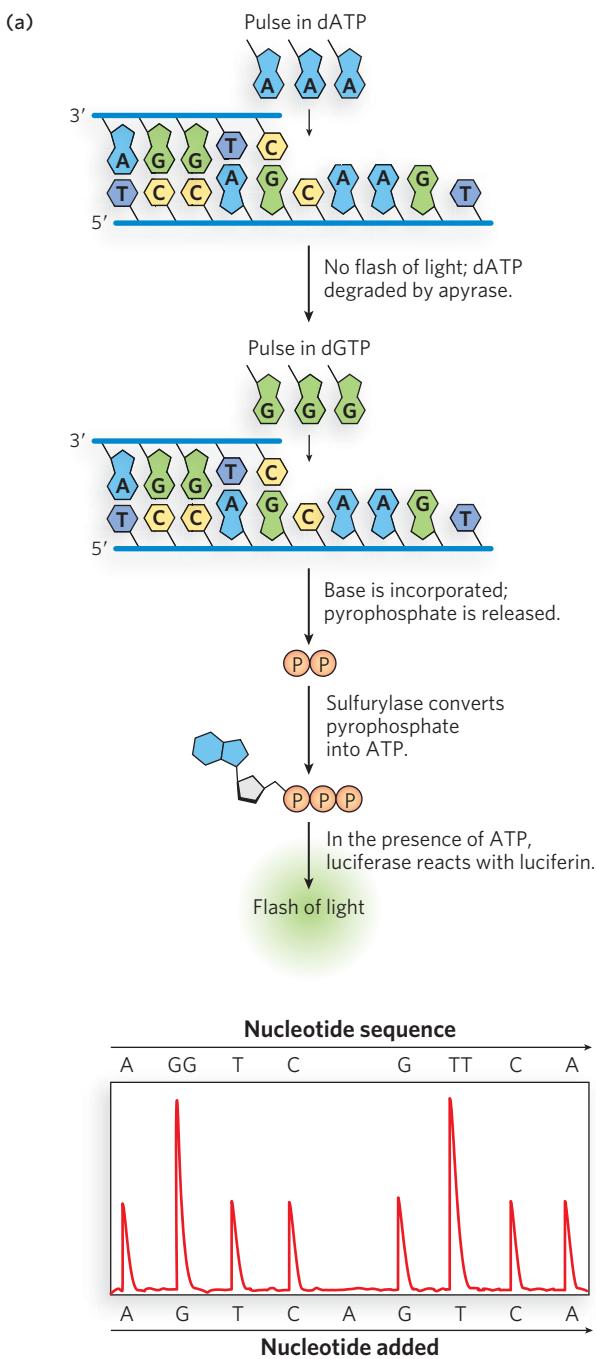
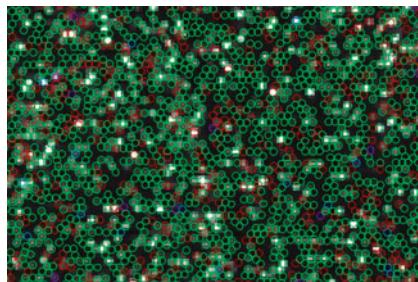


FIGURE 1 (a) Pyrosequencing detects nucleotides by flashes of light. (b) An image of a small part of one cycle of a 454 Sequencing run. Light flashes within circles denote the addition of a nucleotide to the oligonucleotides in the corresponding clusters. [Source: (b) 454 Sequencing™, © 2010 Roche Diagnostics.]



solution bathing the surface is sulfurylase, which converts the pyrophosphate to ATP.

Also present in the medium is the enzyme luciferase and a substrate molecule, luciferin. When ATP is generated, luciferase catalyzes a reaction with luciferin, causing a tiny flash of light. When many of these occur in a cluster, the cumulative flash can be recorded in the captured image. For example, when dCTP is added to the solution, flashes will occur only at clusters where G is present in the template and C is the next nucleotide to be added to the growing DNA chain. If there is a string of 2, 3, or 4 G residues in the template, a similar number of C residues will be added to the growing strand in one cycle. This is recorded as a “flash” amplitude at that cluster that is 2, 3, or 4 times greater than when only one C residue is added. Similarly, when dGTP is added, flashes occur at a different set of clusters, marking those as

clusters where G is the next nucleotide added to the sequence.

The alternative method is the Illumina sequencer, which uses reversible terminator sequencing (Figure 2). A special sequencing primer is added to the clusters, along with fluorescently labeled terminator nucleotides and DNA polymerase. The polymerase adds the appropriate nucleotide to the strands in each cluster, each nucleotide carrying a different fluorescent label. The terminator-blocking group permits only one nucleotide addition to each strand. Next, lasers excite all the fluorescent labels, and an image of the entire surface reveals the color (and thus the identity of the base) added to each cluster. The fluorescent label and the terminator are then chemically or photolytically removed, in preparation for adding a new nucleotide to each cluster. The sequencing proceeds stepwise.

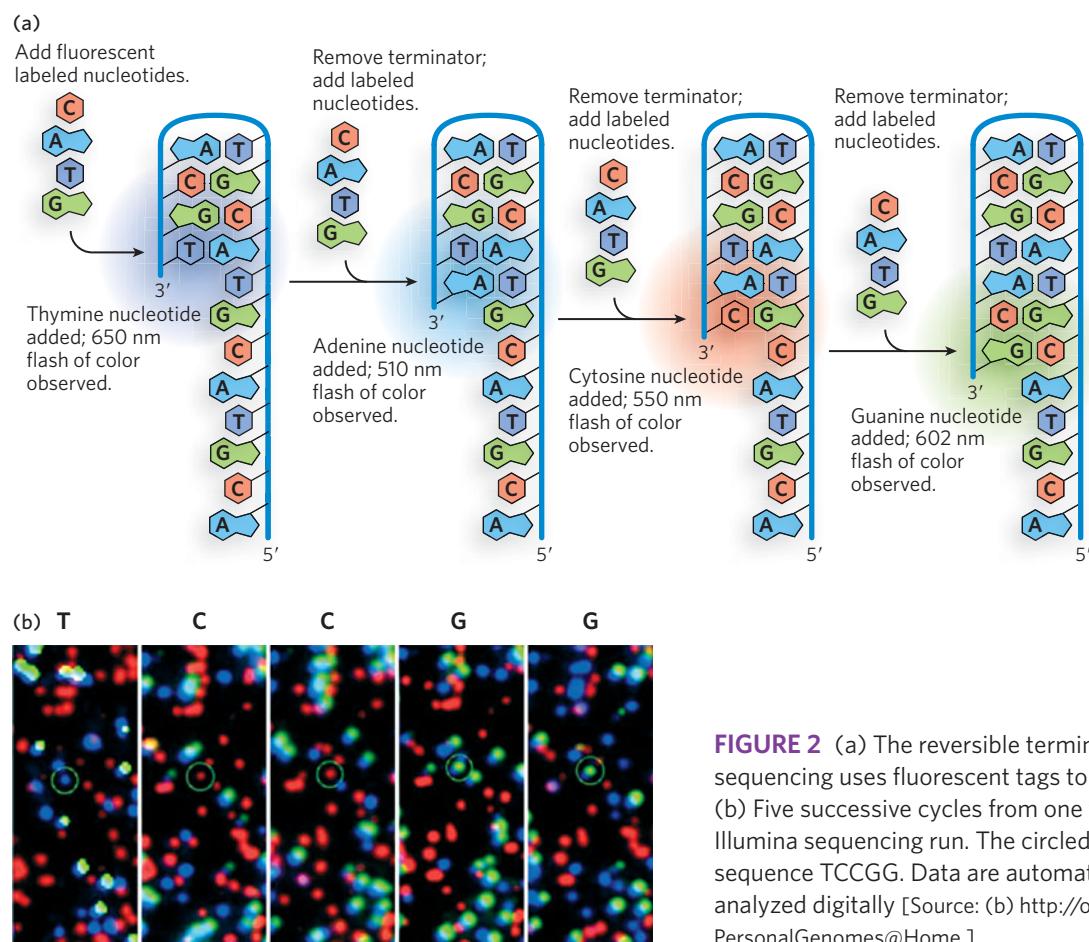


FIGURE 2 (a) The reversible terminator method of sequencing uses fluorescent tags to identify nucleotides. (b) Five successive cycles from one very small part of an Illumina sequencing run. The circled cluster produced the sequence TCCGG. Data are automatically recorded and analyzed digitally [Source: (b) <http://openwetware.org/wiki/PersonalGenomes@Home>.]

bacterial ribosome-binding site to enhance translation of the mRNA derived from the gene (see Chapter 18) or a transcription-termination sequence (Chapter 15).

Many Different Systems Are Used to Express Recombinant Proteins

Bacteria Bacteria, especially *E. coli*, remain the most common hosts for protein expression. The regulatory sequences that govern gene expression in *E. coli* and many other bacteria are well understood and can be harnessed to express cloned proteins at high levels. Bacteria are easy to store and grow in the laboratory, on inexpensive growth media. Efficient methods also exist to get DNA into bacteria and extract DNA from them. Bacteria can be grown in huge amounts in commercial fermentors, providing a rich source of the cloned protein. Problems do exist, however. When expressed in bacteria, some heterologous proteins do not fold correctly, and many do not undergo the postsynthetic modifications (covalent modification, proteolytic cleavage, etc.; see Chapter 18) necessary for their activity. A variety of gene sequence features also can make a particular gene difficult to express in bacteria. For these and many other reasons, some eukaryotic proteins are inactive when purified from bacteria, or cannot be expressed at all.

There are many specialized systems for expressing proteins in bacteria. The promoter and regulatory sequences associated with the lactose operon (see Chapters 5 and 20) are often fused to the gene of interest to direct transcription. The cloned gene will be transcribed when lactose is added to the growth medium. However, regulation in the lactose system is “leaky”: it is not turned off completely when lactose is absent—a potential problem if the product of the cloned gene is toxic to the host cells. Transcription from the Lac promoter is also not efficient enough for some applications.

An alternative system uses a promoter and RNA polymerase found in a bacterial virus called bacteriophage T7. If the cloned gene is fused to a T7 promoter, it is not transcribed by the *E. coli* RNA polymerase, but instead by the T7 RNA polymerase. The gene encoding this polymerase is separately cloned into the same cell in a construct that affords tight regulation (allowing controlled production of the T7 RNA polymerase). The

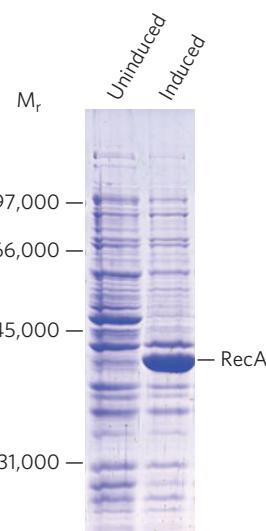


FIGURE 7-14 Regulated expression of RecA protein in a bacterial cell. The gene encoding the RecA protein, fused to a bacteriophage T7 promoter, is cloned into an expression vector. Under normal growth conditions (Uninduced), no RecA protein appears. When the T7 RNA polymerase is induced in the cell (Induced), the *recA* gene is expressed, and large amounts of RecA protein are produced. [Source: Courtesy of Rachel Britt, Department of Biochemistry, University of Wisconsin-Madison.]

polymerase is also very efficient and directs high levels of expression of most genes fused to the T7 promoter. This system has been used to express the RecA protein in bacterial cells (Figure 7-14).

Yeast The yeast *Saccharomyces cerevisiae* is probably the best understood eukaryotic organism, and one of the easiest to grow and manipulate in the laboratory. Like bacteria, this yeast can be grown on inexpensive media. Yeast have tough cell walls that are difficult to breach in order to introduce DNA vectors, so bacteria are more convenient for doing much of the genetic engineering and vector maintenance. Several excellent shuttle vectors exist for this purpose.

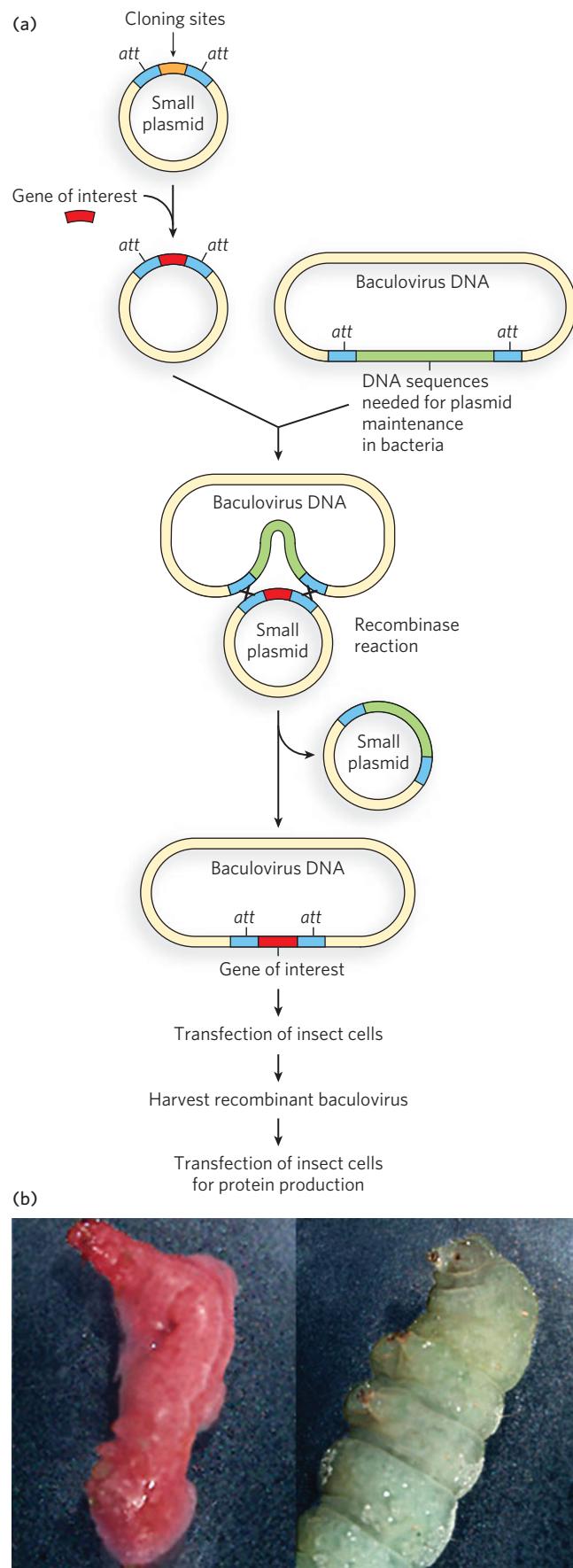
The principles underlying the expression of a protein in yeast are the same as those in bacteria. Cloned genes must be linked to promoters that can direct high-level expression in yeast. For example, the yeast *GAL1* and *GAL10* genes are under cellular regulation such that they are expressed when yeast cells are grown in media with galactose but shut down when the cells are grown in glucose. Thus, if a heterologous gene is expressed using the same regulatory sequences, the expression of that gene can be controlled simply by choosing an appropriate medium for cell growth.

Some of the same problems that accompany protein expression in bacteria also occur with yeast. Heterologous proteins may not fold properly, yeast may lack the enzymes needed to modify the proteins to their active forms, or the expression of proteins may be made difficult by certain features of the gene sequence. However, because *S. cerevisiae* is a eukaryote, the expression of eukaryotic genes (especially yeast genes) is sometimes more efficient in this host than in bacteria. Folding and modification of the products may also be more accurate than for proteins expressed in bacteria.

Insects and Insect Viruses Baculoviruses are insect viruses with double-stranded DNA genomes. When they infect their insect larval hosts, they act as parasites, killing the larvae and turning them into factories for virus production. Late in the infection process, the viruses produce large amounts of two proteins (p10 and polyhedrin)—neither of which is needed for virus production in cultured insect cells, and thus both can be replaced with the gene of a heterologous protein. When the resulting recombinant virus is used to infect insect cells or larvae, the heterologous protein is often produced at very high levels—up to 25% of the total protein present at the end of the infection cycle.

Autographa californica multicapsid nucleopolyhedrovirus (AcMNPV) is the baculovirus most often used for protein expression. It has a large genome (134,000 bp), too large for direct cloning. Virus purification is also cumbersome. These problems have been solved by the creation of **bacmids**, large circular DNAs that include the entire baculovirus genome along with sequences that allow replication of the bacmid in *E. coli* (Figure 7-15). The gene of interest is cloned into a smaller plasmid and combined with the larger plasmid by site-specific recombination in vivo (see Chapter 14). The recombinant bacmid is then isolated and transfected into insect cells (the term **transfection** is

FIGURE 7-15 Cloning with baculoviruses. (a) Shown here is the construction of a typical vector used for protein expression in baculoviruses. The gene of interest is cloned into a small plasmid between two sites (*att*) recognized by a site-specific recombinase, then introduced into the baculovirus vector by site-specific recombination. This generates a circular DNA product that is used to infect the cells of an insect larva. The gene of interest is expressed during the infection cycle, downstream of a promoter that normally expresses a baculovirus coat protein at very high levels. (b) The photographs show (left) an insect larva infected with a recombinant baculovirus vector expressing a protein that produces a red color, and (right) an uninfected larva. [Source: (b) Courtesy of Arthur McIntosh, A. H. McIntosh et al., *J. Insect Sci.* 4:3, 2004.]



used when the DNA used for transformation includes viral sequences and leads to viral replication), followed by recovery of the protein once the infection cycle is finished. A wide range of bacmid systems are available commercially. Baculovirus systems are not successful with all proteins. However, with these systems, insect cells sometimes successfully replicate the protein-modification patterns of higher eukaryotes and produce active, correctly modified eukaryotic proteins.

Mammalian Cells in Culture The most convenient way to introduce cloned genes into a mammalian cell is with viruses. In this way, a molecular biologist can take advantage of the natural capacity of a virus to insert its DNA or RNA into a cell, and sometimes into the cellular chromosome. A variety of engineered mammalian viruses are available as vectors, including human adenoviruses and retroviruses. The gene of interest is cloned so that its expression is controlled by a virus promoter. The virus uses its natural infection mechanisms to introduce the recombinant genome into cells, where the cloned protein is expressed. These systems have the advantage that proteins can be expressed either transiently (if the viral DNA is maintained separately from the host cell genome and eventually degraded) or permanently (if the viral DNA is integrated into the host cell genome). With the correct choice of host cell, the proper posttranslational modification of the protein to its active form can be assured. However,

the growth of mammalian cells in tissue culture is very expensive, and this technology is generally used to test the function of a protein *in vivo* rather than to produce a protein in large amounts.

Transgenic Animals Even large animals can be used for the commercial, large-scale production of recombinant proteins. The strategies are different from those discussed thus far and are designed to generate protein in a low-cost, renewable way, such as purification of a protein from the milk of transgenic dairy cattle (Figure 7-16). The gene of interest is cloned into a special vector, linked to a promoter that directs tissue-specific gene expression. For example, the gene can be placed under the control of regulatory sequences for a mammary gland-specific protein, such as casein lactoglobulin, which is normally secreted in milk in large quantities. The recombinant plasmid is injected into fertilized bovine oocytes, and some of them take up the plasmid and incorporate it into their genome. Genetic analysis or direct demonstration of heterologous protein expression then identifies animals in which the gene transfer has been successful, and these animals are bred. Heterologous proteins expressed in place of casein lactoglobulin can be secreted in the milk at levels above 50% of total milk proteins. Posttranslational protein modifications are not always carried out correctly for proteins expressed in this way, but protein production can be economical once a line of protein-expressing animals is established.



FIGURE 7-16 Cloning in transgenic animals. These cows, grazing in a field in New Zealand, were engineered to produce high levels of a recombinant protein in their milk. [Source: Courtesy of Gotz Laible.]

Alteration of Cloned Genes Produces Altered Proteins

Cloning techniques can be used not only to overproduce proteins but to produce protein products subtly altered from their native forms. Specific amino acids may be replaced individually by **site-directed mutagenesis**. This technique has greatly enhanced research on proteins by allowing investigators to make specific changes in the primary structure and examine the effects of these changes on the protein's folding, three-dimensional structure, and activity. This powerful approach to studying protein structure and function

changes the amino acid sequence by altering the DNA sequence of the cloned gene. If appropriate restriction sites flank the sequence to be altered, researchers can simply remove a DNA segment and replace it with a synthetic one, identical to the original except for the desired change (Figure 7-17a).

When suitably located restriction sites are not present, **oligonucleotide-directed mutagenesis**, coupled to PCR, can create a specific DNA sequence change (Figure 7-17b). Two short, complementary synthetic DNA strands, each with the desired base change, are annealed to opposite strands of the cloned gene within a suitable circular DNA vector. The mismatch of

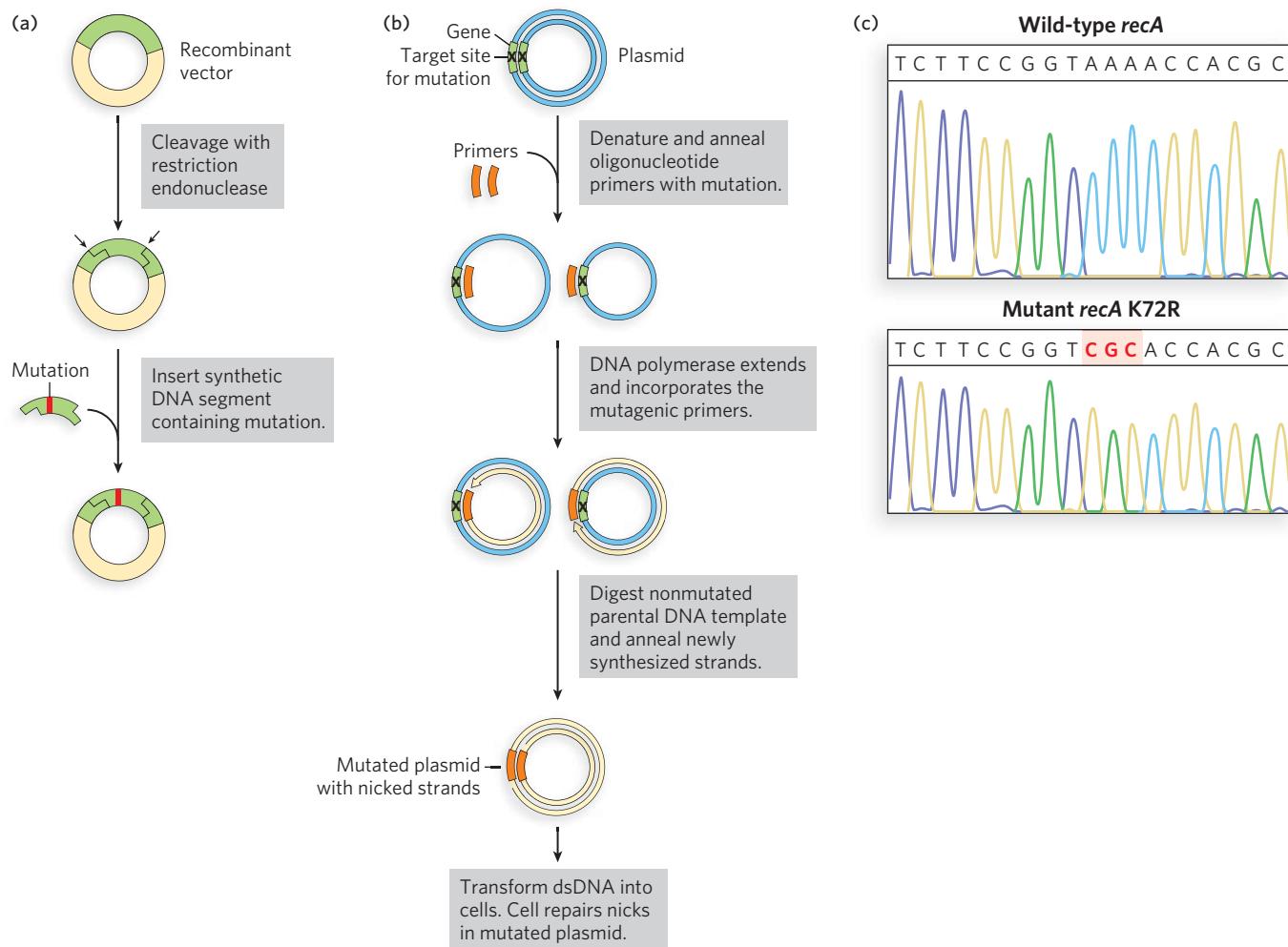


FIGURE 7-17 Two approaches to site-directed mutagenesis.

(a) A synthetic DNA segment replaces a fragment removed by a restriction endonuclease. (b) A pair of synthetic and complementary oligonucleotides with a specific sequence change at one position are hybridized to a circular plasmid with a cloned copy of the gene to be altered. The oligonucleotides act as primers for the synthesis of full-length duplex DNA copies of the plasmid that contain the

specified sequence change. These plasmid copies are then used to transform cells. (c) Results from an automated sequencer (see Figure 7-12), showing sequences from the wild-type *recA* gene (top), and an altered *recA* gene (*recA* K72R, bottom) with the triplet (codon) at position 72 changed from AAA to CGC, specifying an Arg (R) instead of a Lys (K) residue. [Source: (c) Courtesy of Elizabeth Wood, Department of Biochemistry, University of Wisconsin-Madison.]

a single base pair in 30 to 40 bp does not prevent annealing. The two annealed oligonucleotides serve to prime DNA synthesis in both directions around the plasmid vector, creating two complementary strands that contain the mutation. After several cycles of PCR, the mutation-containing DNA predominates in the population and can be used to transform bacteria. Most of the transformed bacteria will have plasmids carrying the mutation. If necessary, the nonmutant template plasmid DNA can be selectively eliminated by cleavage with the restriction enzyme DpnI. The template plasmid, usually isolated from wild-type *E. coli*, has a methylated A residue in every copy of the four-nucleotide palindrome GATC (called a dam site; see Chapter 11). The new DNA containing the mutation does not have methylated A residues, because the replication is done *in vitro*. DpnI selectively cleaves DNA at the sequence GATC only if the A residue in one or both strands is methylated—that is, it breaks down only the template.

For an example, we go back to the bacterial *recA* gene. The product of this gene, the RecA protein, has several activities (see Chapter 13). It binds to and forms a filamentous structure on DNA, aligns two DNAs of similar sequence, and hydrolyzes ATP. A particular amino acid residue in RecA (a 352-residue polypeptide), Lys⁷², is involved in ATP hydrolysis. By changing this Lys residue to an Arg, a variant of RecA protein is created that will bind, but not hydrolyze, ATP (Figure 7-17c). The engineering and purification of this variant RecA protein has facilitated research into the roles of ATP hydrolysis in the functioning of this protein.

Changes can be introduced into a gene that involve far more than one base pair. Large parts of a gene can be deleted by cutting out a segment with restriction endonucleases and ligating the remaining portions to form a smaller gene. Parts of two different genes can be ligated to create new combinations; the product of such a fused gene is called a **fusion protein**. Researchers now have ingenious methods to bring about virtually

any genetic alteration *in vitro*. After reintroducing the altered DNA into the cell, they can investigate the consequences of the alteration.

Terminal Tags Provide Handles for Affinity Purification

Affinity chromatography is one of the most efficient methods for purifying proteins (see Highlight 4-1). Unfortunately, many proteins do not bind a ligand that can be conveniently immobilized on a column matrix. With the use of fusion proteins, almost any protein can now be purified by affinity chromatography.

The gene encoding the target protein is fused to a gene encoding a peptide or protein that binds a simple, stable ligand with high affinity and specificity. The peptide or protein used for this purpose is referred to as a **tag**. Tag sequences can be added to genes such that the resulting proteins have tags at their N- or C-terminus. Table 7-3 lists some of the peptides or proteins commonly used as tags.

The general procedure can be illustrated by focusing on a system that uses the glutathione-S-transferase (GST) tag (Figure 7-18). GST is a small enzyme (M_r 26,000) that binds tightly and specifically to glutathione. When the GST gene sequence is fused to a target gene, the fusion protein acquires the capacity to bind glutathione. The fusion protein is expressed in a host organism such as a bacterium, and a crude extract is prepared. A column is filled with a porous matrix consisting of the ligand (glutathione) immobilized on microscopic beads of a stable polymer such as cross-linked agarose. As the crude extract percolates through this matrix, the fusion protein becomes immobilized by binding the glutathione. The other proteins in the extract are washed through the column and discarded. The interaction between GST and glutathione is tight but noncovalent, allowing the fusion protein to be gently eluted from the column with a solution containing

Table 7-3 Commonly Used Protein Tags

Tag Protein	Molecular Weight	Immobilized Ligand
Protein A	59,000	Fc portion of IgG
His ₆ (His tag)	800	Ni ²⁺
Glutathione-S-transferase (GST)	26,000	Glutathione
Maltose-binding protein	41,000	Maltose
β-Galactosidase	116,000	p-Aminophenyl-β-D-thiogalactoside (TPEG)
Chitin-binding domain	5,700	Chitin

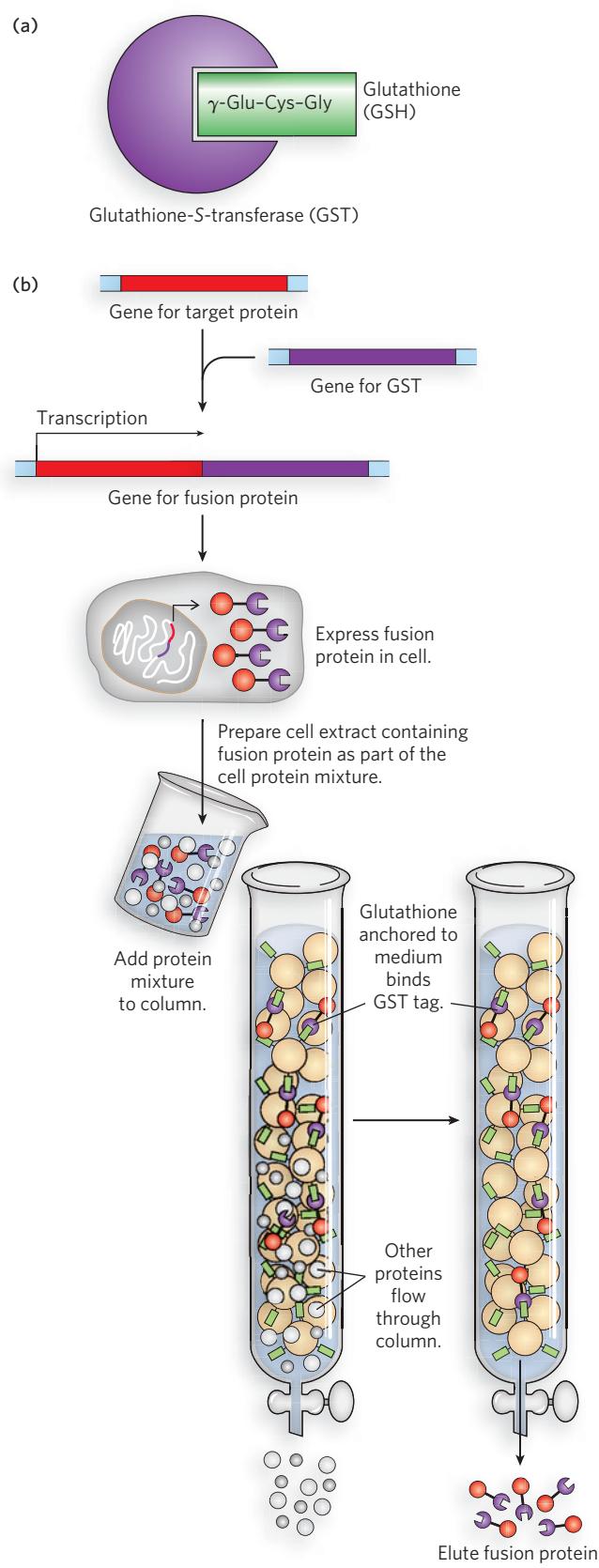


FIGURE 7-18 Use of tagged proteins in protein purification.

The GST tag is fused to the C-terminus of the protein by genetic engineering. The tagged protein is expressed in the cell and is present in the crude extract when the cells are lysed. The extract is subjected to chromatography through a matrix with immobilized glutathione. The GST-tagged protein binds to the glutathione, which retards its migration through the column, while the other proteins are washed through rapidly. The tagged protein is subsequently eluted with a solution containing elevated salt concentration or free glutathione.

either a higher concentration of salts or free glutathione to compete with the immobilized ligand for GST binding. The fusion protein is often obtained with good yield and high purity. In some commercially available systems, the tag can be entirely or largely removed from the purified fusion protein by a protease that cleaves a sequence near the junction between the target protein and its tag.

A shorter tag with widespread application consists of a simple sequence of six or more histidine residues. These histidine tags, or His tags, bind tightly and specifically to nickel ions. A chromatography matrix with immobilized Ni^{2+} can be used to quickly separate a His-tagged protein from other proteins in an extract. Some of the larger tags, such as maltose-binding protein, provide added stability and solubility, allowing the purification of cloned proteins that are otherwise inactive due to improper folding or insolubility.

This technology is powerful and convenient. The tags have been successfully used in thousands of published studies; in many cases, the protein would be impossible to purify and study without the tag. However, even very small tags can affect the properties of the proteins they are attached to, thereby influencing the study results. Even if the tag is removed by a protease, one or a few extra amino acid residues can remain behind on the target protein, which may or may not affect the protein's activity. The types of experiments to be carried out, and the results obtained from them, should always be evaluated with the aid of well-designed controls to assess any effect of a tag on protein function.

SECTION 7.2 SUMMARY

- Genes or other DNA segments can be amplified by the polymerase chain reaction. With specialized adaptations of PCR, investigators can amplify sequences in RNA and quantify the levels of particular RNA molecules in a cell.

- Modern DNA-sequencing methods enable researchers to determine the sequences of entire mammalian genomes in a matter of weeks or even days. Thousands of genomic DNA sequences are now available in public databases.
- Cloned genes can be expressed to provide large amounts of the gene product. Systems have been developed to express genes in bacteria, yeast, insects, mammalian cells, and even entire mammalian organisms.
- Cloned genes can be altered. A gene sequence can be changed, sequences deleted, or sequences added. All changes affect the protein or RNA product of the gene.
- Added sequences can produce protein products that include fused peptide segments, sometimes called tags. With the aid of these tags, the protein can be rapidly purified.

7.3 Understanding the Functions of Genes and Their Products

One of the challenges in molecular biology is to identify the functions of the myriad genes now being discovered in large genome sequencing projects, which we'll survey in Chapter 8. When the complete sequence of an organism's genome becomes available, we often lack functional information for half or more of the defined genes. Modern biotechnology provides some shortcuts to understanding the functions of gene products, and we describe some of the key technologies here (we'll be expanding on their application in Chapter 8).

Protein Fusions and Immunofluorescence Can Localize Proteins in Cells

Often, an important clue to a gene product's function comes from determining its location within the cell. For example, a protein found exclusively in the nucleus could be involved in processes that are unique to that organelle, such as transcription, replication, or chromatin condensation. Researchers often engineer fusion proteins for the purpose of locating a protein in the cell or organism. Some of the most useful fusions involve the addition of marker proteins that allow the investigator to determine the location by direct visualization or by immunofluorescence.

A particularly useful marker is the gene for **green fluorescent protein (GFP)**. A target gene (coding the protein of interest) fused to the GFP gene generates a

fusion protein that is highly fluorescent—it literally lights up when exposed to blue light—and can be visualized directly in a living cell. GFP is a protein derived from the jellyfish *Aequorea victoria* (see How We Know). It has a β -barrel structure (Figure 7-19a), and the fluorophore (the fluorescent component of the protein) is in the center of the barrel. The fluorophore is derived from a rearrangement and oxidation of several amino acid residues. Because this reaction is autocatalytic and requires no other proteins or cofactors (other than molecular oxygen), GFP is readily cloned in an active form in almost any cell. Just a few molecules of this protein can be observed microscopically, allowing the study of its location and movements in a cell. Careful protein engineering, coupled with the isolation of related fluorescent proteins from other marine coelenterates, has made a wide range of these proteins available, with an array of colors (Figure 7-19b) and other characteristics (brightness, stability). With this technology, for example, the protein GLR1 (a glutamate receptor of nervous tissue) has been visualized as a GLR1-GFP fusion protein in the nematode *Caenorhabditis elegans* (Figure 7-19c).

In many cases, visualization of a GFP fusion protein in a live cell is not possible, or is not practical or desirable. The GFP fusion protein may be inactive or may not be expressed at sufficient levels to allow visualization. In this case, **immunofluorescence** is an alternative approach for visualizing the endogenous (unaltered) protein, although fusion proteins are sometimes used. This approach requires the fixation (and thus death) of the cell. The protein of interest is expressed either unaltered or as a fusion protein with an **epitope tag**, a short protein sequence that is bound tightly by a well-characterized, commercially available antibody. Fluorescent molecules (fluorochromes) are attached to this antibody.

More commonly, a second antibody is added that binds specifically to the first one, and it is the second antibody that has the attached fluorochrome(s) (Figure 7-20). A variation of this indirect visualization approach is to attach biotin molecules to the first antibody, then add streptavidin (a bacterial protein closely related to avidin; see Table 5-1) complexed with fluorochromes. The interaction between biotin and streptavidin is one of the strongest and most specific known, and the potential to add multiple fluorochromes to each target protein gives this method great sensitivity.

Highly specialized cDNA libraries (see Figure 7-8) can be made by cloning cDNAs or cDNA fragments into a vector that fuses each cDNA sequence with the sequence for a marker, also called a reporter gene.

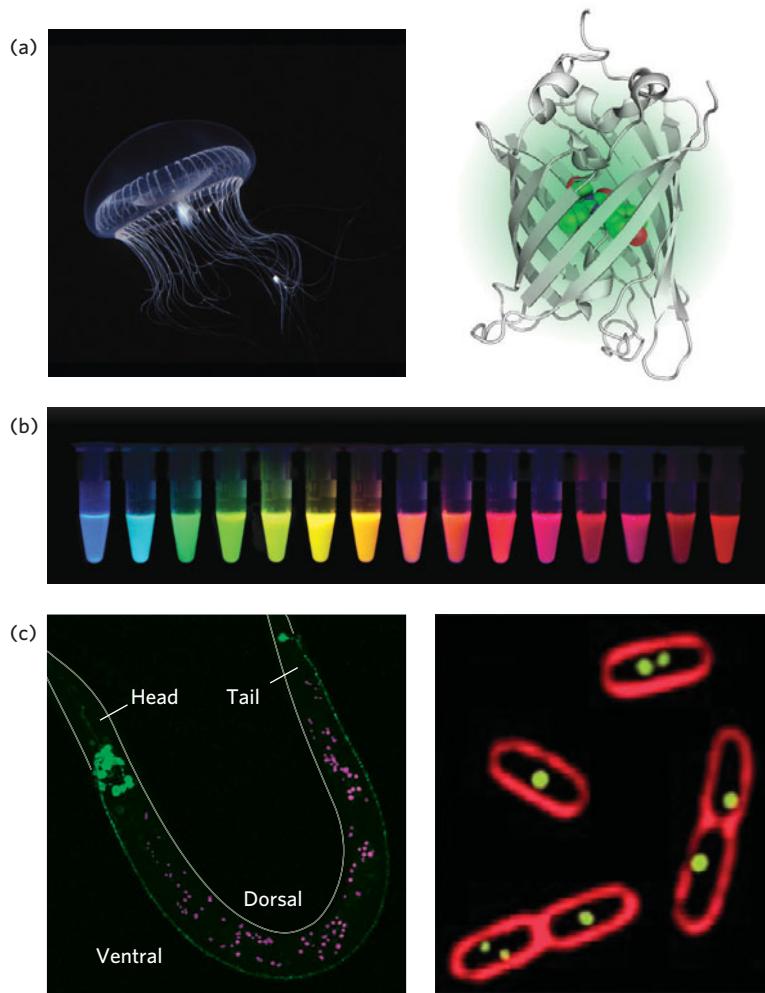


FIGURE 7-19 Green fluorescent protein (GFP). (a) The source of GFP is the jellyfish *Aequorea victoria* (left); the bioluminescent photo-organs are visible. The GFP protein has a β -barrel structure (right); the fluorophore (highlighted) is in the center of the barrel. (b) Variants of GFP are now available in almost any color of the visible spectrum. (c) A GLR1-GFP fusion protein is visible in *Caenorhabditis elegans*, a nematode worm (left). GLR1 is a glutamate receptor of nervous tissue. In the *E. coli* cells, the

membranes are stained with a red fluorescent dye. The cells are expressing a protein that binds to a resident plasmid, fused to GFP. The green spots indicate the locations of plasmids. [Sources: (a) (left) Kevin A. Raskoff. (b) Courtesy of Roger Tsien, B. N. G. Giepmans et al., *Science* 312:217–224, 2006. (c) (left) Courtesy of Penelope J. Brockie and Andres V. Maricq, Department of Biology, University of Utah. (right) courtesy of Joseph A. Pogliano, J. Pogliano et al., *Proc. Natl. Acad. Sci. USA* 98:4486–4491, 2001.]

For example, libraries have been developed in which all the genes in the library are fused to the GFP gene (Figure 7-21). Each cell in the library expresses one of these fused genes. The cellular location of the product of any gene represented in the library can be studied—assuming that the particular fusion protein is expressed at sufficient levels and retains its normal function and location—by examining cells that express the appropriate fused gene, detecting light foci that reveal the protein’s presence.

Proteins Can Be Detected in Cellular Extracts with the Aid of Western Blots

Western blots, also known as immunoblots, make use of antibodies to detect the presence of specific proteins in a biological sample, such as in a certain tissue or at a given time in an organism’s development (Figure 7-22). The antibodies are obtained by purifying the protein of interest, inoculating a chicken or rabbit with the protein, and isolating the resulting antibodies from the

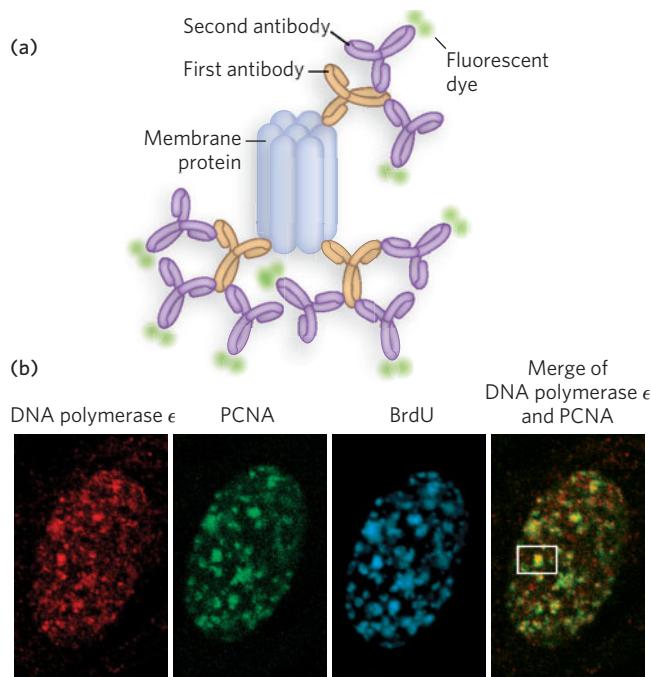


FIGURE 7-20 **Indirect immunofluorescence.** (a) The protein of interest is bound to a first antibody, and a second antibody is added; this second antibody, with one or more attached fluorescent groups, binds to the first. Multiple second antibodies can bind the first antibody, amplifying the signal. If the protein of interest is in the interior of the cell, the cell is fixed and permeabilized, and the two antibodies are added in succession. (b) The end result is an image in which bright spots indicate the location of the protein of interest in the cell. The images show a nucleus from a human fibroblast, stained with antibodies and fluorescent labels for DNA polymerase ϵ , for PCNA, an important polymerase accessory protein, and for bromo-deoxyuridine (BrdU), a nucleotide analog. The patterns of staining show that DNA polymerase ϵ and PCNA colocalize to regions of active DNA synthesis with one example marked by a white box. [Source: (b) J. Fuss and S. Linn, *J. Biol. Chem.* 277:8658–8666, 2002.]

serum. The first few steps are similar to those described for Southern blots (for nucleic acids) in Chapter 6. A protein sample is subjected to electrophoresis in a polyacrylamide gel. The gel is then blotted on a membrane, to which proteins in the gel adhere. The remaining steps are specific for the detection of proteins rather than nucleic acids. The membrane is washed with a solution containing the protein-specific rabbit or chicken antibodies, which bind to the immobilized target protein. A second solution is then added, containing a second antibody that specifically binds to the first (e.g., an antibody derived from goats that binds all rabbit-

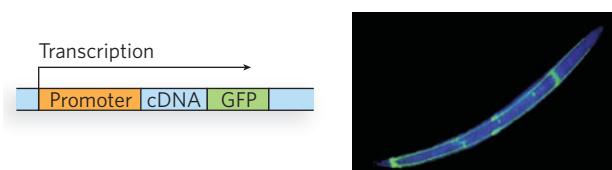


FIGURE 7-21 **Specialized DNA libraries.** Cloning of a cDNA next to the GFP gene creates a reporter construct. Transcription proceeds through the gene of interest (the inserted cDNA) and the reporter gene (here, GFP), and the mRNA transcript is expressed as a fusion protein. The GFP part of the protein is visible with the fluorescence microscope. Although only one example is shown, thousands of genes can be fused to GFP in similar constructs and stored in libraries in which each cell or organism in the library expresses a different protein fused to GFP. If the fusion protein is properly expressed, its location in the cell or organism can be assessed. The photograph shows a nematode worm (*C. elegans*) containing a GFP fusion protein expressed only in the four “touch” neurons that run the length of its body. [Source: Photo courtesy of Kevin Strange, PhD, and Michael Christensen, PhD, Department of Pharmacology, Vanderbilt University Medical Center.]

derived antibodies of the IgG class). The second antibody has an attached radioactive or fluorescent label to allow visualization of protein-antibody complexes. The procedure is sensitive to changes in the amount of target protein present, so increases or decreases in cellular protein levels are readily monitored.

Protein-Protein Interactions Can Help Elucidate Protein Function

Another key to defining the function of a particular protein is to determine what it binds to. In the case of protein-protein interactions, the association of a protein of unknown function with one whose function is known can provide useful and compelling “guilt by association.” The techniques used in this effort are quite varied.

Purification of Protein Complexes With the construction of cDNA libraries in which each gene is fused to an epitope tag, investigators can precipitate the protein product of a gene by complexing it with the antibody that binds the epitope, a process called **immunoprecipitation** (Figure 7-23). If the tagged protein is expressed in cells, other proteins that bind to it precipitate with it. Identifying the associated proteins reveals some of the intracellular protein-protein interactions of the tagged protein. There are many variations of this process. For

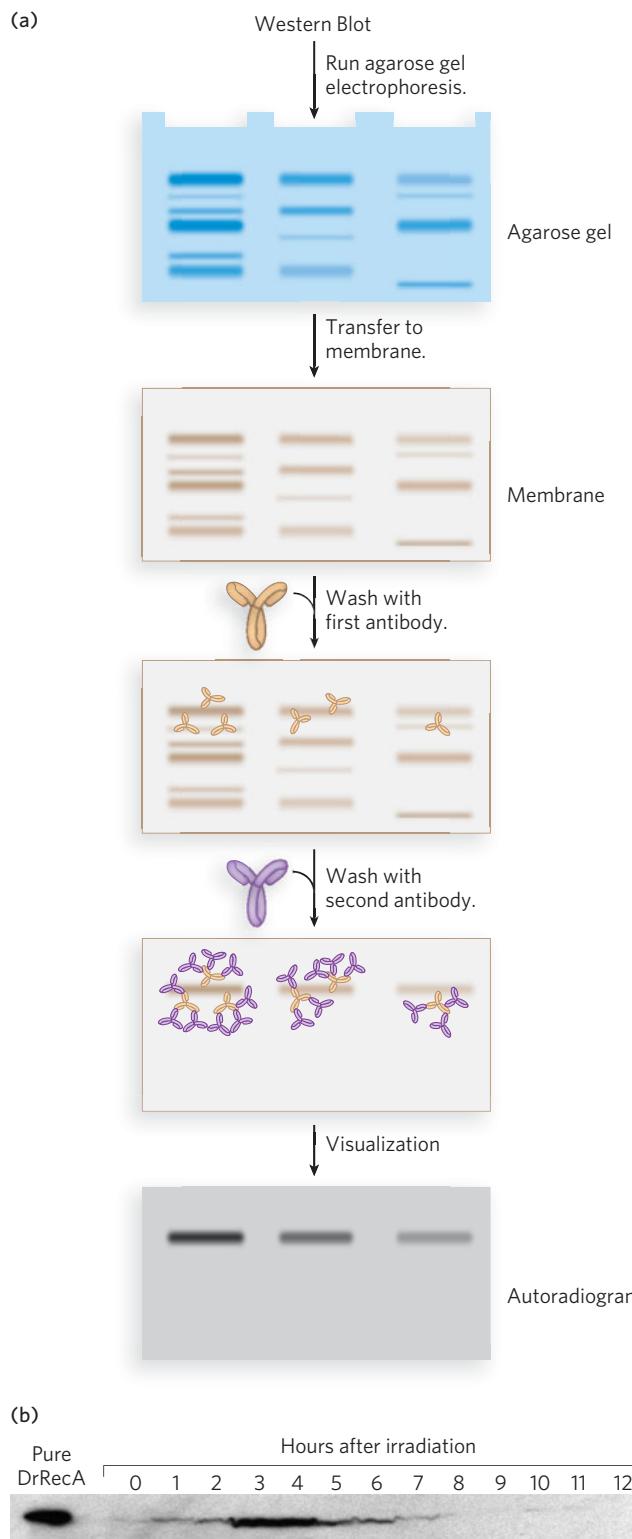


FIGURE 7-22 Western blots. (a) Proteins are subjected to electrophoresis, then transferred from the gel to a membrane. The membrane is washed successively with the first antibody and the second (labeled) antibody, allowing visualization of the protein of interest. (b) A Western blot shows levels of RecA protein in cells of the highly radiation-resistant bacterium *Deinococcus radiodurans*. The *D. radiodurans* protein DrRecA is first induced and then repressed, according to the need for DNA repair, in the hours following high-level irradiation. [Source: (b) Courtesy of Cédric Norais, Department of Biochemistry, University of Wisconsin-Madison.]

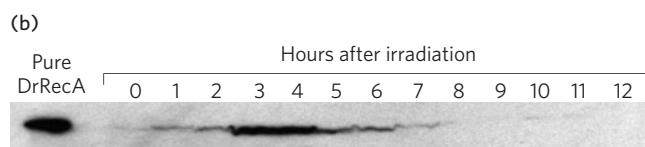
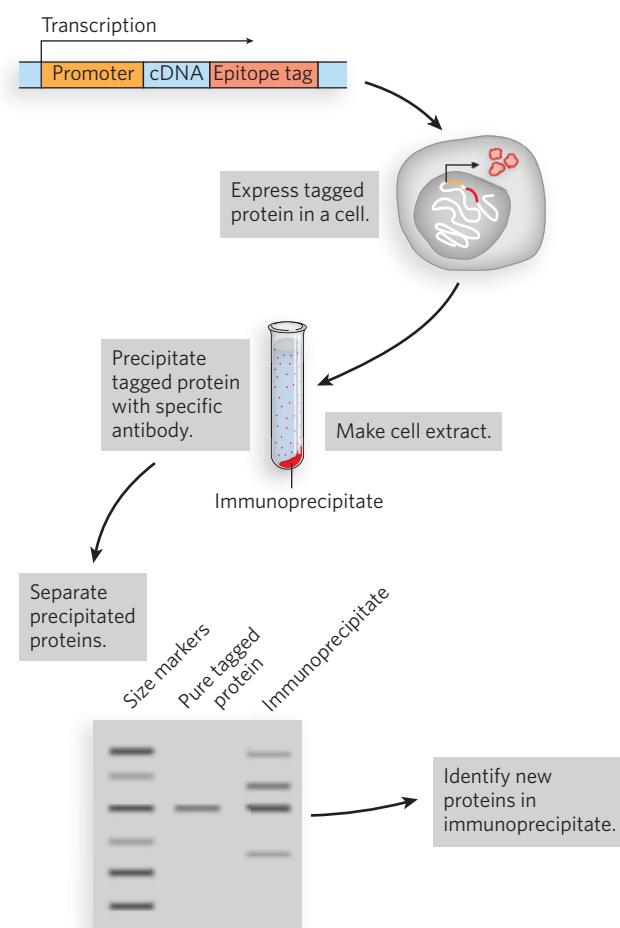


FIGURE 7-23 The use of epitope tags to study protein-protein interactions. The gene of interest is cloned next to a gene for an epitope tag, and the resulting fusion protein is precipitated by antibodies to the epitope. Any other proteins that interact with the tagged protein also precipitate, thereby helping to elucidate protein-protein interactions.

example, a crude extract of cells that express a tagged protein is added to a column containing immobilized antibody. The tagged protein binds to the antibody, and proteins that interact with the tagged protein are sometimes also retained on the column. The connection between the protein and the tag is cleaved with a specific protease, and the protein complexes are eluted from the column and analyzed. Researchers can use these methods to define complex networks of interactions within a cell. In principle, the chromatographic approach to analyzing protein-protein interactions can be used with any type of protein tag (His tag, GST, etc.) that can be immobilized on a suitable chromatographic medium.

The selectivity of this approach has been enhanced with the use of **tandem affinity purification (TAP) tags**. Two consecutive tags, Protein A and calmodulin-binding peptide, are fused to a target protein, and the fusion protein is expressed in a cell (Figure 7-24). A crude extract containing the TAP-tagged fusion protein

is passed through a column matrix with attached IgG antibodies that bind Protein A. Most of the unbound proteins are washed through the column, but proteins associated with the target protein are retained. Protein A is then cleaved from the fusion protein with TEV protease, and the shortened target protein and associated proteins are eluted from the column. The eluent is then passed through a second column containing a matrix of calmodulin beads. Loosely bound proteins are again washed from the column, and the target protein is eluted from the column with its associated proteins. The two consecutive purification steps eliminate any weakly bound contaminating proteins. False positives are minimized, and protein interactions that persist through both steps are likely to be functionally significant.

Yeast Two-Hybrid and Three-Hybrid Analysis A sophisticated genetic approach to defining protein-protein interactions is based on the properties of the Gal4 protein (Gal4p), which activates the transcription of *GAL* genes in yeast (genes encoding the enzymes of galactose metabolism; see Chapter 21). Gal4p has two domains: one that binds a specific DNA sequence and another that activates RNA polymerase to synthesize mRNA from an adjacent gene. The two domains of Gal4p are stable when separated, but activation of RNA polymerase requires interaction with the activation domain, which in turn requires positioning by the DNA-binding domain. Hence, the domains must be brought together to function correctly.

In **yeast two-hybrid analysis**, the protein-coding regions of the genes to be analyzed are fused to the yeast gene for either the DNA-binding domain or the activation domain of Gal4p, and the resulting genes express a series of fusion proteins (Figure 7-25). If a protein fused to the DNA-binding domain interacts with a protein fused to the activation domain, transcription is activated. The reporter gene transcribed by this activation is generally one that yields a protein required for growth or an enzyme that catalyzes a reaction with a colored product. Thus, when grown on the proper medium, cells that contain a pair of interacting proteins are easily distinguished from those that do not. A library can be set up with a particular yeast strain in which each cell in the library has a gene fused to the Gal4p DNA-binding domain gene, and many such genes are represented in the library. In a second yeast strain, a gene of interest is fused to the gene for the Gal4p activation domain. The yeast strains are mated, and individual diploid cells are grown into colonies. This allows for large-scale screening for cellular proteins that interact with the target protein.

The function of many key regulatory proteins, especially in eukaryotic cells, involves specific interactions between the proteins and RNA molecules. A strategy called

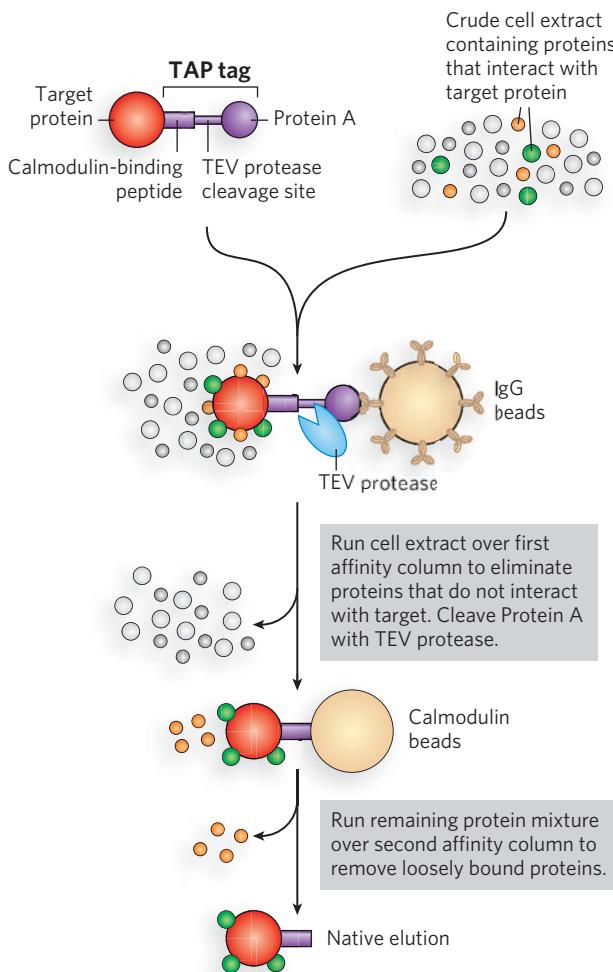


FIGURE 7-24 Tandem affinity purification (TAP) tags. A TAP-tagged protein and associated proteins are isolated by two consecutive affinity purifications. See text for details.

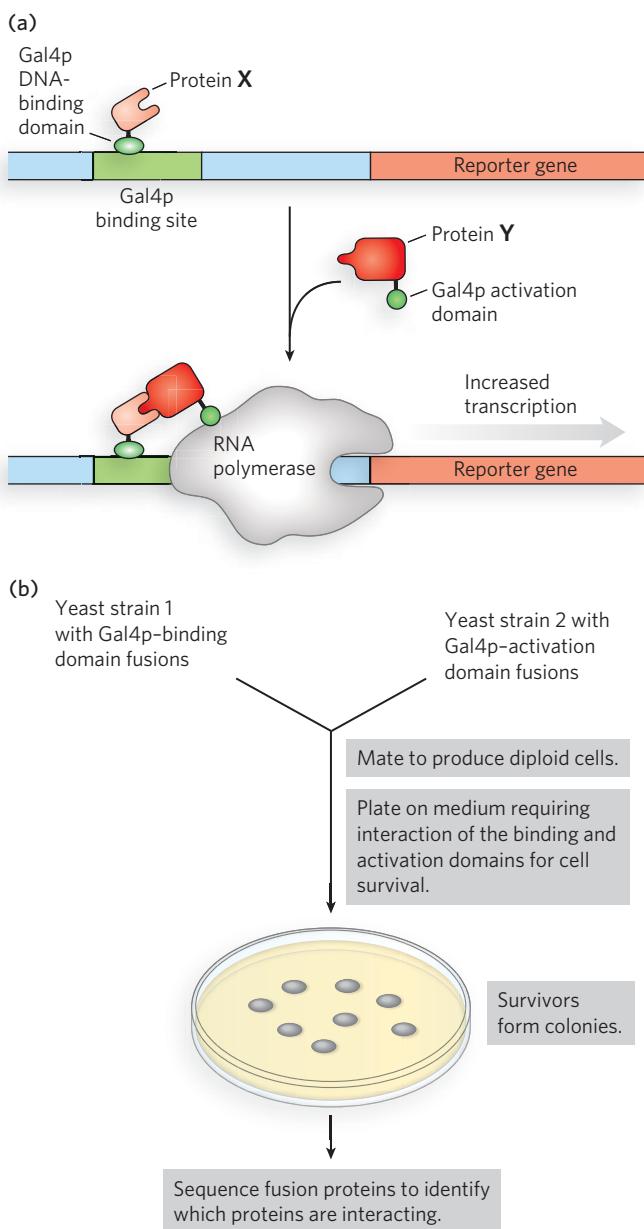


FIGURE 7-25 Yeast two-hybrid analysis. (a) Interaction of X and Y results in the expression of a reporter gene. (b) The two gene fusions are created in separate yeast strains, which are then mated and plated on selective media. Thus, all surviving colonies have interacting fusion proteins.

yeast three-hybrid analysis has been developed to screen for protein-RNA interactions (Figure 7-26). For a known RNA-binding protein, this method yields a rapid identification of all or most of the RNAs that the protein binds. The method uses three engineered elements: two fusion proteins and a plasmid library. The fusion proteins are (1) the protein of interest fused to the Gal4p transcription-activation domain, and (2) the DNA-binding

domain of a protein known as LexA fused to an RNA-binding protein called MS2. The LexA portion binds to a specific DNA sequence, and MS2 binds tightly to an RNA hairpin with a defined sequence. The DNA-binding site for the LexA protein is placed upstream of a reporter gene. The third element, the plasmid library, consists of the gene encoding the RNA hairpin recognized by MS2, fused to random sequences. When transcribed, each MS2 RNA is fused to another RNA segment. If a particular expressed RNA is bound by the protein of interest, the RNA serves

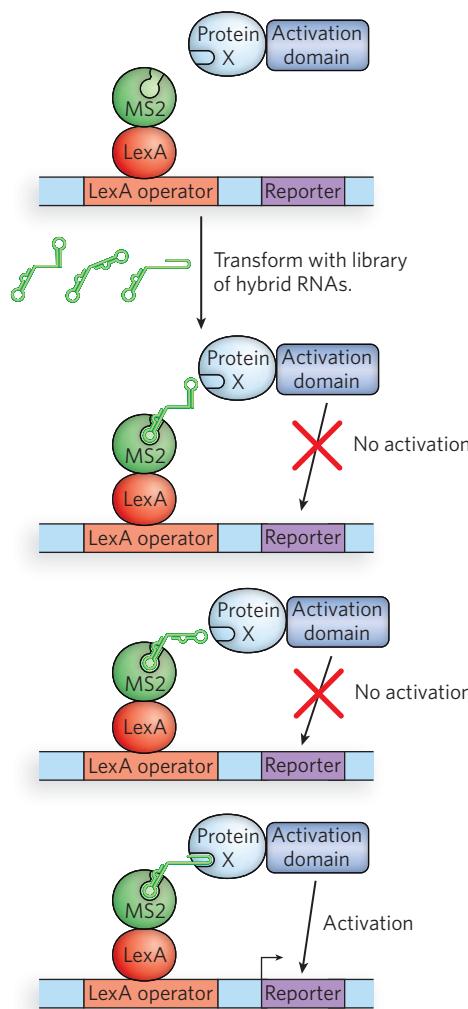


FIGURE 7-26 Yeast three-hybrid analysis. Two fusion proteins bind simultaneously to a hybrid RNA molecule to permit expression of a reporter gene. An RNA library consisting of random-sequence RNA segments fused to a hairpin recognized by MS2 is screened. If the protein of interest (X) binds the random RNA sequence expressed in a given cell, that cell will survive and produce a colony.

as a tether, linking the first fusion protein with the second and activating transcription of the reporter gene. With this method, a few dozen RNA molecules that specifically bind the target protein can be isolated from a library containing millions of cloned RNAs.

These techniques for determining cellular localization and molecular interactions provide important clues to protein function. However, they do not replace classical biochemistry and molecular biology. They simply give researchers an expedited entrée into important new biological problems. When paired with the simultaneously evolving tools of biochemistry and molecular biology, the techniques described here are speeding the discovery not only of new proteins, but of new biological processes and mechanisms.

DNA Microarrays Reveal Cellular Protein Expression Patterns and Other Information

Major refinements of the technology underlying DNA libraries, PCR, and hybridization have come together in the development of **DNA microarrays**, which allow the rapid and simultaneous screening of many thousands of genes. DNA segments from genes of known sequence, a few dozen to hundreds of base pairs long, are amplified by PCR. Robotic devices then accurately deposit nanoliter quantities of the DNA solutions on a solid surface of just a few square centimeters, in a predesigned array, with each of the thousands of spots containing sequences derived from a particular gene. An alternative and increasingly common strategy is to synthesize DNA directly on the solid surface, using photolithography (Figure 7-27).

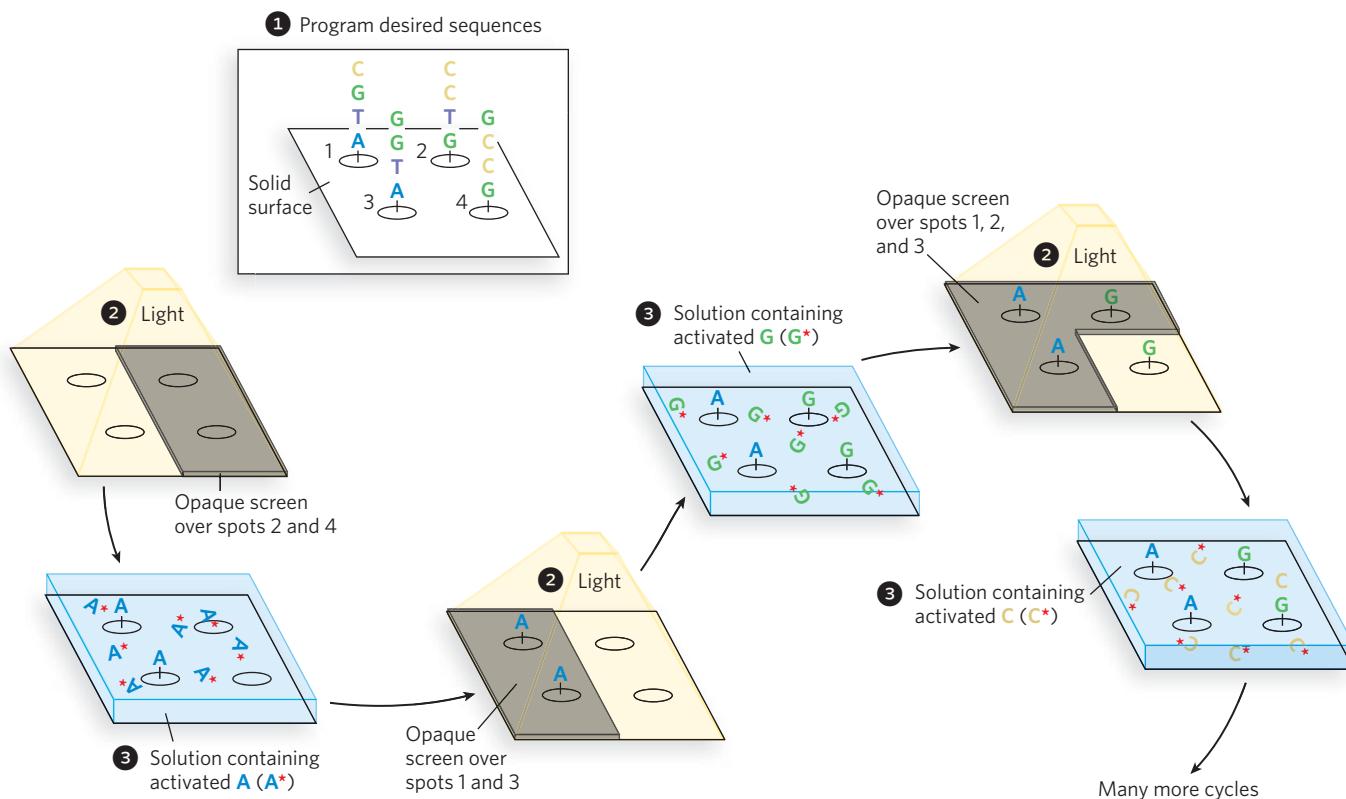


FIGURE 7-27 Photolithography to create a DNA

microarray. (1) A computer is programmed with the desired oligonucleotide sequences. (2) The reactive groups, attached to a solid surface, are initially rendered inactive by photoactive blocking groups, which can be removed by a flash of light. An opaque screen blocks the light from certain groups, preventing their activation. (3) A solution containing one activated nucleotide (e.g., A^*) is washed over the spots. The 5' hydroxyl of the nucleotide is blocked to prevent unwanted reactions, and the nucleotide links to the surface groups at the appropriate spots through its 3'

hydroxyl. The surface is washed successively with solutions containing each remaining activated nucleotide (G^* , C^* , T^*). The 5'-blocking groups on each nucleotide limit the reactions to addition of one nucleotide at a time, and these groups can also be removed by light. Once each spot has one nucleotide, a second nucleotide can be added to extend the nascent oligonucleotide at each spot, using screens and light to ensure that the correct nucleotides are added at each spot in the correct sequence. This continues until the required sequences are built up on each spot on the surface.

The resulting array, or chip, may include sequences derived from every gene of a bacterial or yeast genome, or selected families of genes from a larger genome. Once constructed, the microarray can be probed with mRNAs or cDNAs from a particular cell type or cell culture to identify the genes being expressed in those cells.

A microarray can provide a snapshot of all the genes in an organism, informing the researcher about the genes that are expressed at a given stage in the organism's development or under a particular set of environmental conditions. For example, the total complement of mRNA can be isolated from cells at two different stages of development and converted to cDNA with reverse transcriptase. With the use of fluorescently labeled deoxyribonucleotides, the two cDNA samples can be made so that one fluoresces red, the other green (Figure 7-28). The cDNA from the two samples is mixed and used to probe the microarray. Each cDNA anneals to only one spot on the microarray, corresponding to the gene encoding the mRNA that gave rise to that cDNA. Spots that fluoresce green represent genes that produce mRNAs at higher levels at one developmental stage; those that fluoresce red represent genes expressed at higher levels at another stage. If a gene produces mRNAs that are equally abundant at both stages of development, the corresponding spot fluoresces yellow. By using a mixture of two samples to measure relative rather than absolute sequence abundance, the method corrects for variations in the amount of DNA originally deposited in each spot on the grid, as well as other possible inconsistencies among spots in the microarray. The spots that fluoresce provide a snapshot of all the genes being expressed in the cells at the moment they were harvested—gene expression examined on a

genome-wide scale. For a gene of unknown function, the time and circumstances of its expression can provide important clues about its role in the cell.

An example of this technique is illustrated in (Figure 7-29), which shows the dramatic results that microarray experiments can produce. Segments from each of the roughly 6,500 genes in the completely sequenced yeast genome were separately amplified by PCR, and each segment was deposited in a defined

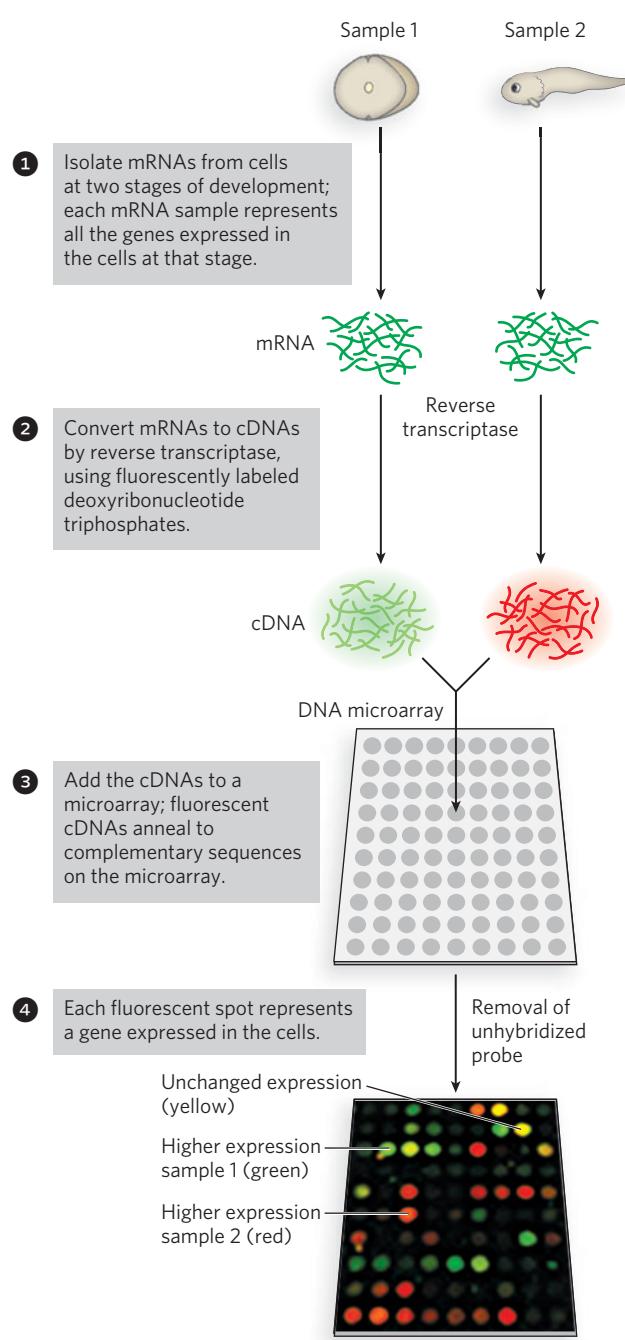


FIGURE 7-28 A DNA microarray experiment. A microarray can be prepared from any known DNA sequence, from any source. Once the DNA is attached to a solid support, the microarray can be probed with other, fluorescently labeled nucleic acids. Here, mRNA samples are collected from cells of a frog at two different stages of development: single-cell stage (left) and a later stage (right). The cDNA probes are synthesized with nucleotides that fluoresce in different colors for each sample; a mixture of the cDNAs is used to probe the microarray. The probes anneal to spots containing complementary DNA. Thus, if the spot lights up, the corresponding gene is represented in the pool of mRNA used to produce the probes. Green spots represent mRNAs more abundant at the single-cell stage; red spots, sequences more abundant later in development. The yellow spots indicate approximately equal abundance at both stages.

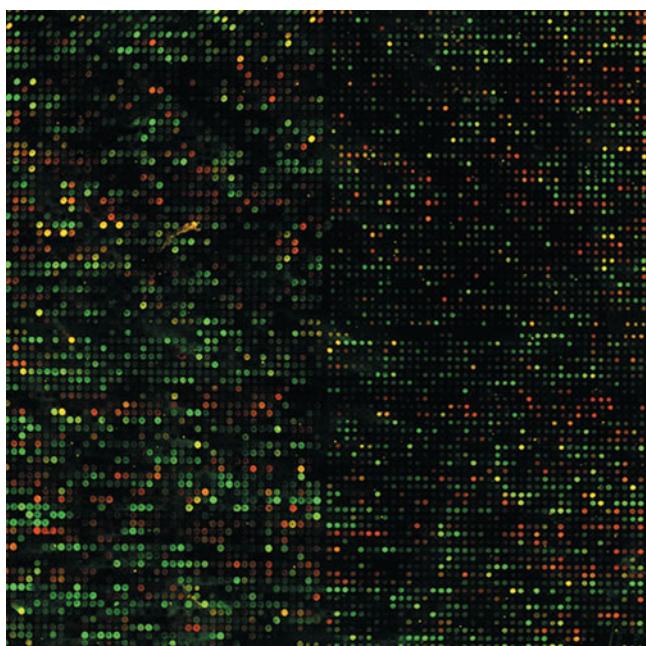


FIGURE 7-29 Enlarged image of a DNA microarray. Each glowing spot contains DNA from one of the roughly 6,500 genes of the yeast (*S. cerevisiae*) genome, with every gene represented in the array. The microarray has been probed with fluorescently labeled nucleic acid derived from mRNAs obtained when the cells were growing in culture (green) and 5 hours after they began to form spores (red). This image is enlarged; the microarray actually measures only 1.8 cm × 1.8 cm. [Source: Courtesy of Patrick O. Brown, Department of Biochemistry, Stanford University School of Medicine.]

pattern to create the microarray. In a sense, this array provides a snapshot of the entire yeast genome. Microarrays are an invaluable tool in systems biology, a topic to be explored in more depth in Chapter 8.

SECTION 7.3 SUMMARY

- The role of a gene of unknown function can be explored by techniques that assess the location of the gene product in the cell, the interactions of the gene product with proteins or RNA, and the expression of the gene under different cellular circumstances.
- By fusing a gene of interest with the genes that encode green fluorescent protein or epitope tags, researchers can visualize the cellular location of the gene product, either directly or by immunofluorescence.

- The presence of particular proteins in biological samples can be detected with the aid of Western blots.
- The interactions of a protein with other proteins or RNA can be investigated with epitope tags and immunoprecipitation or affinity chromatography. Alternatively, interactions in the cell can be probed in yeast two-hybrid and three-hybrid analyses.
- Expression patterns of genes can be probed through the use of microarrays.

Unanswered Questions

As powerful as biotechnology has become, there are still limitations to the reach of molecular biology.

1. Can we make personalized medicine a reality?

The DNA-sequencing methods described in Highlight 7-2 are revolutionary, but they are not adequate to make routine genome sequencing for individual humans practical. The rate at which DNA-sequencing methods are advancing may soon make this statement obsolete. Besides the technological developments, much thought is required on what we will do with all this DNA sequence information when we get it. Ethical considerations must play a role in these decisions.

2. How do we apply biotechnology to routinely and safely alter human genomes? Another limitation faced by molecular biologists concerns the methods available to alter genomes, especially in complex organisms. Genomic alterations are possible even in mammals, but they are laborious. In addition, the desired genomic change is often accompanied by unwelcome, deleterious changes. With more efficient methods, the treatment of genetic disease by human gene therapy may become practical.

3. Can we find out where proteins function in a cell, and what they do? There is also much room for improvement in the methods based on fusion proteins, now used to track the location and function of proteins in cells. Too often, the protein tags used for this work have unintended effects on the activity of the protein to which the tag is attached. The GFP protein works well in many cases, but smaller and brighter tags would be very helpful, along with new ways to link them to proteins.

How We Know

New Enzymes Take Molecular Biologists from Cloning to Genetically Modified Organisms

Cohen, S.N., A.C.Y. Chang, H.W. Boyer, and R.B. Helling. 1973. Construction of biologically functional bacterial plasmids in vitro. *Proc. Natl. Acad. Sci. USA* 70:3240-3244.

Jackson, D.A., R.H. Symons, and P. Berg. 1972. Biochemical method for inserting new genetic information into DNA of simian virus 40: Circular SV40 DNA molecules containing lambda phage genes and the galactose operon of *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* 69:2904-2909.

DNA cloning seems commonplace now, but the idea of combining DNA from two different species created quite a stir in the early 1970s. A convergence of research advances in many laboratories gave rise to a technological revolution.

The first experiment to join two DNA molecules together used a rather laborious strategy. Peter Lobban, a graduate student in Dale Kaiser's lab at Stanford, first used the enzyme terminal transferase to add "tails" of A residues to one DNA fragment and T residues to another. The poly(A) and poly(T) tails annealed and the fragments could be covalently joined with DNA ligase (Figure 1). This experiment linked two segments of DNA derived from a bacterial virus, P-22. Paul Berg's lab soon used the same strategy to link two DNA segments from different species, one from the bacterial virus lambda and the other from the simian virus SV40. The Berg paper was published in 1972.

Meanwhile, Stanley Cohen, also at Stanford, began to study DNA plasmids that made certain disease-causing bacteria resistant to antibiotics (a problem then, as it remains today). His lab soon developed methods for isolating the plasmids and reintroducing them into other bacteria. To get at the genes causing the antibiotic resistance and other features of the plasmid, Cohen wanted to take the plasmids apart and reassemble them. Rather than use the laborious poly(A)-poly(T) tail approach, Cohen relied on restriction enzymes that had recently been discovered. Herbert Boyer and his colleagues at UCSF had shown that one of these enzymes, EcoRI, cleaves DNA asymmetrically at a 6 bp palindrome (see Table 7-2). The resulting sticky ends could guide the rejoicing of the ends by DNA ligase.

At a November 1972 meeting in Hawaii, Cohen and Boyer hatched a collaboration that led to DNA cloning. They settled on a plasmid called pSC101 from the Cohen lab, which encoded a gene that conferred resistance to tetracycline. The collaborators found that Boyer's EcoRI enzyme cut the plasmid only once, and not in a region that affected the sequences needed for either replication

or tetracycline resistance. By early 1973, they had demonstrated that DNA segments from any source, also derived from cleavage by EcoRI, could be linked to this plasmid. The plasmid, in turn, could be reintroduced and propagated in bacteria. These advances, reported in the *Proceedings of the National Academy of Sciences* (USA) in late 1973, gave rise to the first DNA cloning patents and a new company (Genentech). More importantly, they set the stage for the rise of biotechnology.

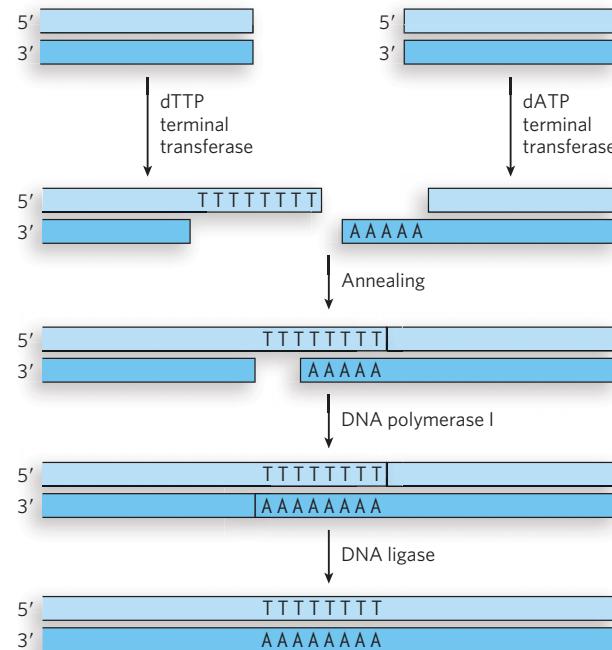


FIGURE 1 A homopolymer tail was added to each DNA segment with the aid of the enzyme terminal transferase. A poly(A) tail was added to one segment, a poly(T) tail to the other. The two complementary tails were annealed. Gaps were filled with the aid of DNA polymerase I, and the DNA was joined by DNA ligase.

A Dreamy Night Ride on a California Byway Gives Rise to the Polymerase Chain Reaction

Brock, T.D., and H. Freeze. 1969. *Thermus aquaticus* gen. n. and sp. n., a nonsporulating extreme thermophile. *J. Bacteriol.* 98:289–297.

Saiki, R.K., S. Scharf, F. Faloona, K.B. Mullis, G.T. Horn, H.A. Erlich, and N. Arnheim. 1985. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle-cell anemia. *Science* 230:1350–1354.

Large advances are sometimes driven by inspiration. In the spring of 1983, Kary Mullis was an employee of the Cetus Corporation in northern California. Hired in 1979 to synthesize oligonucleotides, Mullis discovered that as oligonucleotide synthesis became increasingly automated, he had more and more time to contemplate other projects. He became interested in methods to detect small sequence differences in human DNA, but initially did not make much progress. The idea for the polymerase chain reaction occurred to Mullis one night, in April 1983, as he drove with a friend up the coast. As he described it: he stopped the car and started drawing—DNA molecules hybridizing and lengthening, a chain reaction in which the products of one cycle became the templates for the next.

The first experiment was carried out a few months later, and the first report of the polymerase chain reaction came out in a 1985 paper in *Science* describing a new procedure for detecting the hemoglobin mutation that caused sickle-cell anemia. The method was spelled out in more detail in publications appearing over the next two years. In the early trials, the polymerase used was an active fragment of the *E. coli* DNA polymerase I (see Chapter 11). The heating required to denature the DNA after each PCR cycle inactivated this polymerase, so it had to be added again after each cycle.

A side story shows how basic research can contribute to major advances in surprising ways. About two decades before the development of PCR, microbiologist Thomas Brock (at the University of Wisconsin-Madison) initiated some studies of organisms in the hot springs of Yellowstone National Park (Figure 2). In the fall of 1966, he succeeded in culturing a bacterium from a pool called Mushroom Spring, an organism that grew at higher temperatures than were thought possible for living organisms. The new thermophilic bacterium



FIGURE 2 Hot springs in Yellowstone National Park, one of which is shown here, were the source of the bacterium *Thermus aquaticus*. [Source: Fox71/Dreamstime.]

was subsequently named *Thermus aquaticus*. The heat stability of the proteins in *T. aquaticus* became highly important to the development of PCR. The DNA polymerase from this bacterium (*Taq* polymerase) is stable at very high temperatures and is not inactivated by the heating and cooling cycles that are needed to denature and reanneal the DNA during PCR. The incorporation of *Taq* polymerase into the PCR protocol in the late 1980s allowed the entire procedure to be automated.

PCR methods were quickly optimized, and new protocols gradually expanded the possible applications. By the end of the 1980s, the technology had utterly transformed the biological sciences.

Coelenterates Show Biologists the Light

Chalfie, M., Y. Tu, G. Euskirchen, W.W. Ward, and D. C. Prasher. 1994. Green fluorescent protein as a marker for gene expression. *Science* 263:802-805.

In 1960, shortly after joining the faculty at Princeton University, Osamu Shimomura began to study the bioluminescence produced by the jellyfish *Aequorea victoria*. Traveling regularly to the state of Washington to secure specimens, he would take the photo-organs from 20 or 30 jellyfish and squeeze them through rayon gauze. His "squeezate" was slightly luminescent, and he began to purify the molecules responsible. In 1962, his lab reported the purification of a protein associated with green fluorescent protein, which they called aequorin. The first reference to green fluorescent protein appears in a footnote in that paper, in which Shimomura described isolating from squeezates a protein that formed solutions that looked greenish in sunlight and yellowish under tungsten lights, and showed very bright, greenish fluorescence in UV light. His subsequent studies gradually showed that GFP had a special property: it contained all the chemistry needed to emit fluorescence on its own. Up to that time, most other proteins known to produce bioluminescence, such as firefly luciferase, required the addition of other molecules to do so.

Douglas Prasher, at the Woods Hole Oceanographic Institution, was the first to appreciate the potential of fusing GFP to another protein and using its fluorescence as a cellular marker for that protein. He succeeded in cloning the gene for GFP and determining its sequence, reporting this advance in 1992. Martin Chalfie, at Columbia University, had also seen the potential of GFP. Collaborating with Prasher, he expressed GFP in *E. coli*, reporting the results of the work in *Science* in 1994 (Figure 3). This work realized the potential of the system, showing that GFP expressed in a host organism produced fluorescence without the need for any other protein or factor from the jellyfish.



FIGURE 3 The bacteria on the right side of this agar plate are glowing as a result of the expression of green fluorescent protein. [Source: M. Chalfie et al., *Science* 263:802-805, 1994.]

Sergey Lukyanov, in Moscow, showed that variants of GFP could be cloned from the nonbioluminescent Anthozoa of coral reefs, and he managed to expand the color range of this protein class by cloning a red fluorescent protein.

Much of what we now know about the chemistry and general utility of GFP and other fluorescent proteins has resulted from the subsequent work of Roger Tsien, at the University of California, San Diego. His laboratory has constructed many mutants of GFP that produce the range of colors seen in Figure 7-19b. Many of these mutants also improve on the stability and brightness of the fluorophores in the proteins. Since the mid-1990s, GFP and its variants have become a staple of molecular and cellular biology, illuminating the mechanisms and pathways of countless other cellular proteins and processes.

Key Terms

- DNA cloning, p. 217
 cloning vector, p. 217
 DNA ligase, p. 217
 recombinant DNA, p. 217
 recombinant DNA technology, p. 217
 genetic engineering, p. 217
 Type II restriction endonuclease, p. 218
 plasmid, p. 220
 origin of replication (*ori*), p. 221
 transformation, p. 221
 electroporation, p. 221
 selectable marker, p. 221
- screenable marker, p. 221
 bacterial artificial chromosome (BAC), p. 222
 shuttle vector, p. 223
 yeast artificial chromosome (YAC), p. 223
 pulsed field gel electrophoresis, p. 223
 DNA library, p. 224
 complementary DNA (cDNA), p. 225
 polymerase chain reaction (PCR), p. 226
 Sanger method, p. 228
- expression vector, p. 233
 transfection, p. 237
 site-directed mutagenesis, p. 239
 fusion protein, p. 240
 tag, p. 240
 green fluorescent protein (GFP), p. 242
 immunofluorescence, p. 242
 epitope tag, p. 242
 tandem affinity purification (TAP) tag, p. 246
 yeast two-hybrid analysis, p. 246
 yeast three-hybrid analysis, p. 247
 DNA microarray, p. 248

Problems

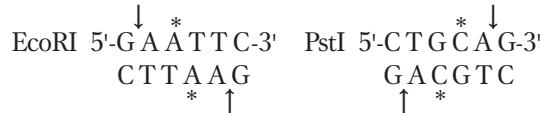
- 1.** When joining two or more DNA fragments, a researcher can adjust the sequence at the junction(s) in a variety of subtle ways, as seen in the following exercises.
- Draw the structure of each end of the linear DNA fragments produced by an EcoRI restriction digest (include the sequences remaining from the EcoRI recognition sequence).
 - Draw the structure resulting from the reaction of these end sequences with a DNA polymerase and the four deoxynucleoside triphosphates.
 - Draw the sequence produced at the junction that arises when the two ends with the structure derived in (b) are ligated.
 - Draw the structure produced when the structure derived in (a) is treated with a nuclease that degrades only single-stranded DNA.
 - Draw the sequence of the junction produced when an end with structure (b) is ligated to an end with structure (d).
 - Draw the structure of the end of a linear DNA fragment produced by a PvuII restriction digest (include the sequences remaining from the PvuII recognition sequence).
 - Draw the sequence of the junction produced when an end with structure (b) is ligated to an end with structure (f).
 - Suppose you can synthesize a short duplex DNA fragment with any sequence you wish. With this synthetic fragment and any of the procedures described in (a) through (g), design a protocol that would remove an EcoRI restriction site from a DNA molecule and insert a new BamHI restriction site at approximately the same location. (See Figure 7-2.)
 - Design four different short, synthetic double-stranded DNA fragments that would permit ligation of structure (a) with a DNA fragment produced by a

PstI restriction digest. In one of these fragments, design the sequence so that the final junction contains the recognition sites for both EcoRI and PstI. In the second and third fragments, design the sequences so that the junction contains only the EcoRI and only the PstI recognition site, respectively. Design the sequence of the fourth fragment so that neither the EcoRI nor the PstI recognition site appears in the junction.

- 2.** The partial sequence of one strand of a double-stranded DNA molecule is:

5' — — GACGAAGTGCTGCAGAAAGTCC
 GCGTTATAGGCATGAATTCTGAGG — — 3'

The cleavage sites for the restriction enzymes EcoRI and PstI are shown below.

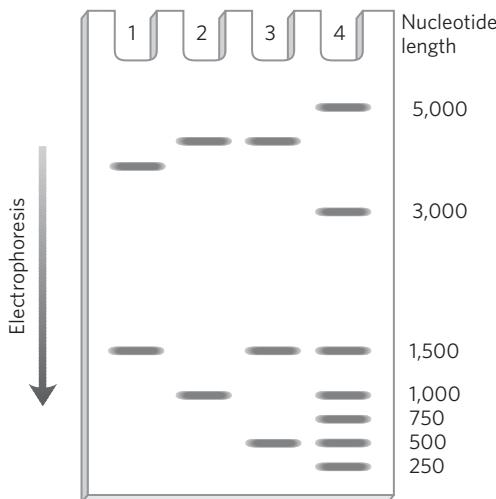


Write the sequence of *both strands* of the DNA fragment created when this DNA is cleaved with both EcoRI and PstI. The top strand of your duplex DNA fragment should be derived from the strand sequence given above.

- 3.** When cloning a foreign DNA fragment into a plasmid, it is often useful to insert the fragment at a site that interrupts a selectable marker (such as the tetracycline-resistance gene of pBR322). The loss of function of the interrupted gene can be used to identify clones containing recombinant plasmids with foreign DNA. With a YAC vector, a selectable marker is not necessary to easily distinguish cells carrying vectors that incorporate large

foreign DNA fragments from those that do not. How are these recombinant vectors identified?

4. The plasmid cloning vector pBR322 (see Figure 7-4) is cleaved with the restriction endonuclease PstI. An isolated DNA fragment from a eukaryotic genome (also produced by PstI cleavage) is added to the prepared vector and ligated. The mixture of ligated DNAs is then used to transform bacteria, and plasmid-containing bacteria are selected by growth in the presence of tetracycline.
- (a) In addition to the desired recombinant plasmid, what other types of plasmids might be found among the transformed bacteria that are tetracycline resistant? How can the types be distinguished?
- (b) The cloned DNA fragment is 1,000 bp long and has an EcoRI site 250 bp from one end. Three different recombinant plasmids are cleaved with EcoRI and analyzed by gel electrophoresis, giving the patterns shown below. What does each pattern say about the cloned DNA? Note that in pBR322, the PstI and EcoRI restriction sites are about 750 bp apart. The entire plasmid with no cloned insert has 4,361 bp. Size markers in lane 4 have the number of nucleotides noted.



5. A new restriction endonuclease is discovered that recognizes and cleaves the palindromic sequence GGATATCC. How often does this sequence appear in a random-sequence DNA in which all four nucleotides are present in equal amounts? In a random-sequence DNA in which the G + C content is 80%, will the frequency with which this site appears increase or decrease?
6. A BAC vector is designed so that large DNA fragments can be inserted into a cleavage site for the enzyme BamHI (see Table 7-2). To prepare chromosomal DNA from a target organism for cloning into this vector, the target DNA is treated briefly with BamHI, not long enough to cleave all of the BamHI sites present. Explain why the BamHI reaction is halted before the chromosomal DNA is completely cleaved.

7. One strand of a chromosomal DNA sequence is shown below. An investigator wants to amplify and isolate a DNA fragment defined by the segment shown in red, using the polymerase chain reaction. Design two PCR primers, each 20 nucleotides long, that can be used to amplify this DNA segment.

5' - - -AATGCCGTAGCCGATCTG**CCTCGAGTCAACTC
GATGCTGGTAACTTGGGGTATAAGCTTACCCATGG
TATCGTAGTTAGATTGATTGTTAGGTTCTTAGGTTA
GGTTTCTGGTATTGGTTAGGGTCTTGATGCTATT
ATTGTTGGTTTGATTTGGCTTTATATGGTTATG
TTTAAGCCGGTTTGTCTGGGATGGTTCGTCTGAT
GTGCGCGTAGCGTGCAGCG - - - 3'**

8. A researcher wants to amplify the same DNA segment as described in Problem 7. However, to aid in cloning, she wants to add a short DNA sequence on each end of the amplified segment that includes the restriction site for the enzyme EcoRI. Design the two PCR primers that this researcher needs, incorporating 20 nucleotides complementary to the appropriate target sequences.
9. Huntington disease (HD) is an inherited neurodegenerative disorder characterized by the gradual, irreversible impairment of psychological, motor, and cognitive functions. Symptoms typically appear in middle age, but onset can occur at almost any age, and the course of the disease can range from 15 to 20 years. The molecular basis of HD is becoming better understood, and the genetic mutation has been traced to a gene that encodes a protein (M_r 350,000) of unknown function. In individuals who will not develop HD, a region of the gene that encodes the N-terminus of this protein has a sequence of CAG codons (for glutamine) repeated 6 to 39 times in succession. In individuals with adult-onset HD, this codon is typically repeated 40 to 55 times. In those with childhood-onset HD, the codon is repeated more than 70 times. Thus, the length of this simple trinucleotide repeat indicates whether an individual will develop HD and at approximately what age the first symptoms will occur.

A small portion of the N-terminal coding sequence of the 3,143-codon HD gene is given below. The nucleotide sequence of the DNA is shown in black, the amino acid sequence corresponding to the gene is in blue, and the CAG repeat is shaded (the numbers at left indicate the starting nucleotide and amino acid residue numbers in that row). Outline a PCR-based test for HD that could be carried out on a blood sample. Assume that each PCR primer must be 25 nucleotides long. By convention, unless otherwise specified, a DNA sequence encoding a protein is written with the coding strand—the sequence identical to the mRNA transcribed from the gene (except for U in the mRNA in place of T)—on top, such that it is read 5' → 3', left to right.

Source: The Huntington's Disease Collaborative Research Group, *Cell* 72:971-983, 1993.

- 10.** In a species of ciliated protist, a segment of genomic DNA is sometimes deleted. The deletion is a genetically programmed reaction associated with cellular mating. A researcher proposes that the DNA is deleted in a type of recombination called site-specific recombination (see Chapter 14), with the DNA on either side of the segment joined together and the deleted segment forming a circular DNA reaction product. Suggest how the researcher might use PCR to detect the presence of the circular, deleted DNA in an extract of the protist.

11. The short DNA shown below is to be sequenced. Using your knowledge of how the Sanger method works, in the gel diagram, draw in the bands that will appear when DNA polymerase is added to the reaction along with the four different nucleotide mixtures indicated. Note that some of these mixtures are *not* what would normally be used in a sequencing reaction. Dideoxynucleotides (ddNTPs) are added in relatively small amounts. The asterisk represents a radioactive label.

*5' - - - 3'-OH

3' - - - ACGACGCAGGACATTAGAC-5'

Nucleotide mixtures:

- A. dATP, dTTP, dCTP, dGTP, ddTTP (given)
 - B. dATP, dTTP, dCTP, dGTP, ddATP
 - C. dTTP, dGTP, dCTP, ddCTP, ddATP
 - D. dATP, dCTP, dTTP, ddGTP



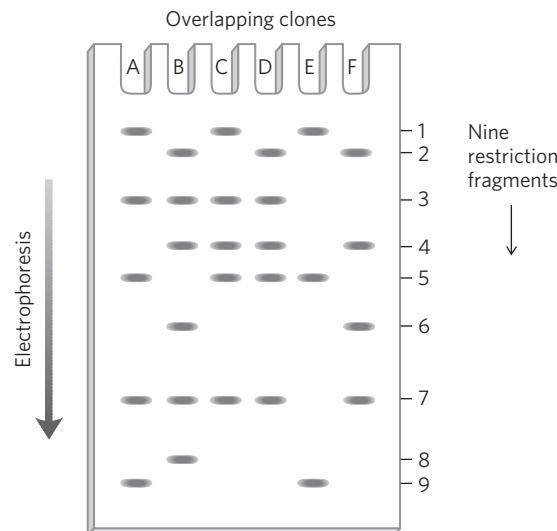
- 12.** To express a cloned gene, the DNA encoding that gene is placed downstream of a bacterial promoter. RNA poly-

merase binds at the promoter and moves away from it in one direction, synthesizing RNA, using one strand of the DNA as a template. The synthesized RNA strand (mRNA) carries the information specifying a protein to the ribosome. The RNA strand is synthesized in the $5' \rightarrow 3'$ direction. The mRNA is identical in sequence (with U replacing T) to one of the DNA strands, and complementary to the other strand. If the orientation of the cloned gene is inverted relative to the promoter, will the same protein still be expressed? Why or why not?

- 13.** An investigator has two systems available for the cloning and expression of proteins: a bacterial plasmid designed for protein expression, and a baculovirus system. Which system would be her best choice to successfully clone and express (a) the gene encoding the *E. coli* RecA protein, and (b) the gene encoding a mammalian DNA polymerase?

14. In the protocol for Western blots, an investigator uses two antibodies. The first binds specifically to the protein of interest. The second is labeled for easy visualization, and it binds to the first antibody. In principle, molecular biologists could simply label the first antibody and skip one step. Why do they use two successive antibodies?

15. A group of overlapping clones, designated A through F, is isolated from one region of a chromosome. Each of the clones is separately cleaved by a restriction enzyme and the pieces resolved by agarose gel electrophoresis, with the results shown below. Nine different restriction fragments can be produced from this chromosomal region, with a subset appearing in each clone. Using this information, deduce the order of the restriction fragments in the chromosome.



- 16.** You are a researcher who has just discovered a new protein in a fungus. To help determine the protein's function, you want to identify the other proteins in the fungal cell with which your protein interacts. How do you design a yeast two-hybrid experiment to address this problem?

- 17.** Figure 7-27 shows the first steps in the process of making a DNA microarray by photolithography. Describe the remaining steps for obtaining the desired sequences (a different four-nucleotide sequence on each of the four

spots) shown in the first panel of the figure. After each step, give the resulting nucleotide sequence attached at each spot (numbered 1 to 4 in Figure 7-27).

Data Analysis Problem

Cohen, S.N., A.C.Y. Chang, H.W. Boyer, and R.B. Helling. 1973. Construction of biologically functional bacterial plasmids in vitro. *Proc. Natl. Acad. Sci. USA* 70:3240–3244.

- 18.** The first recombinant DNA plasmid that combined two different biological activities from different sources and was readily transferred into bacteria was reported in 1973 by the research groups of Stanley Cohen and Herbert Boyer. In this work, the investigators began with a large, naturally occurring circular plasmid called R6-5 (~98,500 bp), which included genes conferring resistance to multiple antibiotics, including tetracycline and kanamycin. Prior to their study, R6-5 had been sheared into fragments, the fragments used to transform *E. coli*, and tetracycline-resistant colonies selected. A smaller circular plasmid had arisen from that work, pSC101 (~9,000 bp), which replicated normally and conferred only the tetracycline resistance. The researchers assumed that the smaller plasmid represented a piece of the larger one that had ligated into a smaller circle, which included a replication origin and the tetracycline-resistance gene. A third plasmid had also been generated when plasmid R6-5 was cleaved with EcoRI, the fragments used to transform *E. coli*, and cells selected for growth on kanamycin. This third plasmid was called pSC102 (~27,000 bp), and it did not confer resistance to tetracycline.

The ultimate objective was to combine pSC101 with fragments from R6-5 or pSC102 to generate a new plasmid that conferred a demonstrable new biological activity. The investigators first cleaved all three plasmids with EcoRI. The cleavage products were subjected to electrophoresis on an agarose gel, and the DNA fragments were visualized by staining with ethidium bromide (a fluorescent molecule that binds to DNA by intercalating between adjacent base pairs). This generated the pattern in Figure 1. Electrophoresis proceeded left to right, so the bands decrease in size from left to right (lane a is pSC102; lane b, R6-5; lane c, pSC101).

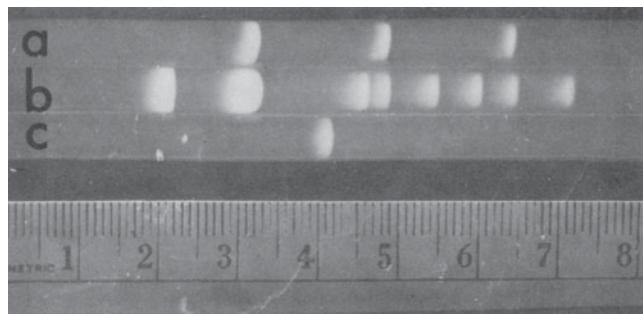


FIGURE 1

- (a) From this pattern, how many EcoRI sites are present in each of the three plasmids? (Note that three small DNA bands from R6-5 are not visible on this gel.)
 (b) Why does the brightness of the bands decrease from left to right?
 (c) In the R6-5 lane (lane b), the second largest band seems to break the general pattern in that it is brighter than the band to its left. How might this be explained?
 (d) The three bands in the pSC102 lane (lane a) comigrate with bands in the R6-5 lane. Explain.

The plasmids pSC101 and pSC102 were cleaved with EcoRI, and the fragments from both plasmids were mixed together and treated with DNA ligase. The ligated mixture was then used to transform *E. coli*, and the cells were grown on plates containing both kanamycin and tetracycline. Colonies appeared, and all of them contained a new plasmid, named pSC105 (~16,000 bp), that conferred resistance to kanamycin and tetracycline. Next, the plasmids pSC101, pSC102, and pSC105 were cleaved with EcoRI and subjected to gel electrophoresis. The results are shown in Figure 2. Electrophoresis proceeded, and fragment sizes decrease, from left to right (lane a is pSC105; lane b, pSC101 + pSC102; lane c, pSC102; lane d, pSC101).

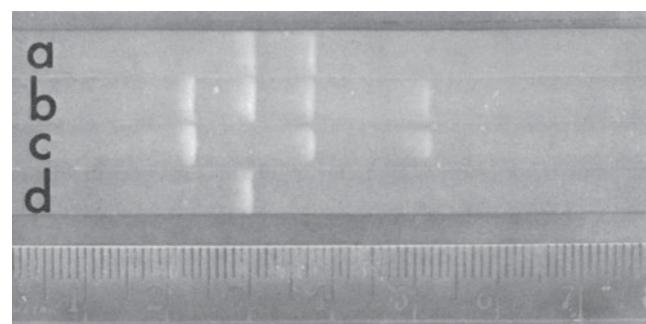


FIGURE 2

- (e) Which of the two fragments of pSC105 (lane a) contains the gene for tetracycline resistance?
 (f) Which of the two fragments of pSC105 contains the gene encoding kanamycin resistance?
 (g) What is the approximate size of the second (smaller) EcoRI cleavage fragment of pSC105?
 (h) How many phosphodiester bonds were created by DNA ligase to produce the circular plasmid pSC105?
 (i) The largest and smallest EcoRI fragments of pSC101 (lane d) do not appear in pSC105, although they were present in the ligation mixture that gave rise to it. Why were these DNA fragments not incorporated into the recombinant plasmid?

Additional Reading

General

- Cohen, S.N., A.C.Y. Chang, H.W. Boyer, and R.B. Helling.** **1973.** Construction of biologically functional bacterial plasmids in vitro. *Proc. Natl. Acad. Sci. USA* 70:3240–3244. The paper that gave rise to biotechnology.
- Jackson, D.A., R.H. Symons, and P. Berg.** **1972.** Biochemical method for inserting new genetic information into DNA of simian virus 40: Circular SV40 DNA molecules containing lambda phage genes and the galactose operon of *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* 69:2904–2909. A report of the first recombinant DNA experiment linking DNA from two species.
- Lobban, P.E., and A.D. Kaiser.** **1973.** Enzymatic end-to-end joining of DNA molecules. *J. Mol. Biol.* 78:453–471. A report of the first recombinant DNA experiment.
- Mullis, K.B.** **1990.** The unusual origin of the polymerase chain reaction. *Sci. Am.* 262(4):36–43. A description of that fateful night.
- Sambrook, J., E.F. Fritsch, and T. Maniatis.** **1989.** *Molecular Cloning: A Laboratory Manual*, 2nd ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press. Although supplanted by more recent manuals, this three-volume set includes much useful background information on the biological, chemical, and physical principles underlying both classical and current techniques.

Isolating Genes for Study (Cloning)

- Terpe, K.** **2006.** Overview of bacterial expression systems for heterologous protein production: From molecular and biochemical fundamentals to commercial systems. *Appl. Microbiol. Biotechnol.* 72:211–222.

Working with Genes and Their Products

- Arnheim, N., and H. Erlich.** **1992.** Polymerase chain reaction strategy. *Annu. Rev. Biochem.* 61:131–156.
- Brock, T.D.** **1997.** The value of basic research: Discovery of *Thermus aquaticus* and other extreme thermophiles. *Genetics* 146:1207–1210. An essay on how basic research can have unexpected benefits.
- Foster, E.A., M.A. Jobling, P.G. Taylor, P. Donnelly, P. de Knijff, R. Mieremet, T. Zerjal, and C. Tyler-Smith.** **1999.** The Thomas Jefferson paternity case. *Nature* 397:32.

The last article in a series about an interesting case study in the use of biotechnology to address historical questions.

- Giepmans, B.N.G., S.R. Adams, M.H. Ellisman, and R.Y. Tsien.** **2006.** The fluorescent toolbox for assessing protein location and function. *Science* 312:217–224.
- Hofreiter, M., D. Serre, H.N. Poinar, M. Kuch, and S. Paabo.** **2001.** Ancient DNA. *Nat. Rev. Genet.* 2:353–359. Successes and pitfalls in the retrieval of DNA from very old samples.
- Ivanov, P.L., M.J. Wadhams, R.K. Roby, M.M. Holland, V.W. Weedin, and T.J. Parsons.** **1996.** Mitochondrial DNA sequence heteroplasmy in the Grand Duke of Russia Georgij Romanov establishes the authenticity of the remains of Tsar Nicholas II. *Nat. Genet.* 12:417–420.
- Lindahl, T.** **1997.** Facts and artifacts of ancient DNA. *Cell* 90:1–3. A good description of how nucleic acid chemistry affects the retrieval of DNA in archaeology.
- MacLean, D., J.D.G. Jones, and D.J. Studholme.** **2009.** Application of “next-generation” sequencing technologies to microbial genetics. *Nat. Rev. Microbiol.* 7:287–296.
- Metzker, M.L.** **2005.** Emerging technologies in DNA sequencing. *Genome Res.* 15:1767–1776.
- Saiki, R.K., S. Scharf, F. Faloona, K.B. Mullis, G.T. Horn, H.A. Erlich, and N. Arnheim.** **1985.** Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle-cell anemia. *Science* 230:1350–1354. The first report of the polymerase chain reaction.
- Shendure, J., R.D. Mitra, C. Varma, and G.M. Church.** **2004.** Advanced sequencing technologies: Methods and goals. *Nat. Rev. Genet.* 5:335–344.

Understanding the Functions of Genes and Their Products

- Budowle, B., M.D. Johnson, C.M. Fraser, T.J. Leighton, R.S. Murch, and R. Chakraborty.** **2005.** Genetic analysis and attribution of microbial forensics evidence. *Crit. Rev. Microbiol.* 31:233–254. A description of how biotechnology is used to fight bioterrorism.
- Stoughton, R.B.** **2005.** Applications of DNA microarrays in biology. *Annu. Rev. Biochem.* 74:53–82.

Genomes, Transcriptomes, and Proteomes



Joe DeRisi [Source: Courtesy of University of California, San Francisco.]

served DNA sequences from all known types of viruses. I recall watching the spots lighting up on the scanner, and we could see that this new virus had sequences similar to coronaviruses from birds, cows, humans—it was some kind of strange new type of coronavirus that had never been seen before!

The second exciting moment came over the next couple of days as we began working nonstop to clone and sequence bits of the virus genome off of the ViroChip array. Sequences were determined by a standard method in which the cloned bits were copied by a polymerase using fluorescently labeled nucleotides. I sat by the sequencer watching as each fluorescent nucleotide was detected, and I wrote down each base by hand because I couldn't wait for the run to finish to see what the sequences were! This level of intensity doesn't happen every day—we were not stopping to sleep or rest, surviving off pizza and Skittles, and it was exhausting—but it was so incredibly exciting to be the very first person to see the actual genome sequence of this new infectious agent! That was so fun. And in the end, we had to obtain enough DNA sequence to predict the encoded protein sequences which positively identified this virus as a new and highly divergent coronavirus.

—**Joe DeRisi**, on his discovery of the SARS virus

8.1 Genomes and Genomics 260

8.2 Transcriptomes and Proteomes 277

8.3 Our Genetic History 282

As an old proverb states, it can be hard to see the forest for the trees. With the birth of biotechnology in the 1970s, the focus was on genes—their form, function, transcription, translation, and applications in medicine and agriculture. The new technologies also facilitated countless advances in our understanding of the RNA and protein products of genes. In truth, though, each gene is like the proverbial tree: a small part of a much larger forest. Understanding individual cellular components requires an examination of their function in the cellular context. At the same time, a host of new molecular questions arises that can be addressed only at the level of an organelle, cell, or organism. How does a cell or organism respond to a change in its environment, and how is the response regulated? How are the replication and transcription of genetic material coordinated with cell division? How can multiple genes be coordinately regulated? How many regulatory, replication, and repair systems for DNA exist in cells, and how does an organism's lifestyle shape its evolution? Over the past few decades, molecular biology has witnessed a constant expansion of its reach. No longer limited to an examination of one or a few genes or gene products at a time, scientists are addressing problems dealing with increasingly complex and interconnected systems.

The shift was made possible by the continuing advances in technology. The sequencing of individual genes has been supplanted by sequencing projects that encompass all of an organism's DNA—its genome. Examining the changes in the sum of a cell's RNA and proteins—its transcriptome and proteome—became a practical pursuit. The resulting efforts have spawned the new subdisciplines of genomics, transcriptomics, and proteomics, which now allow us to investigate questions on a cellular, organismal, or population scale. Huge public databases have been established, packed with information about genetic material and gene products, for species almost too numerous to count. Ongoing initiatives range from the customization of medical treatments based on an individual's genetic makeup to the detailed tracing of human evolution. The subdisciplines and databases we introduce in this chapter are an important legacy of modern molecular biology. They are also a key to its future. Biology in the twenty-first century will move forward with the aid of informational resources undreamed of just a few years ago.

8.1 Genomes and Genomics

The word “genome,” coined by German botanist Hans Winkler in 1920, was derived simply by combining *gene* and the final syllable of *chromosome*. An organism's **genome** is defined as the complete haploid genetic

complement of a typical cell. In essence, a genome is one copy of the hereditary information required to specify the organism. For sexually reproducing organisms, the genome includes one set of autosomes and one of each type of sex chromosome. When cells have organelles that also contain DNA, the genetic content of the organelles is not considered part of the nuclear genome. Mitochondria are found in most eukaryotic cells, and chloroplasts occur in the light-harvesting cells of photosynthetic organisms. Each of these organelles has its own distinct genome. In viruses, which can have genetic material composed of either DNA or RNA, the genome is a complete copy of the nucleic acid required to specify the virus.

In diploid organisms, sequence variations exist between the two copies of each chromosome present in a cell. Subtle as they are, these differences are used to solve crimes, define parentage, and help trace the path of an inherited disease through generations of an affected family. As sequencing methods become more sophisticated, sequences of the complete genetic complement of a diploid cell—diploid genomes—have begun to appear.

The study of complete genomes was rudimentary until the advent of genome sequencing projects in the 1980s. In 1986, Thomas H. Roderick of the Jackson Laboratories in Bar Harbor, Maine, came up with *Genomics* as the name for a new journal, and the word ended up defining a new field. The modern science of **genomics** is dedicated to the study of DNA on a cellular scale. Advances in this field have been propelled by improvements in sequencing technology (see Chapter 7), in computer technologies, and in innovative approaches to the organization and searching of stored information.

Many Genomes Have Been Sequenced in Their Entirety

The genome is the ultimate source of information about an organism. Less than 10 years after the development of practical DNA sequencing methods, serious discussions began about the prospects for sequencing the entire 3×10^9 base pairs (bp) of the human genome. Although the effort was inspired by our natural curiosity about ourselves, it has become much more than a human genome project. Genome sequencing in the twenty-first century is becoming routine. The number of genomes sequenced in their entirety is now in the thousands and includes organisms ranging from bacteria to mammals.

The International Human Genome Project got under way with substantial funding in the late 1980s.

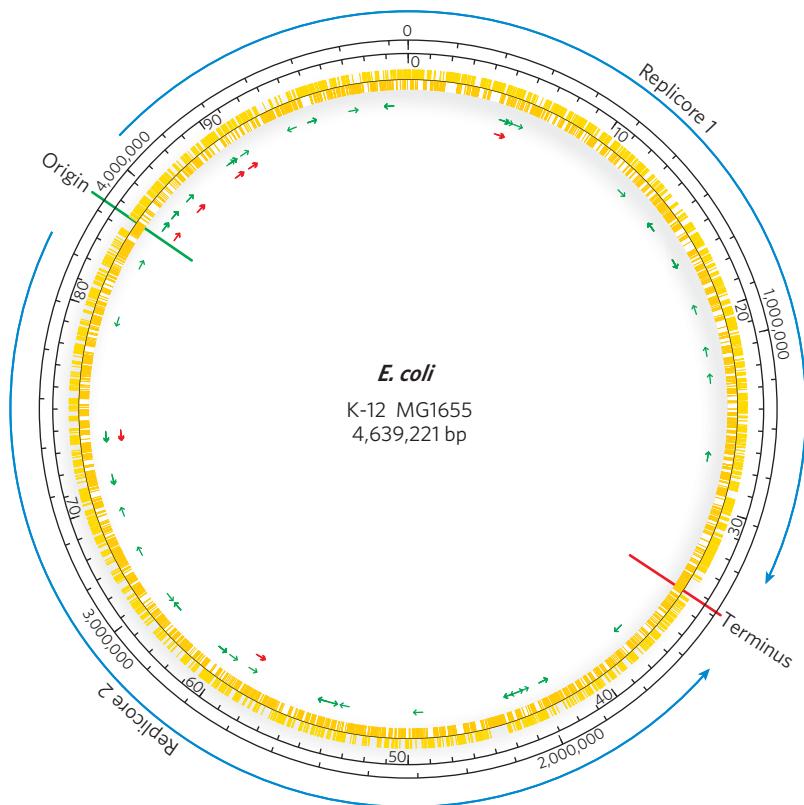


FIGURE 8-1 A snapshot of the *Escherichia coli* genome. K-12 MG1655 is the most commonly used laboratory strain of *E. coli*. The origin and terminus of replication are shown as green and red lines, respectively. Replication proceeds in two directions from the origin, dividing the genome into two regions, or replicores, that are replicated separately. The two outer black circles provide genome reference points in terms of number of base pairs (outer) and minutes (inner). The use of 100 minutes as a genomic yardstick in this chromosome is derived from the approximately 100 minutes it takes to transfer the *E. coli* chromosome from donor cell to recipient during bacterial conjugation. Yellow and orange markings on the inner circle denote protein-coding genes. Green and red arrows are locations of genes for tRNA and rRNA, respectively. [Source: Adapted from F. R. Blattner et al., *Science* 277:1453–1462, 1997.]

Several additional and closely linked projects focused on organisms other than humans. The first complete genome to be sequenced was that of the bacterium *Haemophilus influenzae*, in 1995 (see How We Know). The first eukaryotic genome sequence, that of the yeast *Saccharomyces cerevisiae*, followed in 1996. The genome sequence for the bacterium *Escherichia coli* became available in 1997 (Figure 8-1). The much larger effort directed at the human genome was also accelerating.

The Human Genome Project eventually included significant contributions from 20 sequencing centers distributed among six nations: the United States, Great Britain, Japan, France, China, and Germany. Some general coordination was provided by the Office of Genome Research at the National Institutes of Health in the United States, led first by James Watson and after 1992 by Francis Collins. However, much of the effort relied on the informal, and very successful, international collaborations. At the outset, the task of sequencing a 3×10^9 bp genome seemed to be a titanic job, but it gradually yielded to continued technological innovation. The draft sequence of the human genome was published in February 2001, and the project was completed in April 2003, several years ahead of schedule. The published sequence is actually a composite, derived from several anonymous donors. Although the

DNA of several individuals is represented, the sequence in any given genomic region is generally from one individual.

This advance was the product of a carefully planned effort spanning 13 years. Research teams used restriction enzymes to partially digest the entire human genome, and then cloned suitably long segments into BAC and YAC vectors (see Chapter 7). Overlapping clones in the resulting libraries were identified by hybridization and other methods and organized into long contiguous stretches of chromosomal DNA called **contigs**. Each contig included at least one and usually many identifying sequences that had already been



Francis Collins [Source: Alex Wong/Newsmakers.]

mapped to a particular region of a particular chromosome. These regions were either a unique and previously characterized sequence called a **sequence tagged site (STS)** or a gene whose expression could be monitored, known as an **expressed sequence tag (EST)**. The STS and EST landmarks in the contigs were often sequences that had been roughly mapped along a specific chromosome. These

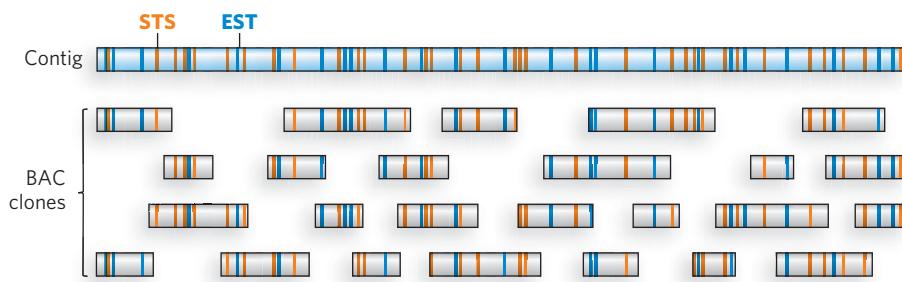


FIGURE 8-2 Mapping BACs into a contig. A contig of a chromosomal region is constructed by identifying the sequence tagged site (STS; orange) and expressed sequence tag (EST; blue) markers in each BAC and mapping them to the contig.

landmark-containing contigs could thus be ordered along each chromosome, gradually defining a physical map of the genome (Figure 8-2).

The contigs were then divided up between the international sequencing centers, and each center began sequencing the mapped BAC or YAC clones corresponding to its assigned segments of the genome. Because many of the clones were more than 100,000 bp long, and contemporary sequencing techniques resolved only 600 to 750 bp of sequence at a time, each clone had to be sequenced in pieces. The strategy was a shotgun approach, in which researchers used powerful new automated sequencers to sequence random segments of a given BAC or YAC clone, then assembled the entire clone by computerized identification of overlaps. Each clone was sequenced at least four to six times to ensure accuracy. The data were made available in the growing genome database.

Construction of the physical map was a time-consuming task, and its progress was followed in annual reports in major journals throughout the 1990s—by the end of which the map was largely in place. Completion of the entire sequencing project was initially projected for 2005, but circumstances and technology intervened to accelerate the process.

A competing commercial effort to sequence the

human genome had been initiated by the newly established Celera Corporation in 1997. Led by J. Craig Venter, the Celera group made use of a different strategy, called **whole-genome shotgun sequencing**, which eliminates the step of assembling a physical map of the genome. Instead, teams sequenced DNA segments from throughout the genome, at random. The sequenced segments were

ordered by the computerized identification of sequence overlaps (with some reference to the public project's detailed and published physical map). Like the public genome project, the Celera effort used DNA from several human donors. About 70% of the sequence comes from one male donor—Craig Venter himself.

At the outset of the Human Genome Project in the 1980s, shotgun sequencing on this scale had been deemed impractical because the sequence assembly was too computationally complex. However, by 1997, advances in computer software and sequencing automation had made the approach feasible. The ensuing race between private and public efforts substantially shortened the timeline for completing the project (Figure 8-3). Publication of the draft human genome sequence in 2001 was followed by two years of follow-up work to eliminate nearly a thousand discontinuities and to provide high-quality sequence data, contiguous throughout the genome.

The human genome is only part of the genome sequencing story, and an increasingly small part. The genomes of many other species have been sequenced in the continuing effort, gradually providing a unique look at genomic complexity throughout the three domains of living organisms: Bacteria, Archaea, and Eukarya (eukaryotes) (see Figure 8-3). Whereas many early sequencing projects focused on species commonly used in research laboratories, the programs have expanded to include a wide range of species of practical, medical, agricultural, and evolutionary interest. As the second decade of the new century begins, the number of completed genomes is almost uncountable. Completed bacterial genomes number in the thousands and include at least one species from virtually every known bacterial family. Completed eukaryotic genomes number in the hundreds. Each completed genome becomes a resource for scientists around the world who use that organism in their research, facilitating the identification of important genes. Multiple individual human genomes have been sequenced, and genome-based,



J. Craig Venter [Source: Mike Theiler/Reuters.]

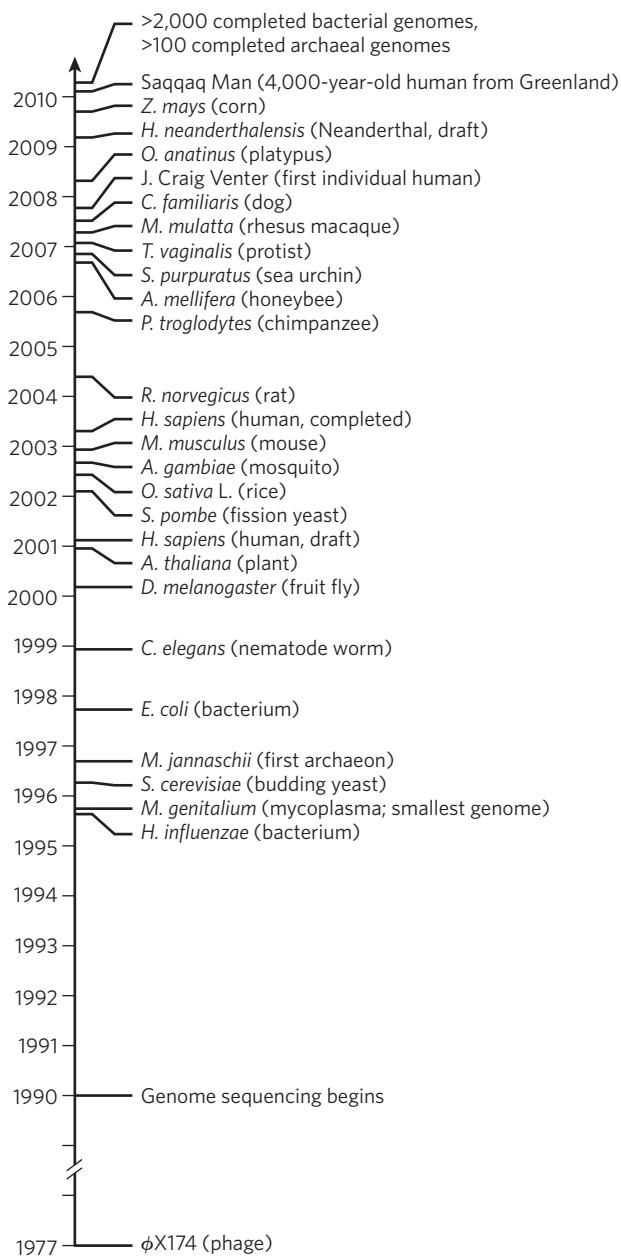


FIGURE 8-3 The genome sequencing timeline. Preparatory work for the Human Genome Project, including extensive mapping to provide genome landmarks, occupied much of the 1990s. The rapid development of sequencing methods and strategies initiated a broad range of additional sequencing efforts. As the draft human genome was announced in 2001, hundreds of other projects were already under way. Many genome projects involving species that are widely used in research have their own websites, which serve as central repositories for the most recent data.

personalized medicine is on the horizon. The sequencing efforts have even expanded to include extinct species such as *Homo neanderthalensis* (Highlight 8-1), as well as humans who died in past millennia. The many

genome sequences provide a source for broad comparisons that help pinpoint both variable and highly conserved gene segments, and allow the identification of genes that are unique to a species or group of species. Efforts to map genes, identify new proteins and disease genes, elucidate genetic patterns of medical interest, and trace our evolutionary history, as well as many other initiatives, are under way.

Annotation Provides a Description of the Genome

A genome sequence is simply a very long string of A, G, T, and C residues. The value of this sequence information depends almost entirely on the manner in which the information is organized when it is stored. The critical process of **genome annotation** yields a listing of information about the location and function of genes and other critical sequences. Genome annotation converts the sequence itself to information that any researcher can use. Much of the effort focuses on genes encoding RNA and protein, because such genes are most often the target of scientific investigations. Every newly sequenced genome includes many genes—often 40% or more of the total—about which little or nothing is known. Here, the annotation exercise is most challenging.

Protein and RNA function can be described on three levels. **Phenotypic function** describes the effects of a gene product on the entire organism. For example, the loss or mutation of a particular protein may lead to slower growth, altered development, or even death. **Cellular function** is a description of the metabolic processes in which a gene product participates and of the interactions of that gene product with other proteins or RNAs in the cell. **Molecular function** refers to the precise biochemical activity of a protein or an RNA, such as the reactions an enzyme catalyzes, the ligands a receptor binds, or the complex formed between a specific RNA and a protein. Each of these functions can be elucidated by computational and experimental approaches. Some of these are described here, and additional techniques are presented in Section 8.2.

Computational approaches involve Web-based programs that are used to define gene locations and assign tentative gene functions (where possible), based on similarity to genes previously studied in other genomes. (Most of these programs are freely available on the Internet.) For investigating the function of a particular gene, resources such as the classic BLAST (Basic Local Alignment Search Tool) algorithm allow a rapid search of all genome databases for sequences related to one that a researcher has just generated. Two other

HIGHLIGHT 8-1 EVOLUTION

Getting to Know the Neanderthals

Modern humans and Neanderthals coexisted in Europe and Asia as recently as 30,000 years ago. The human and Neanderthal ancestral populations permanently diverged about 370,000 years ago, before the appearance of anatomically modern humans. Neanderthals used tools, lived in small groups, and buried their dead. Of the known hominid relatives of modern humans, Neanderthals are the closest. For hundreds of millennia, they inhabited large parts of Europe and western Asia (Figure 1). If the chimpanzee genome can tell us something about what it is to be human, perhaps the Neanderthal genome can tell us more. Buried in the bones and remains taken from burial sites are fragments of Neanderthal genomic DNA. Technologies developed for use in forensic science (see Highlight 7-1) and ancient DNA studies have

been combined to initiate a Neanderthal genome project.

This endeavor is unlike the genome projects aimed at extant species. The Neanderthal DNA is present in small amounts, and it is contaminated with DNA from other animals and bacteria. How does the researcher get at it, and how can we be certain that the sequences really came from Neanderthals? The answers have been revealed by a new application of metagenomics (see Highlight 8-2). In essence, the small quantities of DNA fragments found in a Neanderthal bone or other remains are cloned into a library, and the cloned DNA segments are sequenced at random, contaminants and all. The sequencing results are compared with the existing human genome and chimpanzee genome databases. Segments derived from Neanderthal DNA are readily distinguished from segments derived from



FIGURE 1 Neanderthals occupied much of Europe and western Asia until about 30,000 years ago. Major Neanderthal archaeological sites are shown here.

(Note that this hominid group was named for the site at Neandertal in Germany.)

prominent Internet resources are the NCBI (National Center for Biotechnology Information) site, sponsored by the National Institutes of Health, and the Ensembl site, cosponsored by the EMBL-EBI (European Molecular Biology Laboratory-European Bioinformatics Institute) and the Wellcome Trust Sanger Institute.

The availability of many genome sequences in online databases enables researchers to assign gene functions by genome comparisons, an enterprise referred to as **comparative genomics**. Sequence comparison can be done with DNA, RNA, or protein. Any two genes with a demonstrable sequence similarity, whether or

bacteria or insects by computerized analysis, because they have sequences closely related to human and chimpanzee DNA. Once a collection of Neanderthal DNA segments is sequenced, they can be used as probes to identify sequence fragments in ancient samples that overlap with these known fragments. The potential problem of contamination with the closely related modern human DNA can be controlled for by examining mitochondrial DNA. Human populations have readily identifiable haplotypes (distinctive sets of genomic differences; see text for details) in their mitochondrial DNA, and analysis of Neanderthal samples has shown that their mitochondrial DNA has its own distinct haplotypes. The presence of some base-pair differences in the chimpanzee database but not in the human database is more evidence that nonhuman hominid sequences are being found.

As challenging as the effort is, completion of this endeavor is on the horizon. The draft sequence for the Neanderthal genome was unveiled in early 2009, covering more than 60% of the genomic sequences. A finished sequence will require just a little more time. The data provide evidence that modern humans and the Neanderthals who were the source of this DNA shared a common ancestor about 700,000 years ago (**Figure 2**). Analysis of mitochondrial DNA suggests that the two groups continued on the same track, with some gene flow between them, for about 300,000 more years. The lines split for good long before the appearance of anatomically modern humans.

Expanded libraries of Neanderthal DNA from different sets of remains should eventually allow an analysis of Neanderthal genetic diversity, and perhaps Neanderthal migrations. This look at the hominid past promises to be fascinating.

not they are closely related by function, are called **homologs**. The sequence similarity (homology) implies an evolutionary relationship. Quite often, sequence similarity and a functional relationship go hand in hand. When two genes in different species possess a clear sequence and functional relationship to each

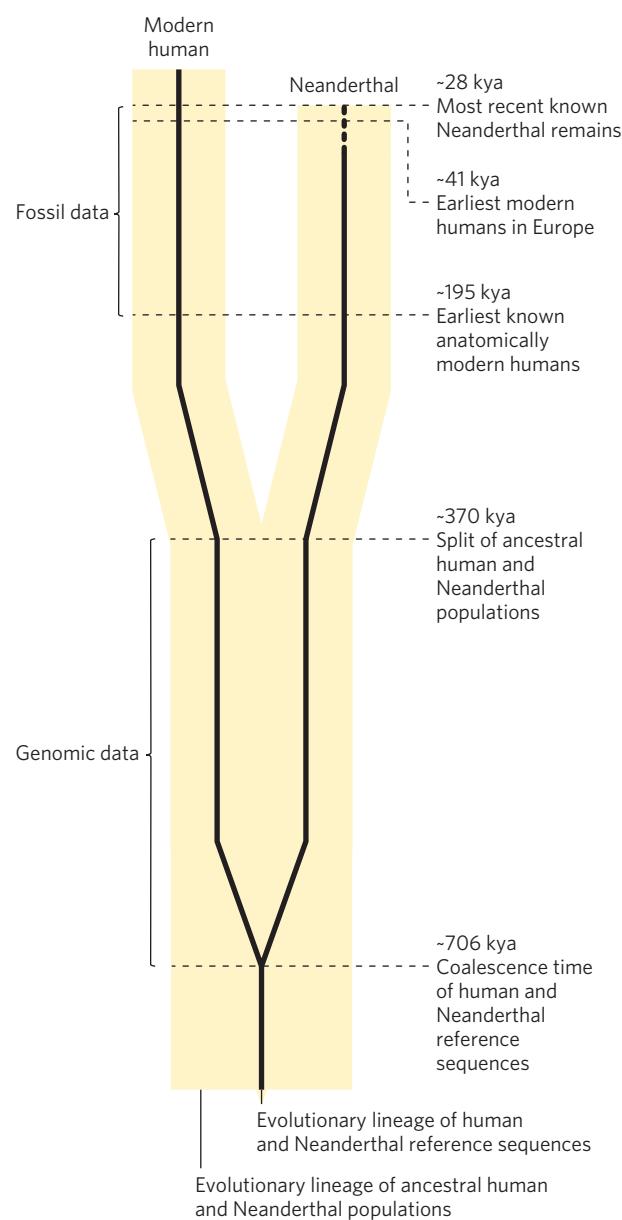


FIGURE 2 This timeline shows the divergence of human and Neanderthal genome sequences (black lines), and of ancestral human and Neanderthal populations (yellow screen). Key events in human evolution are noted (kya indicates thousand years ago. [Source: Adapted from J. P. Noonan et al., *Science* 314:1113–1118, 2006.]

other, they are known as **orthologs**—genes derived from an ancestral gene in the last common ancestor of these two species. **Paralogs** are genes that are similarly related to each other but within a single species; they arise most often from gene duplication in a single genome, followed by specialization of one or both copies

of the gene over the course of evolution. If the function of any gene has been characterized for one species, this information can be used to at least tentatively assign gene function to a related gene in a second species.

Gene identity is often easiest to discern when comparing genomes from closely related species, such as mouse and human, although many clearly orthologous genes have been identified in species as distant as bacteria and humans. In many cases, even the order of genes on a chromosome is conserved over large segments of the genomes of closely related species. Conserved gene order, or **synteny**, provides additional evidence for an orthologous relationship between genes at identical locations in the related segments (Figure 8-4). The distinction between orthologs and paralogs was introduced by Walter Fitch in 1970, and its importance was established with the advent of genome sequencing projects in the 1990s. As the number of known genome sequences increases, many genes and genomic segments can be productively annotated by using automated tools available on the NCBI and Ensembl websites.

In every newly described genome sequence, the many genomic segments and genes that have never been characterized—that unknown 40% or so of the total—represent a special challenge. Elucidation of gene function in these cases will probably take many decades. Some experimental approaches exist, and new ones are being developed. Many of the current approaches again focus on protein-coding genes. For several genomes, such as those of *S. cerevisiae* and the plant *Arabidopsis thaliana*, gene knockout (inactivation) collections have been developed by genetic engineering. Each strain in

an organism's collection has a different inactivated gene, and every genomic gene is represented. If the growth patterns or other properties of the organism change when the gene is inactivated, this provides information on the phenotypic function of the protein product of the gene. In other available libraries, each gene in a specific genome is expressed as a tagged fusion protein (see Chapter 7). The tags may be designed to allow protein isolation, investigate interactions with other proteins, or explore subcellular localization. Some approaches for using tags to determine the function of genes are described in more detail in Section 8.2.

Genome Databases Provide Information about Every Type of Organism

The available genome sequences are assisting research in all biological disciplines. Increasingly, they are inspiring molecular biologists to ask questions that, until now, could not be answered. Just a brief overview will illustrate the utility of expanding the Human Genome Project to essentially all species.

Viruses Viruses are not free-living organisms but obligate intracellular parasites; thus, every virus is a pathogen of some organism. The viruses that are human pathogens—such as SARS (see Moment of Discovery)—are of special interest. However, viruses that infect farm animals, food crops, landscape plants, and many other organisms can be economically important. Bacteria also serve as hosts to viruses, which are generally termed bacteriophages. Viruses fall into seven classes, depending on whether the genomic nucleic acid is RNA or DNA, whether it is single-stranded or double-stranded, and the mechanisms employed to replicate it (Table 8-1). Viruses vary a great deal in genomic complexity, ranging from a mere 2,000 nucleotides (found in a few single-stranded DNA viruses that infect vertebrates) to around 1.2 million bp (in a double-stranded DNA virus that infects amoebas). Thousands of viral genomes have been or are being sequenced, in an effort that will greatly aid future progress in medicine and agriculture.

Bacteria Bacteria inhabit every environment—from polar ice to deserts, from ocean depths to kitchen counters and the soil in your backyard. Some are pathogens. Others help digest our food, convert atmospheric nitrogen to forms that all organisms can use, convert carbon dioxide to oxygen, and carry out myriad other tasks without which all other life forms would perish. With that in mind, molecular biologists are subjecting thousands of representative bacterial species to genome sequencing.

In the past few decades, researchers have realized that a vast number of bacterial species remained

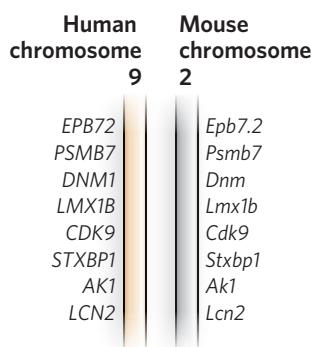


FIGURE 8-4 Synteny in the human and mouse genomes.

Large segments of the two genomes have closely related genes aligned in the same order on the chromosomes. In these short segments of human chromosome 9 and mouse chromosome 2, the genes exhibit a very high degree of homology, as well as the same gene order. The different lettering schemes for the gene names simply reflect the different naming conventions in the two species.

Table 8-1 The Seven Classes of Viruses

Class	Genome	Replication	Examples	Genome Size (kbp or kb)
I	Double-stranded DNA	Nuclear or cytoplasmic; host DNA polymerase	Polyomaviruses Adenoviruses Baculoviruses Papovaviruses Poxviruses 1* P22* T5*	5-1,200
II	Single-stranded DNA	Nuclear in eukaryotic host; host cell DNA polymerases	Circoviruses Geminiviruses Parvoviruses Inoviruses* Microviruses*	2-9
III	Double-stranded DNA	Cytoplasmic; virus-encoded replicases	Birnaviruses Chrysoviruses Cystoviruses Hypoviruses Partitiviruses Reoviruses Totiviruses	3-32
IV	Single-stranded, positive-sense RNA [†]	Virus-encoded replicases	Bromoviruses Coronaviruses Picornaviruses	2-31
V	Single-stranded, negative-sense RNA	Virus-encoded replicases	Arenaviruses Bunyaviruses Bornaviruses Rhabdoviruses Paramyxoviruses	9-19
VI	Single-stranded RNA reverse-transcribing	Virus-encoded reverse transcriptase	Retroviruses (many types)	4-12
VII	Double-stranded DNA	DNA genome generated by reverse transcription of RNA intermediates in viral particle during maturation	Caulimoviruses Hepadnaviruses	3-9

*Bacterial virus.

[†]Most abundant viral genome type. “Positive sense” means that the sequence is identical to the sequence of mRNAs encoded by the virus. (A negative-sense RNA is one with a sequence complementary to the mRNAs encoded by the virus.)

uncharacterized. Many bacteria live in interdependent microbial communities and cannot be cultured in pure form in the laboratory. Examples are found in the human intestine, in the termite gut, and in the effluent of deep-sea steam vents. Many of these bacteria are important to human health, both directly and indirectly. Others are of economic importance. For instance, an understanding of the microbial processes that allow

termites to digest cellulose in their gut could provide new ways to convert grass and other cellulose materials to usable fuels.

The need to know more about these microbial communities has given rise to a subdiscipline of genomics, **metagenomics**. In metagenomics projects, DNA is isolated not from a single bacterial species but from an entire community of microbial species (Highlight 8-2).

HIGHLIGHT 8-2 TECHNOLOGY

Sampling Biodiversity with Metagenomics

Most of the biological diversity on our planet is found in microorganisms. However, we know surprisingly little about Earth's microbial diversity. Much of the research to date has focused on bacteria and viruses that are of medical, agricultural, or commercial interest, and almost all of it on species that can be isolated and cultured in a laboratory. A wealth of microbial diversity remains to be discovered in the world's swamps, deserts, and oceans, involving species that cannot yet be cultured. Assessing the diversity in communities of microorganisms is one goal of the new discipline of metagenomics.

The sampling involves shotgun DNA sequencing on a truly grand scale. Individual species are not isolated. Instead, an entire microbial population is taken from a given environment, and DNA sequences from that population are analyzed at random. Early approaches have looked at a biofilm in an acid mine, soil in Minnesota, water samples from the Sargasso Sea, whale falls (whales that have died and sunk to the sea bottom), and human feces. DNA from the bacteria and/or viruses in the sample is broken into fragments and sequenced at random. Computerized analyses identify any overlapping sequences and link them into longer contigs. These genomic snippets are assembled in a database. The diversity can be measured by focusing on specific genes. For example, the 16S rRNA gene is universal in bacteria and is often used as a benchmark for defining species. When millions of individual

genome segments are sequenced, large numbers of 16S rRNA genes are generally represented in the database. A careful look at this one type of gene can indicate the variety of microbial species in the sample.

A very ambitious metagenomics initiative was started by Craig Venter and his coworkers at the J. Craig Venter Institute (Rockville, Maryland) in the spring of 2003. A 30-meter sailing sloop, the *Sorcerer II*, was converted into an oceanic research vessel. After a trial run in the Sargasso Sea, the Global Ocean Sampling (GOS) expedition was launched in March 2004. Beginning in Halifax, Canada, and circumnavigating the globe (Figure 1), the voyage continued until January 2006. Samples of ocean water were taken every 200 nautical miles. Microorganisms were strained from the water with a series of filters and sent back to the lab for extraction of DNA and sequencing.

The result is the largest database of DNA sequences derived from marine organisms yet released into the public domain. More than 7.7 million sequences are included in the GOS database, encompassing more than 6.3×10^9 bp of DNA. Within these sequences are 4,125 distinct 16S rRNA gene sequences, representing more than 800 species—perhaps half of them not previously discovered. The genes for 6.12 million bacterial proteins have been found, including nearly 4,000 separate protein families, of which 1,700 are novel. The effort has given rise to new technologies for archiving and analyzing massive sequence databases.

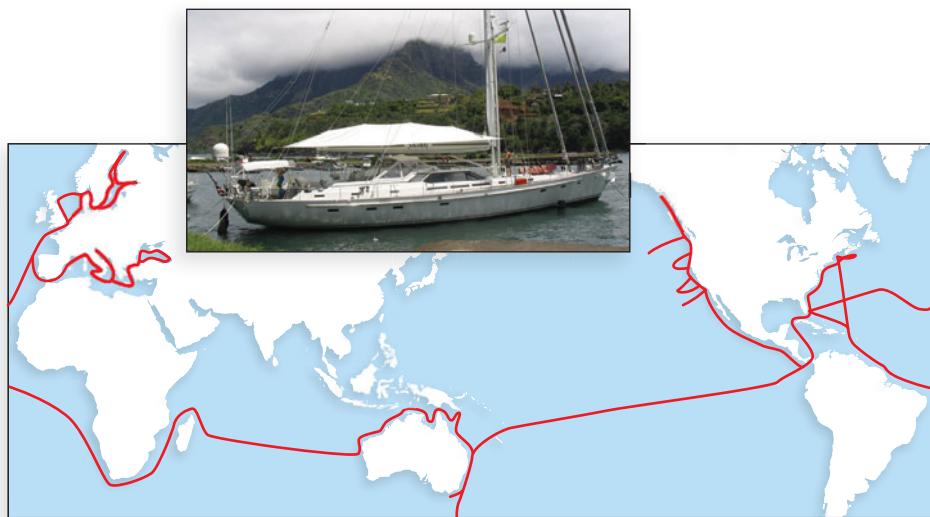


FIGURE 1 The route taken by *Sorcerer II*. [Source: Photograph courtesy of J. Craig Venter Institute.]

The DNA is sequenced by a shotgun technique, and the researcher uses computer programs to assemble overlapping segments derived from individual genomes. When the community includes only a handful of species, researchers have been able to reconstruct multiple genomes from these mixed samples. The process has some unique complexities, many related to the close evolutionary relationship of many bacterial species. This results in similarities over large stretches of genomic nucleic acid that can complicate the computerized assembly of genomes from short sequence “reads.” Assembling genomes from microbial communities with hundreds of different species will require further advances in both sequencing technologies and assembly programs.

Archaea In 1977, Carl Woese and his colleagues introduced the world to a new domain of living organisms, the Archaeabacteria, now renamed Archaea. A careful study of 16S ribosomal RNA sequences led to the discovery of this previously unsuspected group. The archaea (singular, archaeon) are single-celled organisms, look a lot like bacteria, and like bacteria are ubiquitous. However, many of the most interesting species are extremophiles, inhabiting hot springs or water with very high salinity or other unusual environments. Sharing some properties with both bacteria and eukaryotes, archaea have nevertheless evolved as an independent line. Their contributions to the chemistry of the biosphere make them important targets for study and genome sequencing.

Eukaryotes Eukaryotic genomes can be considerably larger than the genomes in the other two domains. Nevertheless, the sequencing of even very large eukaryotic genomes is becoming routine. Databases already contain complete genomes ranging from single-celled eukaryotes such as *S. cerevisiae* to nematodes to plants, insects, and mammals. Orthologs of genes involved in important processes and disease states in humans can almost always be found in the genomes of model organisms, facilitating laboratory research into gene function. Specialized databases have been developed for the genomes of organisms that are of particular interest to science, including mouse, fruit fly, mustard weed, and yeast (see Model Organisms Appendix). Other databases are being established that focus on plant and animal species critical to agriculture, such as corn, rice, and cattle. Some databases focus on specific types of genes. All of these databases are easily found on an Internet search or via links on the NCBI and Ensembl websites. Individual human genomes are also available online, including those of James Watson and Craig Venter!

The Human Genome Contains Many Types of Sequences

All of these rapidly growing databases have the potential not only to fuel advances in biology but to change the way we think about ourselves. What does our own genome, and its comparison with those of other organisms, tell us?

In some ways, we are not as complicated as we once imagined. Decades-old estimates that humans had about 100,000 genes within the approximately 3.2×10^9 bp of the human genome have been supplanted by the discovery that we have only about 25,000 protein-encoding genes—less than twice the number in a fruit fly (13,601 genes), not many more than in a nematode worm (19,735 genes), and fewer than in a rice plant (38,000 genes).

In other ways, however, we are more complex than we previously realized. The study of eukaryotic chromosome structure, and more recently the sequencing of entire eukaryotic genomes, has revealed that many, if not most, eukaryotic genes contain one or more intervening segments of DNA that do not code for the amino acid sequence of the polypeptide product. These nontranslated inserts interrupt the otherwise colinear relationship between the gene's nucleotide sequence and the amino acid sequence of the encoded polypeptide. Such nontranslated DNA segments are called **intervening sequences**, or **introns**, and the coding segments are called **exons** (Figure 8-5). Few bacterial genes contain introns. The process of removing introns from a primary RNA transcript to generate a transcript that can be translated contiguously into a protein product is known as splicing (see Chapter 16). An exon often (but not always) encodes a single domain of a larger, multi-domain protein. Humans share many protein domain types with plants, worms, and flies, but we use these domains in more complex arrangements. Alternative modes of gene expression and RNA splicing permit the production of alternative combinations of exons, leading to the production of more than one protein from a single gene. Humans and other vertebrates engage in this process far more than do bacteria, worms, or any other form of life—thereby allowing greater complexity in the proteins generated.

In mammals and some other eukaryotes, the typical gene has a much higher proportion of intron DNA than exon DNA; in most cases, the function of introns is not clear. Only about 1.5% of human DNA is “coding” or exon DNA, carrying information for protein or RNA products (Figure 8-6a). However, when the much larger introns are included in the count, as much as 30% of the human genome consists of genes. Several efforts

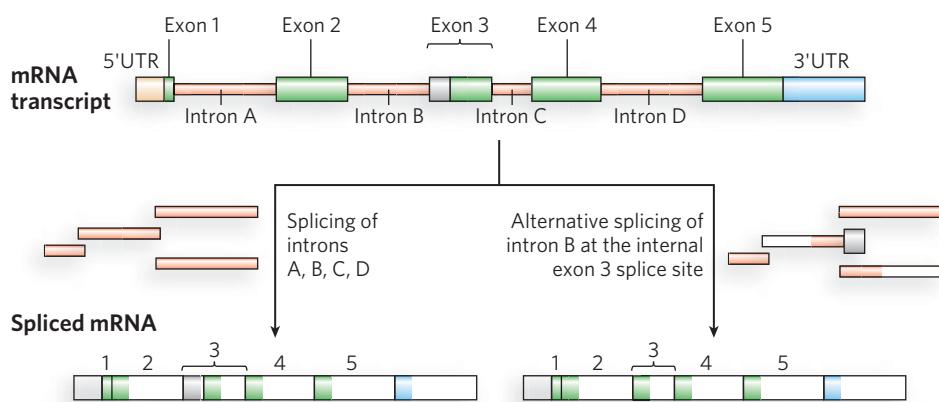


FIGURE 8-5 Introns and exons. The gene for human growth hormone 1 (*GH1*) contains five exons and four introns, along with 5' and 3' untranslated regions (5'UTR and 3'UTR). Of the several alternative patterns of splicing,

are under way to categorize the protein-encoding genes by function (Figure 8-6b).

The relative paucity of genes in the human genome leaves a lot of DNA unaccounted for. Much of the non-gene DNA is in the form of repeated sequences of several kinds. Perhaps most surprising, about half the human genome is made up of moderately repeated sequences that

two are shown here. Alternative splicing allows cells to synthesize different variants of a given protein from one gene. [Source: Adapted from J. J. Kopchick et al., *Nat. Clin. Pract. Endocrinol. Metab.* 3:355–368, 2007.]

are derived from transposable elements—segments of DNA, ranging from a few hundred to several thousand base pairs long, that can move from one location to another in the genome. Originally discovered in corn by Barbara McClintock, transposable elements (**transposons**) are a kind of molecular parasite. They efficiently make their home in the genomes of essentially every organism. Many transposons contain genes encoding the proteins that catalyze the transposition process itself, as described in more detail in Chapter 14. There are multiple classes of transposons in the human genome. Many are strictly DNA segments, increasing in number slowly with time as a result of replication events coupled to the transposition process. Some, called **retrotransposons**, are closely related to retroviruses, transposing from one genomic location to another via RNA intermediates that are reconverted to DNA by reverse transcription. Some transposons in the human genome are active, moving at a low frequency, but most are inactive, evolutionary relics altered by mutations. Although transposons generally do not encode proteins or RNAs that are used in human cells, they have played a major role in human

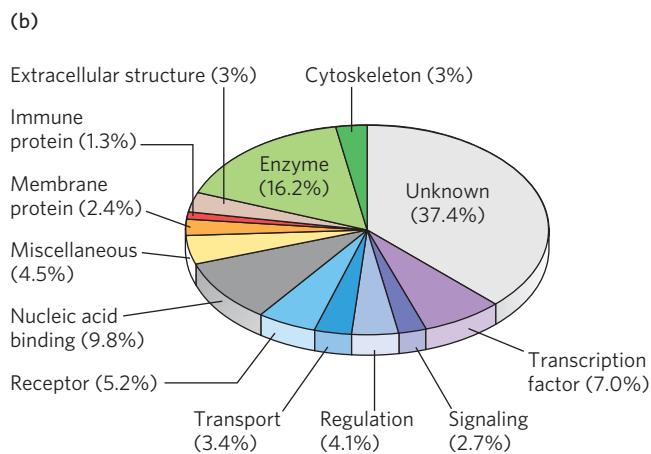
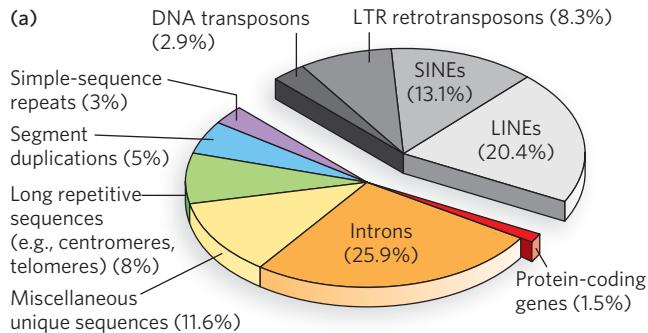


FIGURE 8-6 A snapshot of the human genome. (a) This pie chart shows the proportions of various types of sequences in our genome. The classes of transposons that represent nearly half of the total genomic DNA are indicated in shades of gray. LTR retrotransposons are retrotransposons with long terminal repeats. Long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs) are special classes of particularly common DNA transposons (see Chapter 14). (b) The approximately 25,000 protein-coding genes in the human genome can be classified by the type of protein encoded. [Source: (a) Adapted from T. R. Gregory, *Nat. Rev. Genet.* 6:699–708, 2005.]



Barbara McClintock,
1902–1992 [Source: National
Institutes of Health.]

evolution; transposon movement can lead to the redistribution of other genomic sequences.

Once the protein-coding genes (including exons and introns) and transposons are accounted for, perhaps 25% of the total DNA remains. The largest portion of this consists of unique sequences found between protein-coding genes. As described in Chapters 16 and 19 through 22, it is slowly becoming clear that virtually all of

these DNA segments are transcribed into RNA in at least some human cells. New classes of functional RNAs—encoded by genes whose existence was previously unsuspected—are being discovered at a rapid pace. Many genes encoding functional RNA are difficult to identify by automated methods, particularly when the RNA products have not been characterized. However, the RNA-coding genes are clearly a prominent feature of these otherwise uncharted genomic regions.

Another 3% or so of the human genome consists of highly repetitive sequences referred to as **simple-sequence repeats (SSRs)**. Generally less than 10 bp long, an SSR is sometimes repeated millions of times per cell and has identifiable functional importance in human cellular metabolism. The most prominent examples of SSR DNA occur in centromeres and telomeres (see Chapter 9). However, long repeats of simple sequences also occur throughout the genome.

What does all this information tell us about the similarities and differences among individual humans? Within the human population there are millions of single-base variations, called **single nucleotide polymorphisms**, or **SNPs** (pronounced “snips”). Each human differs from the next by, on average, 1 in every 1,000 bp. Many of these variations are in the form of SNPs, but a wide range of larger deletions, insertions, and small rearrangements occur in the human population as well. From these often subtle genetic differences comes the human variety we are all aware of—differences in hair color, stature, eyesight, allergies to medication, foot size, and (to some unknown degree) behavior.

The process of genetic recombination during meiosis tends to mix and match these small genetic variations so that different combinations of genes are inherited (see Chapter 13). However, groups of SNPs and other genetic differences that are close together on a chromosome are rarely affected by recombination

and are usually inherited together; these groupings are known as **haplotypes**. Haplotypes provide convenient markers for certain human populations and individuals within populations.

Defining a haplotype requires several steps. First, positions that contain SNPs in the human population are identified in genomic DNA samples from multiple individuals (Figure 8-7a). Each SNP may be separated from the next by many thousands of base pairs. Second, SNPs that are inherited together are compiled into haplotypes (Figure 8-7b). Each haplotype consists of the particular bases found at the various SNP positions in the defined haplotype. Finally, tag SNPs—a subset of the SNPs that define the entire haplotype—are chosen to uniquely identify each haplotype (Figure 8-7c). By sequencing just these tag positions in genomic samples from human populations, researchers can quickly identify which of the haplotypes are present in each individual. Especially stable haplotypes exist in the mitochondrial genome (which never undergoes meiotic recombination) and on the male Y chromosome (only 3% of which is homologous to the X chromosome and thus subject to recombination). As we'll see in Section 8.3, haplotypes can be used as markers to trace human migrations.

Genome Sequencing Informs Us about Our Humanity

A primary purpose of most genome sequencing projects is to identify conserved genetic elements of functional significance, such as conserved exon sequences, regulatory regions, and other genomic features (centromeres, telomeres, etc.). The primary purpose of sequencing the human genome is quite distinct. Here, we are interested in the differences between our genome and those of other organisms. These differences can reveal the molecular basis of human genetic diseases and can help identify genes, gene alterations, and other genomic features that are unique to the human genome and thus likely to contribute to definably human characteristics.

As the genome projects have made clear, the human genome is very closely related to other mammalian genomes over large segments of every chromosome. However, for a genome measured in billions of base pairs, differences of a few percent can add up to millions of genetic distinctions. Searching among these, and utilizing comparative genomics techniques, we can begin to explore the molecular basis of our large brain, language skills, tool-making ability, or bipedalism.

The genome sequence of our closest biological relative, the chimpanzee, offers some important clues and can illustrate the comparative process. Humans

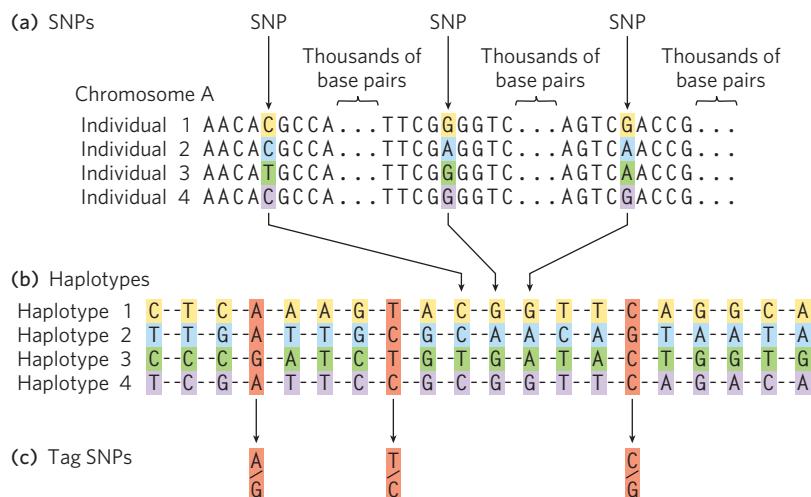


FIGURE 8-7 Haplotype identification. (a) SNPs are identified in genomic samples, and (b) groups of SNPs are compiled into a haplotype. The SNPs will vary in the overall human population, such as in the four fictional individuals shown here. However, the SNPs chosen to define a haplotype will often be the same in most individuals of a particular population. (c) Haplotype-defining SNPs (tag SNPs) can be used to simplify the process of identifying an individual's

haplotype (by sequencing 3 instead of 20 loci). If the positions shown are sequenced, an ATC sequence might be characteristic of a population native to one location in northern Europe, whereas GTC might be found in a population in Asia. Multiple haplotypes of this kind are used to trace prehistoric human migrations. See text for details. [Source: Adapted from International HapMap Consortium, *Nature* 426:789–796, 2003.]

and chimpanzees shared a common ancestor about 7 million years ago. Genomic differences between the two species fall into two types: base-pair changes (SNPs) and larger genomic rearrangements of many types. SNPs in the protein-coding regions often result in amino acid changes that can be used to construct a phylogenetic tree (Figure 8-8a), as described in Section 8.3. Segments of chromosomes may become inverted as a result of a segmental duplication, transposition of one copy to another arm of the same chromosome, and recombination between them (Figure 8-8b). Such inversions have occurred in the human lineage on chromosomes 1, 12, 15, 16, and 18. Chromosome fusions can also occur. In the human lineage, two chromosomes found in other primate lineages have been fused to form human chromosome 2 (Figure 8-8c). The human lineage thus has 23 chromosome pairs rather than the 24 pairs typical of simians. Once this fusion appeared in the line leading to humans, it would have represented a major barrier to interbreeding with other primates that lacked it.

If we ignore transposons and large chromosomal rearrangements, the published human and chimpanzee genomes differ by only 1.23% at the level of base pairs (compared with the 0.1% variance from one human to another). Some variations are at positions where there is a known polymorphism in either the

human or the chimpanzee population, and these are unlikely to reflect a species-defining evolutionary change. When we also ignore these positions, the differences amount to about 1.06%, or about 1 in 100 bp. This might seem a small number, but in large genomes it translates into more than 30 million base-pair changes, some of which affect protein function and gene regulation.

The genome rearrangements that help distinguish chimpanzees and humans include 5 million short insertions or deletions involving a few base pairs each, as well as a substantial number of larger insertions, deletions, inversions, or duplications that can involve many thousands of base pairs. When transposon insertions—a major source of genomic variance—are added to the list, the differences between the human and chimpanzee genomes increase. The chimpanzee genome has two classes of retrotransposons that are not present in the human genome (see Chapter 14). Other types of rearrangements, especially segmental duplications, are also common in primate lineages. Duplications of chromosomal segments can lead to changes in the expression of genes contained in the segments. There are about 90 million bp of such differences between human and chimpanzee, representing another 3% of these genomes. In effect, each species has segments of DNA, constituting 40 to 45 million bp, that are entirely unique

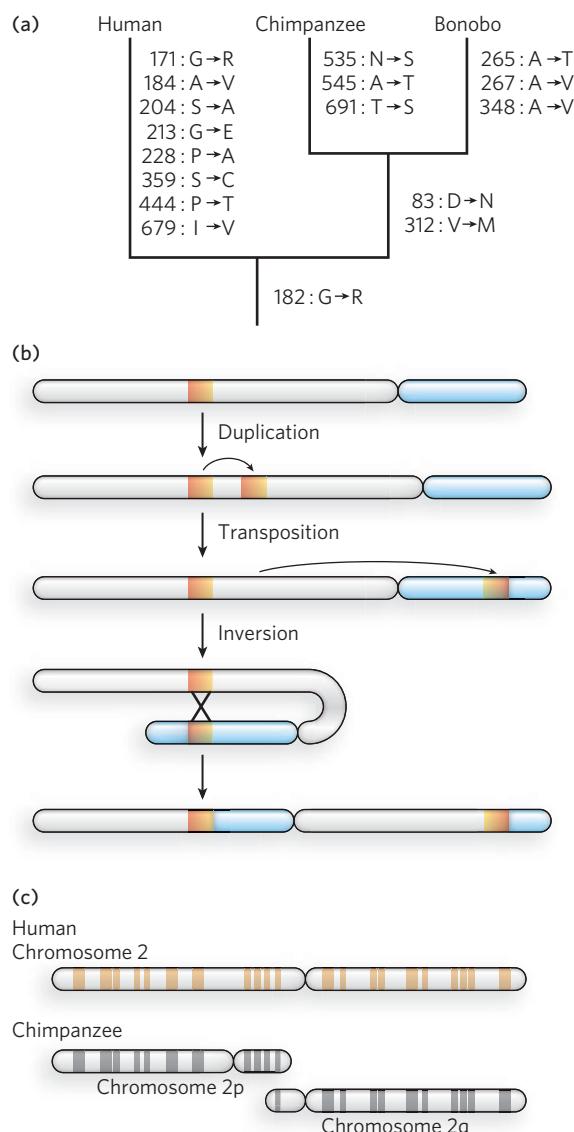


FIGURE 8-8 Genomic alterations in the human lineage.

(a) This evolutionary tree is for the progesterone receptor, which helps regulate many events in reproduction. The gene encoding this protein has undergone more evolutionary alterations than most. Amino acid changes associated uniquely with human, chimpanzee, and bonobo are listed beside each branch (with the residue number). (b) One of the multistep processes that can lead to the inversion of a chromosomal segment. A gene or segment of the chromosome is duplicated, then moved to another chromosomal location by transposition. Recombination of the two segments may result in inversion of the chromosomal DNA between them. (c) The genes on chimpanzee chromosomes 2p and 2q are homologous to those on human chromosome 2, implying that two chromosomes fused into one at some point in the line leading to humans. [Source: (a) Adapted from C. Chen et al., *Mol. Phylogenet. Evol.* 47:637–649, 2008.]

to that particular genome, with larger chromosomal insertions, duplications, and other rearrangements affecting more base pairs than do single-nucleotide changes. Thus, the total genomic difference between chimpanzee and human amounts to about 4% of their genomes.

Sorting out which genomic distinctions are relevant to features that are uniquely human is a daunting task. If the two species share a common ancestor, then, logically, half the changes represent chimpanzee lineage changes and half represent human lineage changes (if one assumes a similar rate of evolution in both lines). When you see a difference, how do you tell which variant was the one present in the common ancestor? One way is to compare both genome sequences with those of more distantly related organisms referred to as **outgroups**. Consider a locus, X, where there is a difference between the human and chimpanzee genomes (Figure 8-9). The lineage of the orangutan, an outgroup, diverged from that of chimps and humans prior to the chimpanzee/human common ancestor. If the

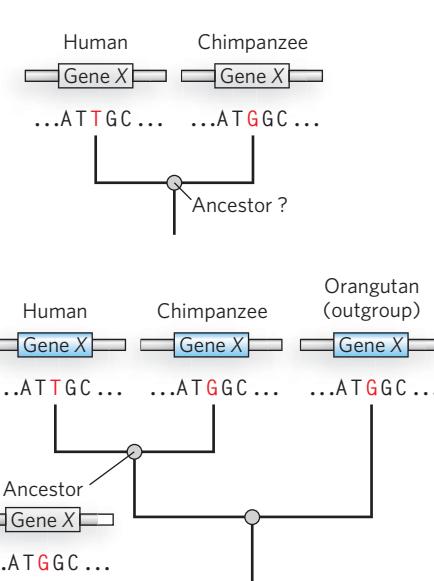


FIGURE 8-9 Determination of sequence alterations unique to one ancestral line. (a) Sequences from the same hypothetical gene in human and chimpanzee are compared. The sequence of this gene in their last common ancestor is unknown. (b) The orangutan genome is used as an outgroup. The sequence of the orangutan gene is found to be identical to the chimpanzee gene. This means that the mutation causing the difference between human and chimpanzee almost certainly occurred in the line leading to modern humans, and the common ancestor of human and chimpanzee (and orangutan) had the variant now found in chimpanzee.

sequence at locus X is identical in orangutan and chimpanzee, this sequence was probably present in the chimpanzee/human ancestor, and the sequence seen in human is specific to the human lineage. Sequences that are identical in human and orangutan can be eliminated as candidates for human-specific genomic features. The importance of comparisons with closely related outgroups has given rise to new efforts to sequence the genomes of orangutans, macaques, and many other primate species.

The search for the genetic underpinnings of special human characteristics, such as our enhanced brain function, can benefit from two complementary approaches. The first searches for genomic regions where extreme changes have occurred, such as genes that have been duplicated many times or large genomic segments not present in other primates. The second approach looks at genes known to be involved in relevant human diseases. For brain function, for example, one would examine genes involved in cognition, such as those that contribute to mental disorders when mutated.

Several factors, such as the development of human-specific life history traits (e.g., a greater age of sexual maturity and a longer generation time), have led to an approximately 3% slower accumulation of genomic changes in the ancestral line leading to humans than in the line leading to chimpanzees. Evolution has occurred somewhat faster in other primate lines. Observed genetic changes are sometimes concentrated in a particular gene or region. In principle, human-specific traits could reflect changes in protein-coding genes, in regulatory processes, or both. A few classes of protein-coding genes exhibit evidence of accelerated divergence (more amino acid substitutions than normal). These include genes involved in chemosensory perception, immune function, and reproduction. In these cases, rapid evolution is evident in virtually all primate lines, reflecting physiological functions that are critical to all primate species. Another class of genes showing evidence of accelerated evolution is those encoding transcription factors, proteins involved in the expression of other genes (see Chapter 21).

Notably, analyses of the human lineage have not detected an enrichment of genetic changes in protein-coding genes involved in brain development or size. Guided in part by the results obtained for transcription factor genes, the focus of such analyses has gradually shifted to changes in gene expression. In primates, most genes that function uniquely in the brain are even more highly conserved than genes functioning in other tissues. This may reflect some special constraints related to brain biochemistry. However, some differences in gene expression are observed. For example, the gene

encoding the enzyme glutamate dehydrogenase, which plays an important role in neurotransmitter synthesis, has an increased copy number due to gene duplication. When changes in genomic regions related to gene regulation are analyzed, genes involved in neural development and nutrition are disproportionately affected. A variety of RNA-coding genes, some with expression concentrated in the brain, also show evidence of accelerated evolution (Figure 8-10). The many new classes of RNA that are now being discovered (see Chapter 22) are likely to radically change our perspective on how evolution alters the workings of living systems.

Genome Comparisons Help Locate Genes Involved in Disease

One of the motivations for the Human Genome Project was its potential for accelerating the discovery of genes underlying genetic diseases. That promise has been fulfilled; well over 1,600 human genetic diseases have been mapped to particular genes. Some disease-gene hunters caution that, so far, the work may have uncovered mostly the relatively easy cases, with many challenges remaining.

The main approach during the past two decades uses a method called **linkage analysis**. In brief, the gene involved in a disease condition is mapped relative to well-characterized genetic polymorphisms that occur throughout the human genome, using methods rooted in evolutionary biology. The search often begins with one or more large families that include several individuals affected by a particular disease. We'll illustrate by describing the search for one gene involved in Alzheimer disease. About 10% of all cases of Alzheimer's in the United States result from an inherited predisposition. Several different genes have been discovered that, when mutated, can lead to early onset of the disease. One such gene (*PS1*) encodes the protein presenilin-1, and its discovery made heavy use of linkage analysis. Two of the many family pedigrees used to search for this gene in the early 1990s are shown in Figure 8-11a. In studies of this type, DNA samples are collected from both affected and unaffected family members. Researchers first localize the region associated with a disease to a specific chromosome. This effort makes use of a set of genomic locations where common SNPs or other mapped genomic alterations occur in the human population, as identified by the Human Genome Project. Using a panel that includes several well-characterized SNP loci mapped to each chromosome, investigators compare the genotypes of individuals with and without the disease, focusing especially on close family members. By identifying the

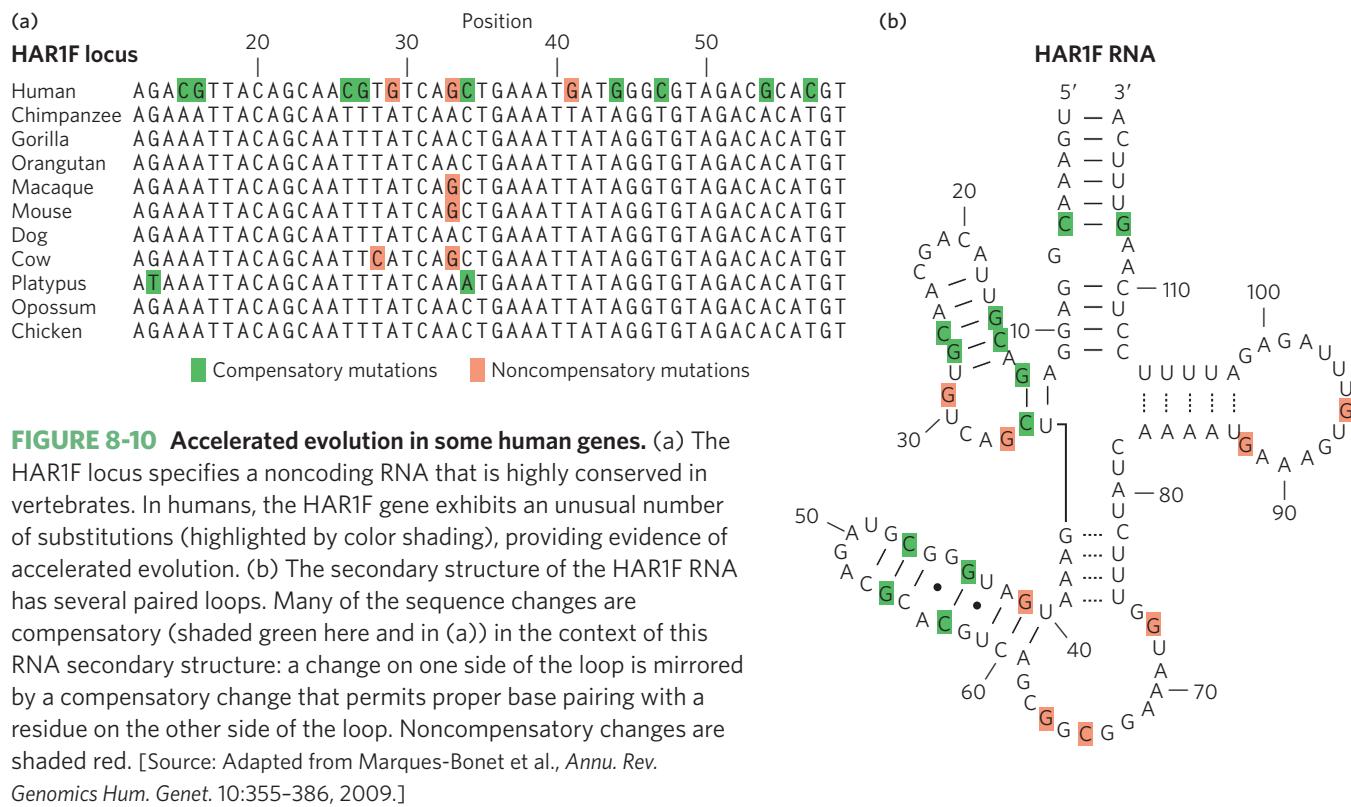


FIGURE 8-10 Accelerated evolution in some human genes. (a) The HAR1F locus specifies a noncoding RNA that is highly conserved in vertebrates. In humans, the HAR1F gene exhibits an unusual number of substitutions (highlighted by color shading), providing evidence of accelerated evolution. (b) The secondary structure of the HAR1F RNA has several paired loops. Many of the sequence changes are compensatory (shaded green here and in (a)) in the context of this RNA secondary structure: a change on one side of the loop is mirrored by a compensatory change that permits proper base pairing with a residue on the other side of the loop. Noncompensatory changes are shaded red. [Source: Adapted from Marques-Bonet et al., *Annu. Rev. Genomics Hum. Genet.* 10:355–386, 2009.]

particular SNPs that are most often inherited with the disease-causing gene, the responsible gene can gradually be localized to a single chromosome. In the case of the *PS1* gene, coinheritance was strongest with markers on chromosome 14 (Figure 8-11b).

Chromosomes are very large DNA molecules, and localizing the gene to one chromosome is only a small part of the battle. On that chromosome is a mutation that is giving rise to the disease. However, in every individual human genome, thousands of SNPs and other changes are present on every chromosome—representing alterations of all kinds relative to the reference sequence in the human genome database. Simply sequencing the entire chromosome would be unlikely to reveal the SNP or other change associated with the disease. The more detailed localization of a disease-causing gene on a chromosome relies instead on an even more elaborate application of linkage analysis. Statistical methods can correlate the inheritance of additional, more closely spaced polymorphisms with the occurrence of the disease, focusing on a denser panel of polymorphisms known to occur on the chromosome of interest. The more closely a marker is located to a disease gene, the more likely it is to be inherited along with that gene. This process can pinpoint a region of the chromosome that contains the gene. However, the region may still contain a long length of DNA encompassing many genes. In our example, linkage analysis

indicated that the disease-causing gene was somewhere near a SNP locus called D14S43 (Figure 8-11c).

The final steps again use the human genome databases. The local region containing the gene is examined and the genes within it are identified. DNA from many individuals, some who have the disease and some who do not, is sequenced over this region. This process, with an increasing number of individuals analyzed, gradually leads to the identification of gene variants consistently present in individuals with the disease state, and not in unaffected individuals. The search can be aided by an understanding of the function of the genes in the target region, because particular metabolic pathways may be more likely than others to produce the disease state. In 1995, the chromosome 14 gene associated with Alzheimer disease was identified as gene *S182*. The product of this gene was given the name presenilin-1, and the gene itself was subsequently renamed *PS1*.

Many human genetic diseases are caused by mutations in a single gene, and the defect is inherited in Mendelian patterns (see Chapter 2). Several different mutations in a particular gene, all leading to the same or related genetic condition, may be present in the human population. There are several variants of *PS1*, for example, all giving rise to a much increased chance of early-onset Alzheimer's. Another, more extreme example is the several genes encoding different

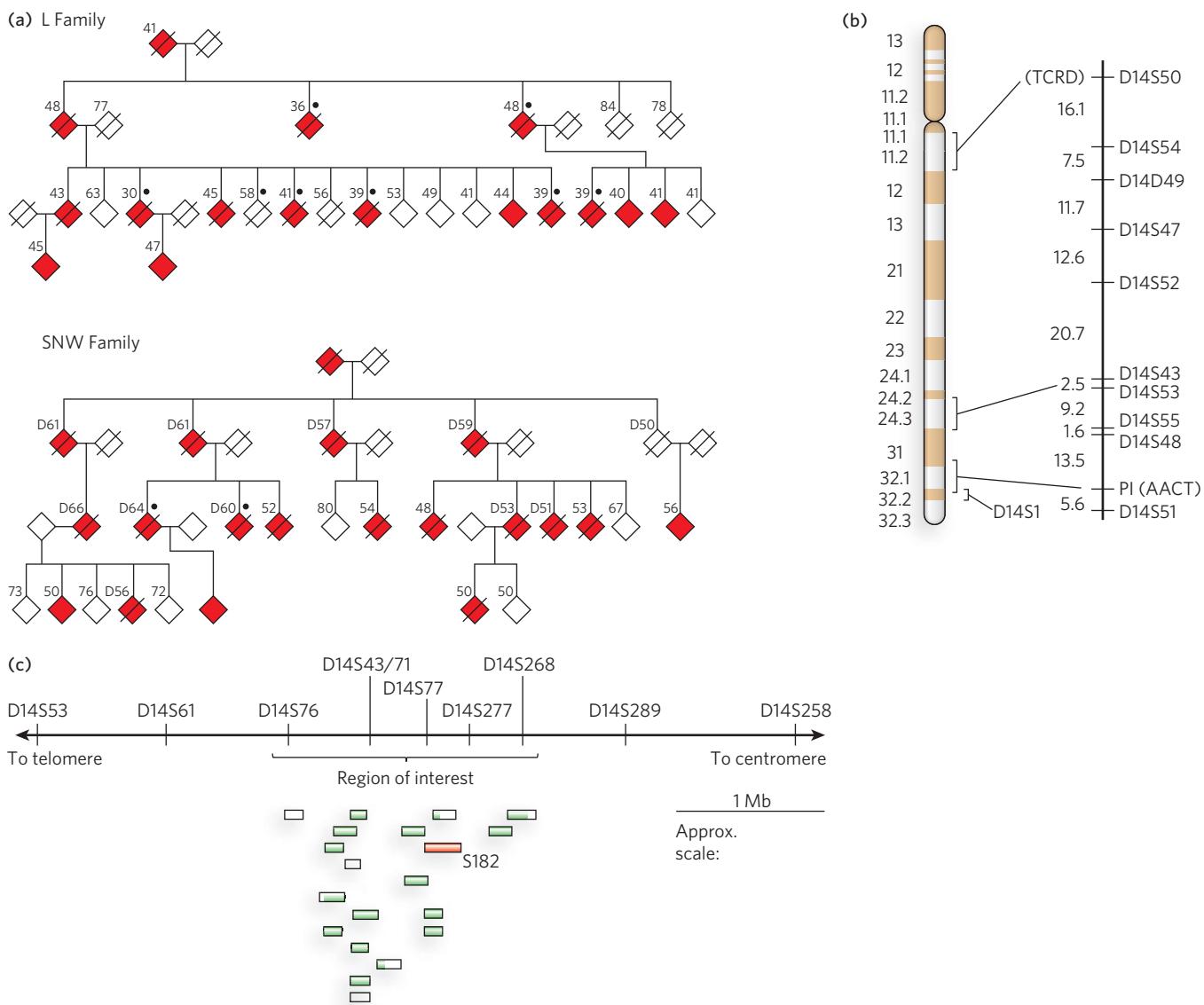


FIGURE 8-11 Linkage analysis in the discovery of disease genes.

genes. (a) These pedigrees for two families affected by early-onset Alzheimer disease are based on the data available at the time of the study. Filled symbols represent affected individuals; slashes indicate deaths. The number above each symbol is the person's age at onset of symptoms (for affected individuals), at time of the study (for living unaffected individuals), or at time of death (for deceased unaffected individuals and others marked by "D"). Black dots indicate an autopsy was done to verify the presence of Alzheimer disease. To protect family privacy, gender is not indicated. (b) Chromosome 14, with

bands created in metaphase by certain dyes, has marker positions shown at the right, with the genetic distance between them in centimorgans. TCRD (T-cell receptor delta) and PI (AACT, α 1-antichymotrypsin) are genes with variations in the human population that were used as markers, along with SNPs, in chromosome mapping. (c) A region of interest that contains 19 expressed genes was eventually defined near marker D14S43. The gene labeled S182 (red) encodes presenilin-1. [Sources: (a), (b) Adapted from G. Schellenberg et al., *Science* 258:668, 1992. (c) Adapted from R. Sherrington et al., *Nature* 375:754, 1995.]

hemoglobins: more than 1,000 known mutational variants are present in the human population. Some of these variants are innocuous; some cause diseases ranging from sickle-cell anemia to thalassemias. The inheritance of particular mutant genes may be concentrated in families or in isolated populations.

More complex are cases where a disease condition is caused by the presence of mutations in two different genes (neither of which, alone, causes the disease), or where a particular condition is enhanced by an otherwise innocuous mutation in another gene. Identifying the genes and mutations responsible for such digenic

diseases is exceedingly difficult, and these diseases are sometimes possible to document only within small, isolated, and highly inbred populations.

Modern genome databases are opening up alternative paths to the identification of disease genes. In many cases, we already have biochemical information about the disease. In the case of Alzheimer's, an accumulation of the amyloid β -protein in limbic and association cortices of the brain is at least partly responsible for the symptoms. Defects in presenilin-1 (and in a related protein, presenilin-2, encoded by a gene on chromosome 1) lead to elevated cortical levels of amyloid β -protein. Focused databases are being developed that catalog such functional information on the protein products of genes, and on protein-interaction networks (determined by methods described in Section 8.2), SNP locations, and other data. The result is a streamlined path to the identification of candidate genes for a particular disease. If a researcher knows a little about the kinds of enzymes or other proteins likely to contribute to a disease, these databases can quickly generate a list of genes known to encode proteins with relevant functions, additional uncharacterized genes with orthologous or paralogous relationships to the genes in this list, a list of proteins known to interact with the target proteins or orthologs in other organisms, and a map of gene positions. With the aid of data from some selected family pedigrees, a short list of potentially relevant genes can often be determined rapidly.

These approaches are not limited to human diseases. The same methods can be used to identify the genes involved in diseases—or genes that produce desirable characteristics—in other animals and in plants.

SECTION 8.1 SUMMARY

- A genome is one copy of the complete genetic complement of an organism. Thousands of complete genome sequences are now available. The Human Genome Project was undertaken by two competing teams that used different strategies of shotgun sequencing.
- The sequencing of a genome is followed by genome annotation, an attempt to summarize the locations and functions of genes and other sequences.
- The human genome contains approximately 25,000 genes, fewer than expected. Only 1.1% to 1.4% of the human genome encodes proteins; the remainder is made up of transposons, functional RNA-encoding genes, introns, sequences involved

in gene regulation, and tandem repeats of short sequences.

- The sequencing of multiple primate genomes is opening windows into human evolution. Genomic alterations that are specific to the human lineage occupy about 4% of our genome, with large genomic rearrangements such as transposon insertions and segmental duplications playing a larger role than single-nucleotide polymorphisms.
- Genome sequence databases facilitate the search for genes that specifically contribute to particular traits, and for genes involved in disease.

8.2 Transcriptomes and Proteomes

A gene is not simply a DNA sequence; it is information that is converted into a useful product—a protein or functional RNA molecule—when and if needed by the cell. We now turn to methods that contribute to our understanding of the functions of these gene products. The methods can be applied to efforts to study the response of a cell or organism to particular events or changes in the environment. They can also be used to help identify the functions of the many genes in every genome for which we know very little about their roles in the cell.

The study of complex interconnected processes in biology is called **systems biology**. Genome sequencing contributes to systems biology by providing information about all the genes in an organism. The methods we now address contribute more directly by examining the expression of genes or the interactions of many kinds of proteins under specified sets of conditions. Many of the methods were described in Chapter 7. Here, we focus on increasingly complex problems in cellular metabolism.

Special Cellular Functions Are Revealed in a Cell's Transcriptome

Only a subset of the many genes in a genome is expressed in any given cell. That subset may change in response to changes in the cellular environment or to extracellular signals of many kinds. The genes expressed in a cell under a given set of conditions constitute its **transcriptome**. Studies of the transcriptome, carried out by researchers in the subdiscipline of **transcriptomics**, can help reveal new cellular processes, as well as identify the genes and gene products involved in known processes. If the function of a gene is

not known, an understanding of the circumstances that result in expression of that gene can provide an important functional clue.

Transcriptome analysis was first made practical with the advent of microarray technologies (see Figures 7-28 and 7-29). Microarrays can reveal the genes that are newly induced when a cell is subjected to heat shock, variations in expression patterns in different regions of a mammalian brain, changes that occur when a pathogenic bacterium invades a host organism, and so on. The growing use of microarray-based transcriptome analysis has led to the development of online databases, some specific to a single organism, that make data available to the entire scientific community. As the quality of transcriptome data improves, the transcriptomes themselves become more than a list of expressed genes. They are also a kind of fingerprint that characterizes a class of cell under a given set of conditions. These databases are rapidly finding uses not only in basic research but in medicine as well.

For instance, the cells making up a tumor exhibit characteristic patterns of gene expression—a transcriptional profile—that may differ greatly from one tumor to the next. These profiles can provide a kind of tumor fingerprint, which can be used to predict a patient's prognosis and/or select the most beneficial therapies. The value of these tools to oncologists and patients will only increase as the technologies become more widespread.

Recent progress in the diagnosis and treatment of breast cancer illustrates the potential of the technology. Broad clinical studies over the past decade have used microarrays to develop transcriptional profiles of many thousands of breast cancers. Treatment protocols have been tracked, and the successes and failures carefully documented. Researchers are gradually identifying specific genes and groups of genes that, when expressed at higher levels and in certain combinations, serve as prognostic indicators. The result is a growing database of correlations that allows the use of transcriptional profiles to develop both prognoses and treatments.

High-Throughput DNA Sequencing Is Used in Transcriptome Analysis

Microarrays have some disadvantages for transcriptome analysis. They can provide inaccurate information about relative levels of transcription for genes that are expressed at very low or very high levels. In addition, they can miss any RNAs that are not homologous to genes included on the microarray. A newer method, called **RNA-Seq**, has been developed to ad-

dress these shortcomings, taking advantage of modern high-throughput DNA sequencing technologies (see Highlight 7-2).

A typical RNA-Seq experiment is shown in **Figure 8-12**. RNA is isolated from the cell or tissue to be analyzed. In most cells, rRNA is by far the most abundant RNA, but it is usually other types of RNA that are of most interest. Thus, most protocols include a step involving subtractive hybridization of the rRNA, using complementary probes that allow removal of the hybridized material. The remaining RNA is then converted to cDNA with the enzyme reverse transcriptase (see Figure 7-8). The cDNA is fragmented to an appropriate average length. Short adapter DNA segments that provide target sequences for the primers needed for DNA sequencing are ligated to both ends. Each cDNA is then “read” by DNA sequencing. Huge numbers of these short sequencing reads (typically 30 to several hundred base pairs, depending on the sequencing technology used) are produced. The gene from which each sequencing read is derived is determined by computerized alignment with the same sequence in the relevant genome database. Genes expressed at high or low levels are represented by correspondingly high or low levels of sequence reads. Gene expression levels can be mapped across genes, chromosomes, and entire genomes.

RNA-Seq provides information on gene expression levels with a much greater dynamic range and has proved highly accurate when compared with more laborious and quantitative methods. The direct sequencing also provides additional information, showing the exact transcriptional boundaries of genes and revealing how exons are linked together in transcripts. In genes that have alternative splicing patterns (see Figure 8-5), the method can also reveal which exons within a single gene are being expressed at higher levels. In some organisms, RNA transcripts are edited, producing new sequences not present in the DNA genes (see Chapter 16). These sequence changes are directly revealed in RNA-Seq. As the costs of high-throughput DNA sequencing decrease, RNA-Seq may well replace microarrays as the method of choice for transcriptome analysis.

The Proteins Generated by a Cell Constitute Its Proteome

The word “proteome” first appeared in the research literature in 1995. A cell’s **proteome** is the complement of proteins present in that cell under a given set of conditions, and the subdiscipline of **proteomics** includes efforts to define the proteome. More broadly,

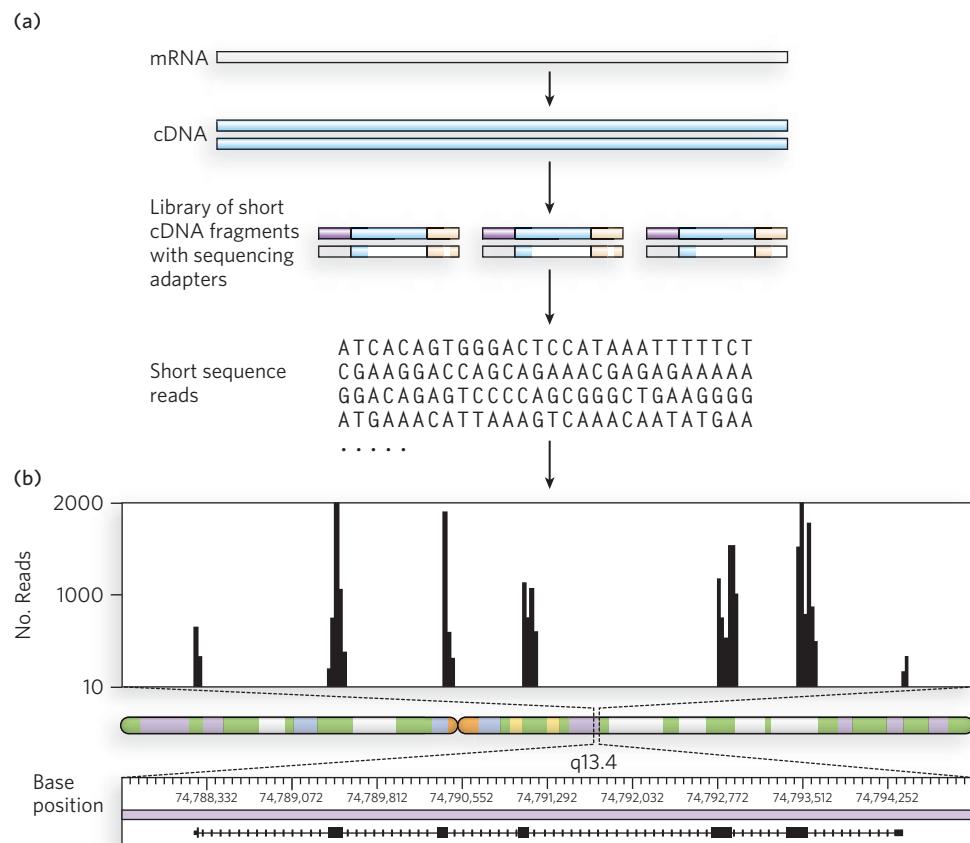


FIGURE 8-12 Use of RNA-Seq to examine transcriptomes (a) An mRNA is isolated, transcribed to cDNA, fragmented into smaller pieces (e.g., by shearing or nuclease digestion), and ligated to adapter oligonucleotides that provide targets for sequencing primers. Sequencing then follows, using one of the methods described in Highlight 7-2. (b) The number of times a sequence from a given gene or segment of a gene appears in a sequencing read (i.e., the number of reads containing all or part of that sequence) is plotted. The number of reads from a given genomic region reflects the relative level of mRNA produced from that region. Data for a small portion of human chromosome 11, segment q13.4, are shown here.

any effort to analyze a complex mixture of proteins, whether or not it applies to all the proteins in a cell, falls under the proteomics umbrella. For example, some studies are directed at the proteins in a specific organelle or the proteins embedded in the cytoplasmic membrane.

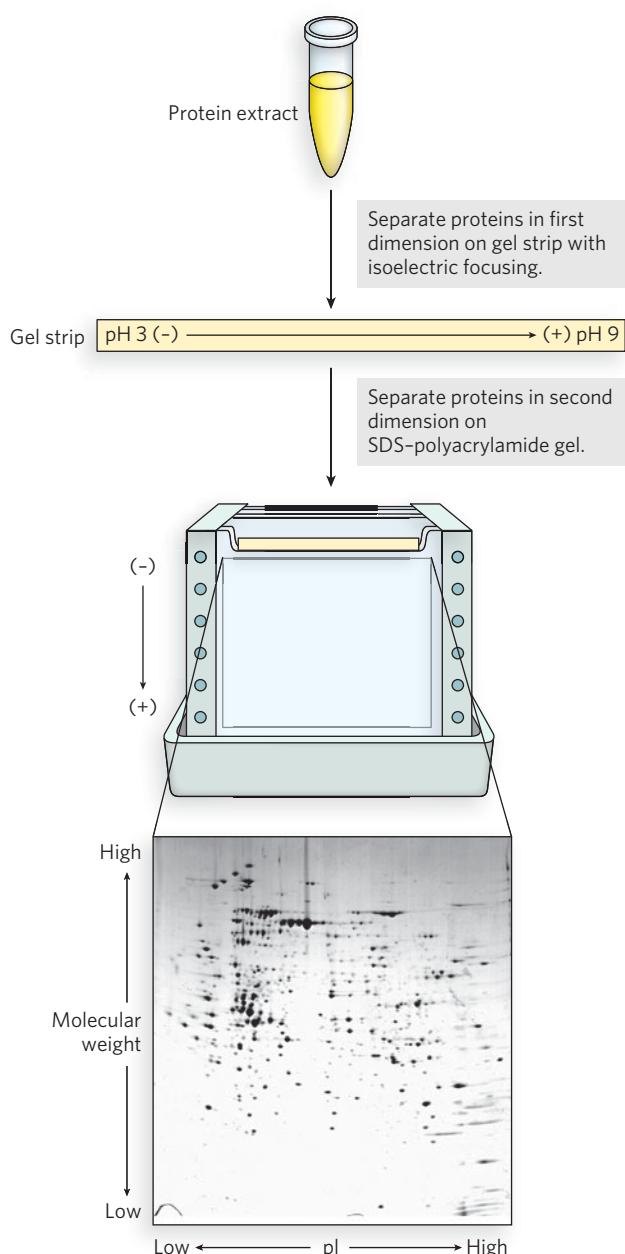
The problems that proteomics researchers explore can be straightforward to describe, but the solutions often are not. Each genome presents us with thousands of protein-encoding genes. We wish to know which proteins are present and contributing to cellular metabolism under every possible set of circumstances. Analysis includes the structure, posttranslational modifications, cellular localization, and detailed function of all those proteins, and how the various proteins interact. Given that many proteins can still reveal surprises even after years of study, the investigation of an entire proteome or any complex

protein mixture is daunting. Simply discovering the functions of newly described proteins requires intensive work. Biochemists can now apply shortcuts in the form of a broad array of new and updated technologies and databases that address protein function on a cellular level.

Transcriptome information tells us about RNA levels in a cell, but this does not necessarily inform us about protein levels. The expression of many genes, particularly in eukaryotes, is regulated at the level of translation. Messenger RNAs for particular genes can be stored in the cell in an inactive state until the protein product of that gene is needed. In addition, many proteins are initially synthesized in an inactive state, their function dependent on posttranslational modifications. A complete understanding of a proteome requires all this information about the modification status of its complement of proteins.

Electrophoresis and Mass Spectrometry Support Proteomics Research

The resolution power of polyacrylamide gel electrophoresis can be amplified by carrying out two electrophoretic steps in succession, separating proteins on the basis of different properties; the technique is known as **two-dimensional gel electrophoresis** (Figure 8-13). The first step (or dimension) uses isoelectric focusing, a method that separates proteins on the basis of their isoelectric point, or pI (the pH at which the net charge of a protein is zero). Polyacrylamide gels containing an immobilized pH gradient are commercially available.



Voltage is applied across the gel. Proteins migrate through the gel, halting where the pH of the strip equals the pI of the protein. In the second step, the gel strip is laid on top of another gel, and electrophoresis is carried out at 90° to the first step (i.e., in the second dimension), this time using an SDS-polyacrylamide protocol to separate proteins according to size.

This technique allows the separation and display of up to 1,000 different proteins on a single gel. After staining to visualize the proteins, the gel can be compared with similar gels displaying the proteins in extracts taken from the same types of cells but under different conditions. The appearance (or disappearance) of particular protein spots in different samples can help define the cellular function of these proteins. Individual spots on the gel can be digested with the protease trypsin. **Mass spectrometry** can then be used to partially sequence individual peptides derived from the spots and to assign each to a protein. Many sophisticated protocols coupling electrophoresis and mass spectrometry lie at the heart of proteomics research.

The mass spectrometer has long been an indispensable tool in chemistry. Molecules to be analyzed, referred to as **analytes**, are first ionized in a vacuum. When the newly charged molecules are introduced into an electric and/or magnetic field, their paths through the field are a function of their mass-to-charge ratio, m/z . This measured property of the ionized species can be used to deduce the mass (M) of the analyte with very high precision.

Mass spectrometry provides a wealth of information for proteomics research, enzymology, and protein chemistry in general. Because the techniques require only minuscule amounts of sample, they are readily applied to the small amounts of protein that can be extracted from a two-dimensional electrophoretic gel. The accurately measured molecular mass of a protein is one of the critical parameters in its identification. Once a protein's mass is accurately known, mass spectrometry is also a convenient and accurate method for detecting

FIGURE 8-13 Two-dimensional gel electrophoresis of a complex mixture of proteins. In the first dimension, proteins are separated according to pI (i.e., charge) on a gel strip by isoelectric focusing. The strip is then laid on top of an SDS-polyacrylamide gel, and the proteins are separated according to size by electrophoresis. The original protein complement is thus spread in two dimensions, aiding the separation of similar proteins into individual spots. The spots can be cut out of the gel and the proteins identified by mass spectrometry. [Source: A. Mogk, T. Tomoyasu, et al., *EMBO J.* 18:6934–6949, 1999. Photo courtesy of Axel Mogk.]

changes in mass due to bound cofactors, bound metal ions, covalent modifications, and so on.

Even a short sequence is often enough to permit unambiguous association of a protein with its gene, if the gene sequence is known. If the protein is fragmented, short peptides can be sequenced by mass spectrometry. This is ideal for proteomics research aimed at cataloging the hundreds of cellular proteins that might be separated on a two-dimensional gel. In the coming decades, detailed genome sequence data will become available for hundreds, and eventually thousands, of organisms. The ability to rapidly associate proteins with genes through the use of mass spectrometry will greatly facilitate the exploitation of this extraordinary information resource.

Computational Approaches Help Elucidate Protein Function

With the number and size of databases increasing rapidly, the information required to answer a biological question may be right at one's fingertips. Increasingly, data-mining is complementing experimentation as a highly productive path to mechanistic and functional insights about genes, RNAs, and proteins.

Sequence Relationships A wide range of conserved amino acid sequences associated with structural motifs involved with particular functions can be identified within a protein or set of proteins (see Chapter 4). The motifs often correspond to binding (e.g., ATP, nucleic acids, NAD⁺, metals) or catalytic activities (e.g., helicase, polymerase, ATPase). The presence of a structural motif may suggest, for example, that the protein catalyzes ATP hydrolysis, binds to DNA, or forms a complex with zinc ions, thus helping define molecular function. Resources for conducting such searches are available on the NCBI and Ensembl websites.

Structural Relationships Accurate determination of the three-dimensional structure of proteins is not always successful, but efforts are so common that structural databases are replete with protein structures of all types. To further the assignment of function based on structural relationships, the research community has initiated a large-scale structural proteomics project. The goal is to crystallize and determine the structure of as many proteins and protein domains as possible, in many cases with little or no existing information about protein function. The project has been assisted by the automation of the tedious empirical screening of hundreds of solution conditions, as required to crystallize a particular

protein (see Chapter 4). As these structures are solved, they are made available in the structural databases (see Highlight 4-2). This effort should help define the extent of variation in structural motifs. When a newly discovered protein is found to have structural folds that are clearly related to motifs with known functions in the databases, this information can suggest a molecular function for the protein.

Comparisons of Genome Composition Although not evidence of direct association, the mere presence of combinations of genes in certain genomes can hint at protein function. One can simply search the genome databases for specific genes, then determine which other genes are present in the same genomes—a process known as **phylogenetic profiling** (Figure 8-14). (Phylogenetics is explained in more detail in Section 8.3.) The consistent appearance of two genes together in a genome suggests that the proteins they encode may be functionally related. Such correlations are most useful if the function of at least one of the proteins is known.

Phylogenetic profiling is often carried out on hundreds or thousands of genes at once, in broad studies that complement approaches such as linkage analysis. The search for a gene called *BBS5*, involved in Bardet-Biedl syndrome (BBS), provides an example. Bardet-Biedl syndrome is a serious genetic condition characterized by retinal degeneration, obesity, a variety of physical deformities, and learning disabilities. Six *BBS* genes discovered before *BBS5* were found to be involved in the function of the flagellar and basal body. *BBS5* had been localized to a region in chromosome 2. To facilitate identification of the gene in this region, the

Protein	Species			
	1	2	3	4
P1	+	-	+	+
P2	-	-	+	-
P3	+	+	-	+
P4	+	-	+	-
P5	+	-	-	-
P6	+	+	-	+
P7	+	+	+	-

FIGURE 8-14 Use of comparative genomics to identify functionally related genes. This example of phylogenetic profiling shows gene comparisons for four organisms. P1 through P7 indicate proteins encoded by each species. The + or – indicates presence or absence of the protein. The technique does not require homologous proteins. Because proteins P3 and P6 always appear together in a genome (red shading), they may be functionally related. In particular, they may have a function that is found in species 1, 2, and 4, but not in species 3. Further testing would be needed to confirm this inference.

researchers did a phylogenetic profile, comparing genes of human and the green alga *Chlamydomonas*, species that possess flagellar and basal bodies, with the plant *Arabidopsis* that lacks this cellular structure. They generated a list of 688 genes present in human and *Chlamydomonas* but absent in *Arabidopsis*. The region of chromosome 2 that interested the researchers had a total of 230 genes, but only 2 of them were on the list of 688 generated by the phylogenetic profile. One of these turned out to be *BBS5*.

Experimental Approaches Reveal Protein Interaction Networks

Every protein functions by interacting with other molecules, from small metabolites to nucleic acids and other proteins. One of the strongest clues to protein function is a knowledge of what other proteins that protein interacts with. For example, if a protein of unknown function interacts with an RNA polymerase, there is a good chance that the protein is also involved in transcription. Powerful new technologies are providing information on protein interaction networks in cells.

Protein Chips For large-scale studies, proteins, like nucleic acids, can be immobilized on a solid surface to form protein chips. These can be used to detect the presence or absence of other proteins in a sample. For example, an array of antibodies to a particular set of proteins is immobilized as individual spots on a solid surface. A sample of proteins is added, and any proteins that bind an antibody on the chip can be detected by a variety of methods. However, whereas DNA is consistent in its physicochemical properties and is readily immobilized on silicon chips, proteins vary a great deal in their properties, and the construction of protein chips can be challenging. The conformation of many proteins depends on solution conditions, and immobilization on a silicon chip may inactivate some in a manner that is not always predictable. Nevertheless, many successful efforts have been reported.

Probing Macromolecular Interactions In Vivo The study of protein-protein interactions by the two-hybrid method and the study of protein-RNA interactions by the three-hybrid method rely on macromolecular interactions that occur *in vivo*. Both are important avenues for examining protein interaction networks in proteomics research. A somewhat different approach to detecting protein interactions *in vivo* involves immunoprecipitation of proteins from cell extracts. Antibodies are used to precipitate a given protein, and the precipitate is examined to identify any other proteins that were associated with the target

protein in the cell and are thus precipitated with it. All of these techniques are described in Chapter 7.

Miscellaneous Approaches The proteomics literature is replete with examples of creative approaches to dissecting protein interaction networks. One approach is the search for “Rosetta stone” fusions. Sometimes, two proteins that exist as separate entities in species 1 may have orthologs in species 2 that are the product of two fused genes. This fusion in species 2 makes it highly likely that the two proteins in species 1 interact. Another approach simply mines the biochemical literature, focusing on proteins that are mentioned together in the same article. If two proteins are mentioned together in a large number of publications, the assumption is made that the two may interact.

SECTION 8.2 SUMMARY

- A transcriptome is a listing of the genes that are expressed in a given cell under a defined set of conditions. The transcriptome may change in response to environmental changes or cellular signals.
- Microarrays provide one picture of cellular transcriptomes. The RNA-Seq approach is even more effective in generating a detailed transcriptome.
- A proteome is a compilation of all the proteins present in a given cell under a defined set of conditions. Computational and experimental techniques explore the proteome, and the functions of the proteins it encompasses, on a cellular scale.
- The most common approach to examining a cellular proteome under a defined set of conditions involves two-dimensional gel electrophoresis coupled to protein identification by mass spectrometry.
- The generation of protein interaction networks is one of the goals of proteomics research. Techniques include protein chips, two-hybrid and three-hybrid methods, immunoprecipitation, and protein fusions.

8.3 Our Genetic History

Perhaps more than any scientific discipline that preceded it, genomics provides the doorway to an especially informative and often quantitative study of evolution. Genomics research carries important implications that go to the heart of human existence. Where did we come from? How did we get to where we are now? Our rapidly

increasing understanding of genomes has greatly advanced the scientific answer to these and many other fundamental questions. As interesting as the quest may be, however, it can sometimes seem to be simply an academic exercise. Yet a better understanding of how new species evolve and how we are related to one another and to other species is highly relevant to advancing knowledge in areas ranging from ecosystems to pandemics. Answers can pay huge dividends in medicine, agriculture, resource management, and general quality of life.

All Living Things Have a Common Ancestor

One goal of modern genomics and evolutionary biology is the reconstruction of the evolutionary tree that traces the origin of every extant species. We can approach this problem from both ends—the first living creature and the current list of living organisms—and explore how genomics can help trace the path between them.

The successful living entity that gave rise to all non-viral life now on Earth is referred to as **LUCA, the last universal common ancestor**. Although its biological form and genome are obscured by billions of years of evolution, there are several approaches to thinking about LUCA. The first approach attempts to assemble the list of genes and other features that are currently shared by every living creature, an effort greatly facilitated by modern genome sequencing efforts. These features were probably present in LUCA. The second approach is an effort to define the minimum set of genes necessary to support a living cell. Such a minimalist cell would help define the essence of the free-living state and would provide a more complete understanding of the basic problem of life and the threshold of complexity that would have to be breached by the first viable cell.

As revealed by genomics and work in many other biological fields, current organisms share several features that permit a trace to a common ancestor. The core components of the translation and transcription machinery in all cells are demonstrably related. The use of the D isomers of sugars in cells and the L isomers of amino acids in protein synthesis is also universal. At this point, the generalities start to break down. All organisms consist of cells surrounded by lipid-containing membranes, but the composition and structure of those membranes can vary greatly from one group to another. For example, bacteria have membranes consisting mostly of fatty acid esters, whereas membranes in the archaea consist of isoprene ethers. All organisms replicate their DNA, but the replication machinery also varies in important ways.

Current estimates for the number of genes shared by all known species vary from 80 to about 500. The

lower number focuses on genes with clearly identifiable orthologs in all organisms, based on sequence comparisons. The higher number includes genes required for processes that are found in all organisms, but for which many mechanistic and sequence similarities have been obscured by evolutionary time. A cell with 500 components would be simpler than most existing life forms, but still very complex. There must have been many intermediates of gradually increasing complexity in the process that led to LUCA.

The search for the minimal genome begins with the assumption that this cell will grow in a stress-free laboratory environment with abundant resources and constant temperature. The goal is to define the minimum group of components for supporting life, without the specialized functions required for particular world environments. Bacteria with small genomes are a useful place to start.

The bacterium *Mycoplasma genitalium*, a parasite of the genital and respiratory tracts of primates, has the smallest genome sequenced thus far for a defined organism. Its 580,000 bp DNA includes 521 genes, 482 of which encode proteins. Directed efforts to disrupt individual genes has revealed that the bacterium can dispense with only 97 of them and still retain viability in the laboratory, giving a minimal complement of 385 genes. The small genome of this *Mycoplasma* reflects its sheltered environment as a parasite.

Similar experiments have indicated that the minimal genome for an organism living autonomously includes about 1,350 genes. Attempts to define a minimal gene complement are fueling efforts to create an artificial cell from nonliving chemical components, an achievement that would mark a new level of understanding of living systems.

Genome Comparisons Provide Clues to Our Evolutionary Past

One goal of modern biology is to reconstruct the complete tree of life, relating all existing species to their ancestors and ultimately to LUCA. It is a massive project involving many laboratories, and is a natural outgrowth of genome sequencing projects. The evolutionary relationship among species, populations, or genes is known as a **phylogeny**, and the study of such relationships is called **phylogenetics**.

Phylogenetics helps biologists classify organisms. It can also reveal important information about the evolution of traits in an organism or the appearance of new pathogens. It can even aid in criminal investigations (Highlight 8-3). Phylogenies are usually described with the aid of phylogenetic trees, which can be based on

HIGHLIGHT 8-3 EVOLUTION

Phylogenetics Solves a Crime

In the summer of 1994, a nurse in Lafayette, Louisiana, broke off a messy 10-year affair with a physician. The nurse had donated blood to a local blood bank on several occasions; she was tested and found negative for HIV in October 1992, May 1993, and April 1994. The physician had been giving the nurse vitamin shots for fatigue. He gave her one more of these shots, somewhat against her will, in August 1994, after the breakup. In late 1994, the nurse became ill and tested positive for both HIV-1 and hepatitis C, although she had no history of contacts that could have led to the infections. The nurse accused the doctor of infecting her with HIV.

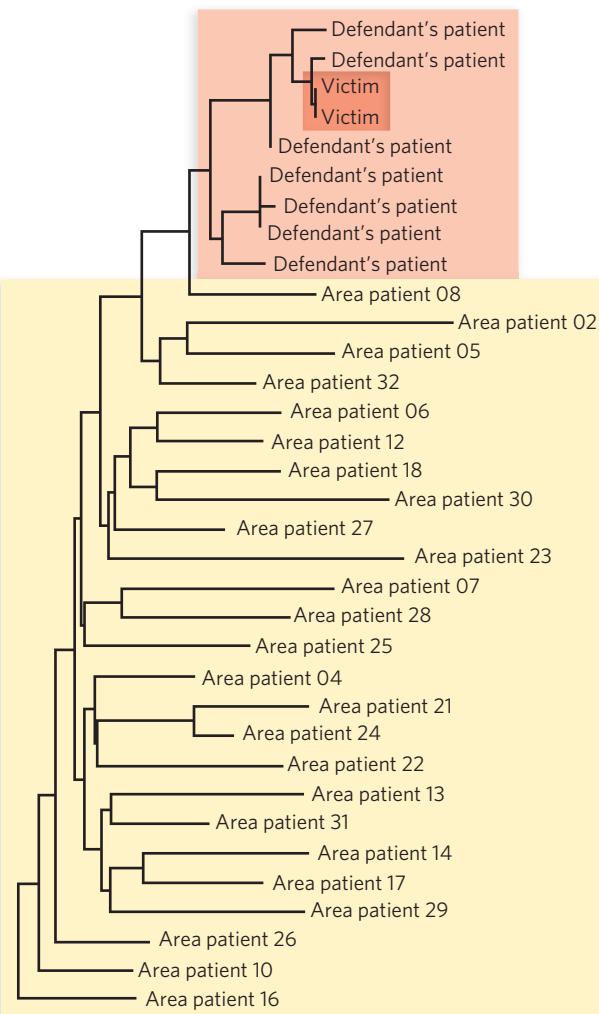
Investigators found records indicating that the doctor had treated and drawn blood from his only HIV-infected patient and a hepatitis C-infected patient just before giving the nurse the injection in August 1994. But how does one link the patients' blood to the nurse-victim in this case? The subsequent trial of the doctor was the first to use phylogenetics in a court case. The investigation focused on the HIV infection.

Once HIV begins to replicate in a new host, the virus mutates rapidly, an evolution that occurs within one infected individual. Samples taken from a person with HIV years after infection can be used to build a phylogenetic tree that can trace the evolution of the virus in that individual. Blood samples were collected from the doctor's HIV-infected patient and from the nurse. Control samples were collected from 30 different HIV-positive patients selected at random in the Lafayette area. The HIV in the samples was sequenced and analyzed independently by two different laboratories at Baylor University and the University of Michigan. Both analyses yielded the

FIGURE 1 A phylogenetic tree reveals the diversity of HIV samples in the Lafayette area. The part of the tree derived from the doctor's HIV-positive patient is highlighted, with the nurse-victim's DNA clearly nested within this set of sequences. [Source: Adapted from M. L. Metzker et al., *Proc. Natl. Acad. Sci. USA* 99:14,292–14,297, 2002.]

same result. The phylogenetic analysis of the victim's HIV strains showed that they were most closely related to, and nested within, the strains from the doctor's patient (Figure 1).

With this and other evidence, the doctor was convicted of attempted second-degree murder in 1998. The verdict was upheld by a Louisiana appeals court in 2000, and the U.S. Supreme Court refused to hear the case in 2002, ending court proceedings. The same methodology has since been used in rape and child abuse cases.



sequence information or on other attributes of a species, such as morphological characteristics. The construction of evolutionary trees was imprecise and descriptive until the 1950s, when mathematical biologists began to systematize the process. That work continues.

On one level, the branched evolutionary trees often depicted in the literature are almost self-explanatory. However, they are based on a host of underlying assumptions and conventions, and the structures in the tree have specific meanings (Figure 8-15a). The subjects

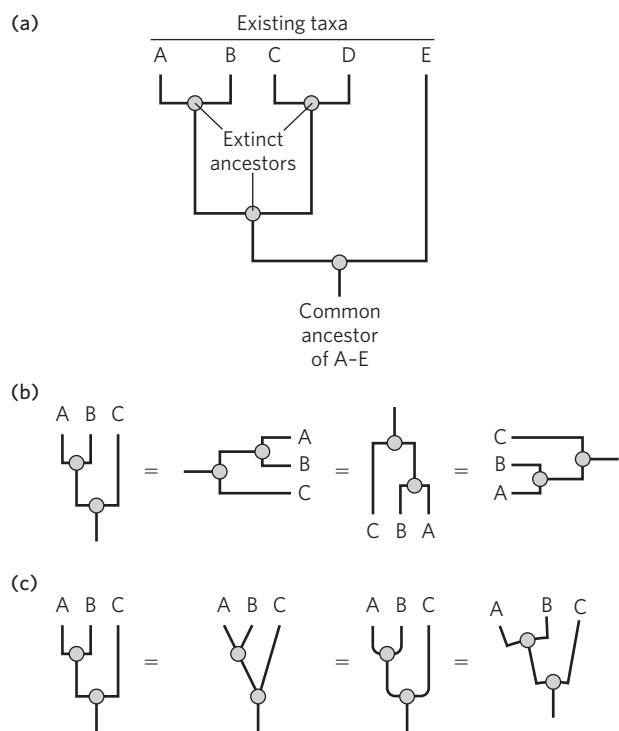


FIGURE 8-15 Phylogenetic trees. An evolutionary tree consists of branches (usually bifurcating) connected by nodes. (a) Basic conventions. The ends of external (upper) branches represent existing taxa, the nodes represent extinct ancestors, and the root end represents the common ancestor of the taxa included in the tree. (b) The orientation of the tree does not matter, and (c) there are several common (and equivalent) depiction styles.

of an evolutionary tree are groupings of organisms known as **taxa**. A taxon is any such grouping, and it can refer to an individual species (*Homo sapiens*), a genus (*Homo*), a class (*Mammalia*), particular populations of a single species, and so on. The tips or ends of the branches on the tree represent the taxa being studied and often reflect species or groups of species now in existence. Each branch point, or node, represents an extinct ancestral species common to the two connected branches. The node at the base of the tree signifies the common ancestor of all the taxa represented in the tree. This is sometimes called the root of the tree. As shown later, not all trees are rooted. Generally, one species is thought to give rise to two, leading to a bifurcating tree. Uncertainty about some evolutionary relationships can lead to the generation of a tree with multiple branches at a single node; such a node is called a polytomy. Node orientation and rotation are arbitrary (Figure 8-15b). Many different tree depictions are in use, with different branch shapes (Figure 8-15c). The

choice of representation is made simply for convenience or personal preference.

Branch lengths can be meaningless. However, they often represent some measure of evolutionary time, such as numbers of altered morphological features, or (more commonly in modern trees) numbers of mutations in one or several genes or numbers of genomic alterations in a region of the genome (Figure 8-16a). For example, we can look at the differences in a gene found in both human and chimpanzee and determine which variant existed in the common ancestor, such as by examining outgroups, as described in Section 8.1. Once the sequence of the gene in the common ancestor is determined, the common ancestor becomes a node in the tree. The lengths of the branches leading from ancestor node to human and chimpanzee reflect the number of changes occurring between that ancestor and the living species.

Numbers next to branches on a tree usually reflect the level of confidence the investigator has in the information contained in that branch (see Figure 8-16a, right). A common method for setting confidence limits is the bootstrap analysis, which gives a range from 100 (very high confidence) to 0 (no confidence). In brief, the bootstrap is a statistical method that starts with the

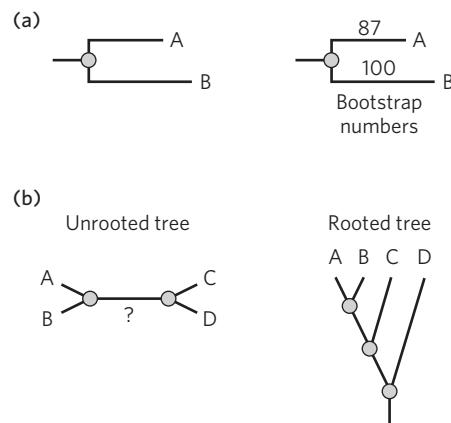


FIGURE 8-16 Time depictions and rootedness in phylogenetic trees. (a) The length of tree branches can be meaningless or, if specified, can represent some unit of evolutionary time. For example, the relative lengths of the branches correspond to differences in a time measure leading to taxa A and B. When numbers are given (right), these indicate the investigator's confidence in the information in that branch, based on statistical tests; these numbers are typical of the common bootstrap analysis. See text for details. (b) In an unrooted tree, some of the relationships between taxa may be evident, but there is uncertainty about the common ancestor of all the taxa. In a rooted tree, all taxa can be traced to a common ancestor.

set of sequences used to generate the original tree. For example, let's say a particular sequence of gene X is used to construct a tree for 100 species that all have gene X. A computer program randomly samples the original 100 sequences to create a new set of 100 sequences. However, in this new set, some of the original set may be missing and other sequences may be included multiple times. A new phylogenetic tree is generated from each of the created datasets. The number of times the same branch configuration arises for a cluster of species is tallied. The score reflects the absence (high confidence) or presence (lower confidence) of viable branching alternatives.

An unrooted tree is one for which the positioning of the common ancestor is uncertain (Figure 8-16b). In such a tree, the direction of evolution for parts of the tree might be unknown.

A wide range of problems arises in the construction of evolutionary trees. Mutation rates are often assumed to be constant, but this assumption is flawed. Mutation rates can be affected by environmental factors. For example, reactive oxygen species are the most common source of mutagenic DNA lesions (see Chapter 12). Thus, aerobic organisms are subject to more DNA dam-

age and potential mutagenesis than anaerobic organisms. Exposure to DNA-damaging agents such as UV light can vary greatly, depending on the ecological niche occupied by a given species. Certain regions of a gene may accommodate mutations better than others, depending on the functional importance of a given segment. An occasional back-mutation to the original base or amino acid could obscure the actual mutation rates. Finally, not all DNA in an organism is inherited linearly from parents to offspring. In individuals, genes can be lost, such as by genomic deletions due to DNA replication errors, and genes can be gained.

Gene gain can result from a process called **horizontal gene transfer**, which is common in bacteria and archaea (witness the spread of genes encoding antibiotic resistance in human bacterial pathogens). Early viruses may have transferred genes from one bacterium to another, and from one species to another, resulting in the sudden appearance (rather than the gradual evolution) of a gene in a particular evolutionary line. Large genome rearrangements might abruptly break up a pattern of synteny in one ancestral line, complicating the analysis. Sorting out these patterns is the job of increasingly sophisticated computer algorithms.

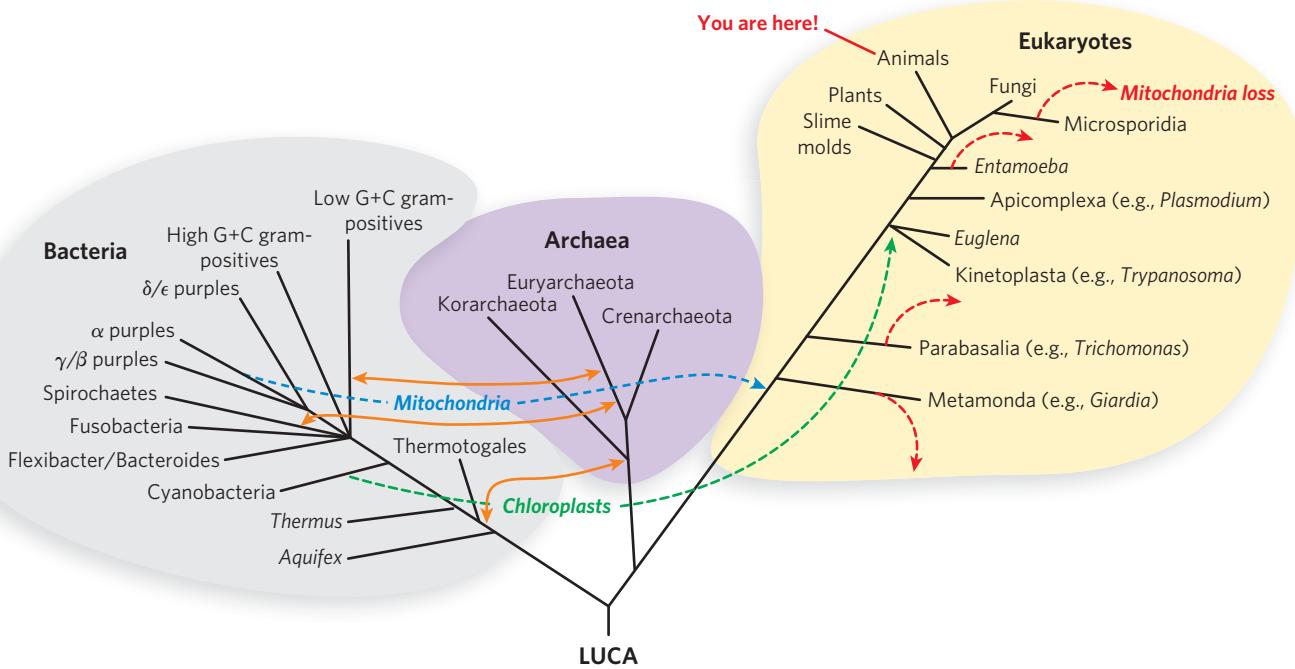


FIGURE 8-17 The tree of life. (a) This relatively simple tree includes only a few of the many sequenced genomes, but it illustrates some of the complexities of generating a complete tree of life. One crucial complicating factor is the tendency of living systems to occasionally incorporate DNA into their

genomes from other sources by horizontal gene transfer (orange arrows). Other factors are the assimilation of bacteria as organelles (mitochondria and chloroplasts; blue and green dashed arrows, respectively) and the subsequent loss of such organelles in some evolutionary lines (red dashed arrows).

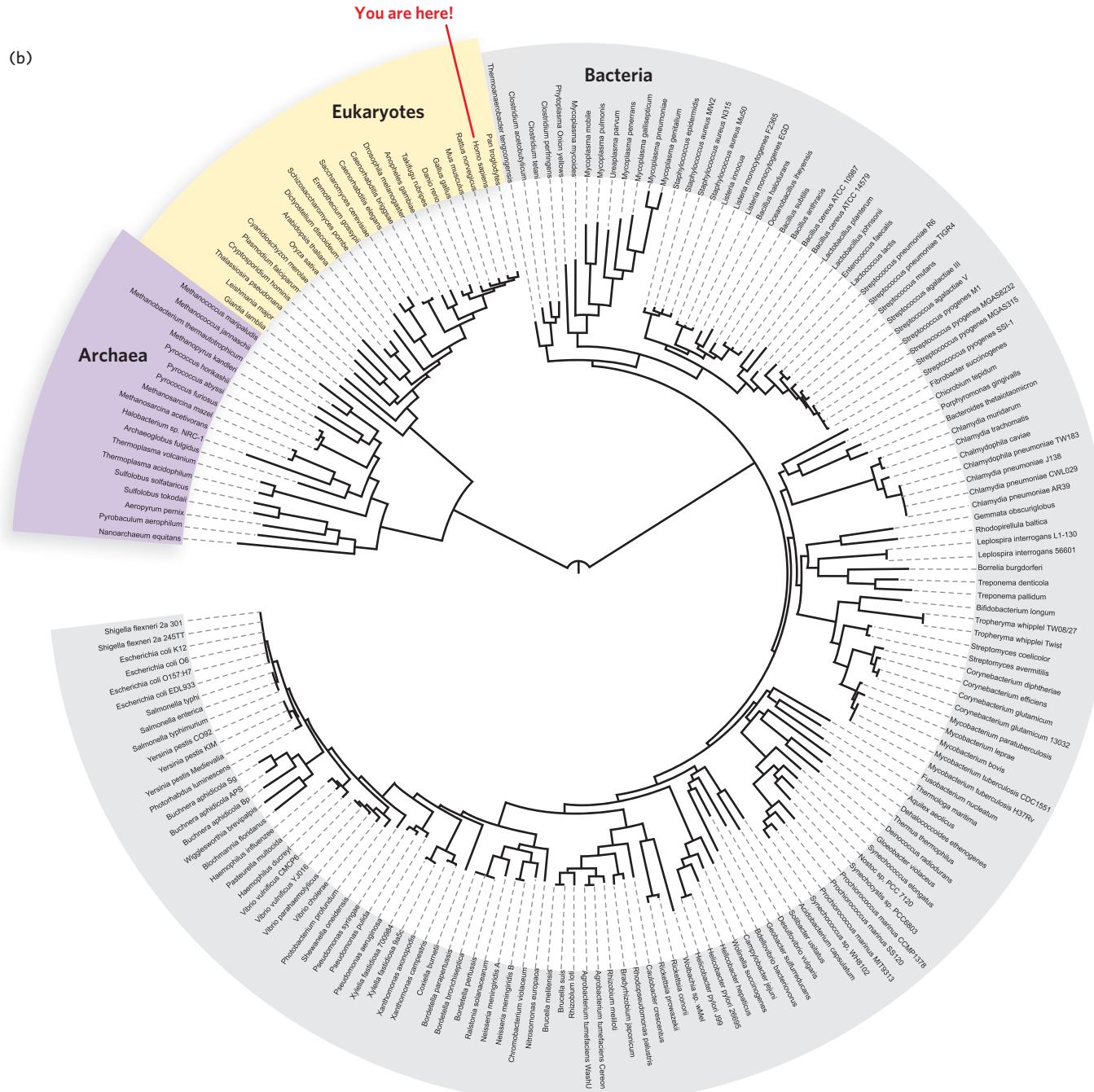


FIGURE 8-17 (continued) (b) A more complex evolutionary tree developed from 191 species with sequenced genomes. [Sources: (a) Adapted from J. R. Brown, "Universal tree

The complexity of the problem is evident in a current tree of life, two versions of which are shown in **Figure 8-17a and b**. These trees are based on analyses of particular genes and patterns in fully sequenced genomes. They are probably not correct in every detail. Corrections, additions, and updates will continue to be made for decades, perhaps centuries, to come.

of life," in *Encyclopedia of Life Sciences*, Wiley InterScience (online), 2005. (b) Adapted from F. D. Ciccarelli et al., *Science* 311:1283-1287, 2006.]

The Human Journey Began in Africa

Four major factors affect the evolution of any group of organisms. **Mutation rates** determine the extent of genetic diversity. **Natural selection** affects which genomic changes are inherited in a population. However, many mutations are relatively neutral and do not

undergo positive or negative selection. Neutral mutations are subject to a third evolutionary factor called **genetic drift**, in which the frequency of particular mutations in a population changes more or less randomly over time. Genetic drift is affected by such variables as the number of reproducing individuals in a population and the number of offspring generated. Finally, when groups of organisms colonize new regions and environments, their **migrations** may subject them to new and different selective pressures. These forces shaped the evolution of *Escherichia coli*; they also shaped the evolution of *Homo sapiens*.

About 7 million years ago, the common ancestor of chimpanzees and humans lived in Africa. Groups of that ancestral species followed divergent lines of evolution, one leading to chimpanzees and bonobos, and one leading to humans (Figure 8.18). The path to humans first generated a series of species in a genus dubbed *Australopithecus*. The Australopithecines remained in Africa, giving rise, about 3 million years ago, to *Homo habilis*, the first species of our own genus. The archaeological record indicates that *H. habilis* was the first species to use stone tools. About 1.7 million years ago, a successor to *H. habilis* emerged—*Homo erectus*. The hominids were a bit more adventurous than the Australopithecines. Armed with better tools and a mastery of fire, *H. erectus* spread from Africa to virtually all of Eurasia. The fossil record provides evidence of many other *Australopithecus* and *Homo* species during the past 3 million years. These species probably arose by **allopatric speciation**: geographic isolation of a group of individuals followed by evolution to form a distinct species that no longer can interbreed with the original one. All of these species ultimately became extinct, except for one.

Homo sapiens evolved about 500,000 years ago. The history of our species is written in our DNA. For decades, scientists argued about two possible human origins. The multiregional theory proposed that humans evolved gradually in many places, with gene flow occurring constantly between the various populations. This would entail a direct evolution of *H. erectus* into *H. neanderthalensis* into *H. sapiens* occurring simultaneously in Eurasia and Africa. The alternative, “out of Africa” theory posits that the *H. erectus* and *H. neanderthalensis* expansions into Eurasia were independent of the *H. sapiens* expansion, and that the former two species represented separate evolutionary branches. Modern genomics has definitively resolved the debate in favor of the “out of Africa” theory.

A woman who lived in sub-Saharan Africa 140,000 to 200,000 years ago, sometimes called Eve by the

scientists who study our ancestral tree, is the female genetic ancestor of all living humans. She was not the only human female then living, but she is the only one whose DNA has been inherited in the modern human lineage. All mitochondrial DNA is inherited maternally, deposited in the egg prior to fertilization. Mitochondrial DNA is also not subject to the scrambling effects of recombination. Thus, stable haplotypes of mitochondrial genome polymorphisms can be reliably traced back in time. The current human lineage traces back to mitochondrial Eve.

There is also a genomic Adam, but Eve never knew him. All males now living on Earth are descended from a male who lived in Africa about 60,000 years ago. Again, Adam was not the only male member of his species present. He is simply the one whose DNA survives. Our information about this individual comes from analyses of haplotypes in Y chromosome DNA, most of which is not subject to recombination.

Human Migrations Are Recorded in Haplotypes

About 50,000 years ago, a small group of humans looked out across the Red Sea to Asia. Perhaps encouraged by some innovation in small boat construction, or driven by conflict or famine, or simply curious, they crossed the water barrier. That initial colonization, involving maybe 1,000 individuals, began a journey that did not stop until humans reached Tierra del Fuego (at the southern tip of South America) many thousands of years later. In the process, the established population from a previous hominid expansion into Eurasia, including *H. neanderthalensis*, was displaced. The Neanderthals disappeared, just as *H. erectus* had disappeared before them.

This journey can be traced by looking at our genomic polymorphisms. Efforts are under way to survey genetic diversity in human populations around the globe. One such endeavor is the International HapMap (haplotype map) project; another is the Human Genome Diversity Project. Both are international efforts to sequence thousands of human genomes taken from carefully selected populations around the world and to accumulate information about tens of thousands of genome polymorphisms in the sequenced individuals. These enterprises are every bit as large and complex as the original Human Genome Project. The results have helped define the mitochondrial Eve and the Y chromosome Adam, and they are telling us a great deal more. Complementary analyses of mitochondrial DNA from Neanderthals have established that they were on a separate evolutionary line.

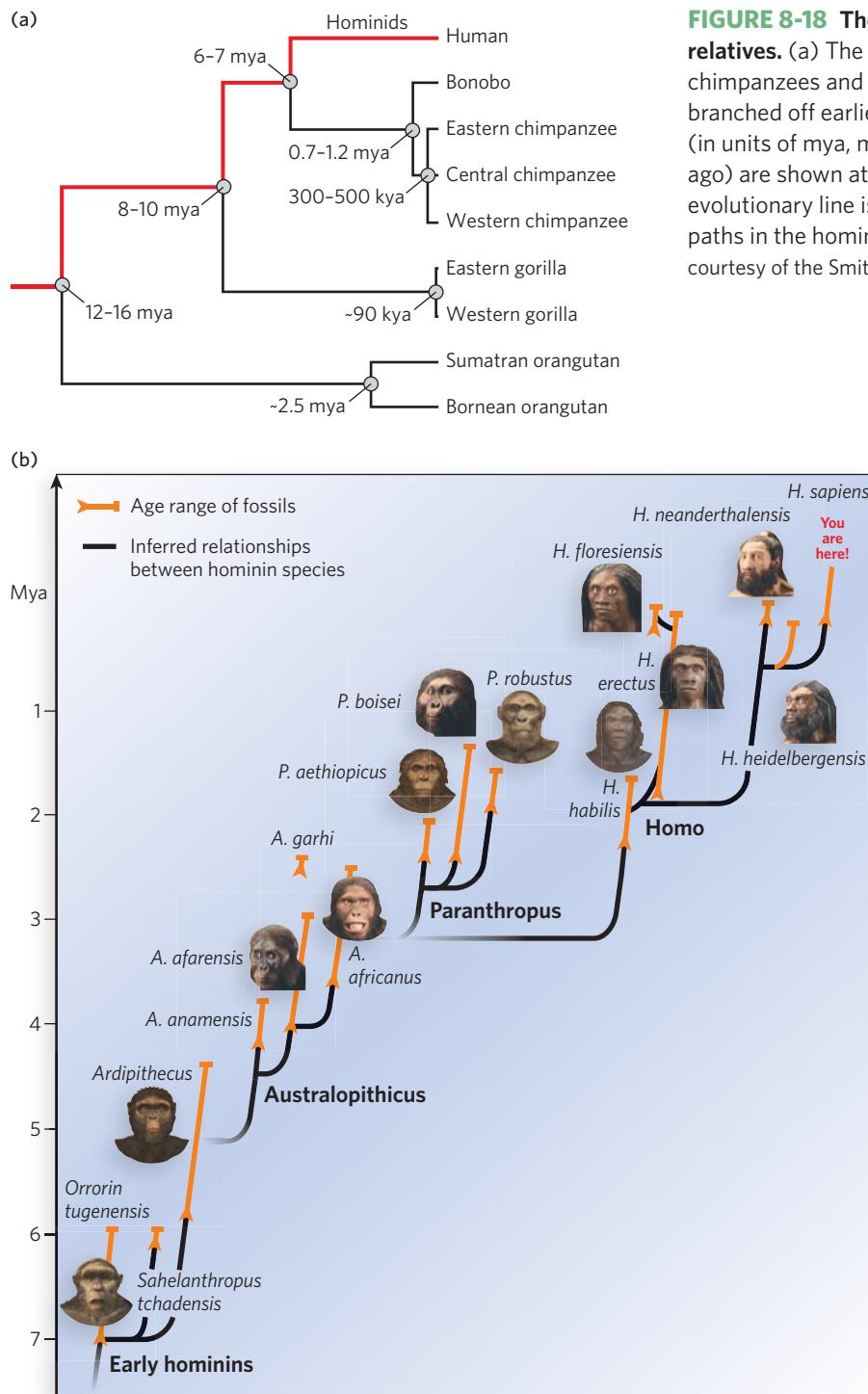


FIGURE 8-18 The evolution of humans and their close relatives. (a) The closest living relatives of humans are chimpanzees and bonobos. The orangutan and gorilla lines branched off earlier. Estimated times of species divergence (in units of mya, million years ago, and kya, thousand years ago) are shown at the branch points. The hominid evolutionary line is in red. (b) Details of the evolutionary paths in the hominid line. [Source: (b) Reconstruction photos courtesy of the Smithsonian Institution.]

Phylogenetic analyses of species evolution generally rely on gene mutations that are fixed in a given species; that is, all the members of species X have one gene sequence, and all the members of species Y have a different sequence. The analysis of genetic polymorphisms within a species increasingly relies on a different kind of mathematical analysis, called **coalescent**

theory. Even though it is not subject to recombination, the sequence of mitochondrial DNA, like that of chromosomal DNA, changes slowly with time because of mutations. If a mutation occurred recently, it will appear in the relatively few individuals descended from that female. If the mutation appeared much earlier, it is found in many individuals across broad

geographic regions. With mathematical models that take into account estimated mutation rates, selection, genetic drift, and other factors, various polymorphisms are traced back to the ancestor in which they first appeared—a coalescence.

Overall genetic diversity in the human lineage is lower than that in chimpanzees. This is one of several pieces of evidence indicating that early human populations went through evolutionary bottlenecks a few hundred thousand years ago, when only a few thousand or tens of thousands of individuals existed. Our mitochondrial Eve and Y chromosome Adam lived in times when there were far fewer humans than today. More than 85% of the polymorphisms in the human population appear at the same frequency in all human populations worldwide, indicating that they arose before the appearance of the first modern humans. The remaining 15% tell us about human migrations.

Genetic diversity, in terms of haplotypes that do not occur uniformly across world populations, is by far the greatest in extant African populations. When that wanderlust-possessing population of humans first colonized Asia, they took only a subset of the variable human haplotypes with them. That first colony expanded in population size, and at some point, additional migrations led to new colonies farther away. The new colonies would reflect a subset of the haplotypes present in the previous colony, but sometimes (every few thousand or tens of thousands of years) a new col-

ony would pick up a new haplotype, due to a random new mutation, that would spread exclusively in that group (a founder event). As humans dispersed across the planet, the farthest spread (into the Americas) is characterized by the lowest overall haplotype diversity, while at the same time being marked by a few unique haplotypes picked up relatively late in the migratory process. This prevalence of key haplotypes in various populations lets us trace the path of human migrations (Figure 8-19).

Of course, the same methods can be used to analyze the history of any species, from viruses to mammals. For example, these methods allow the tracing of viral evolution associated with human pandemics and reveal the types of mutational events that occurred in the past and are therefore possible in the future. Analysis of the genomic history of maize or rice could reveal lost genetic diversity in common production strains that might prove useful to agriculture.

The ongoing analysis of worldwide human genetic diversity, increasingly enriched by the completed genome sequencing of thousands of individuals and new methods that incorporate information about haplotypes throughout the genome, will yield increasingly detailed information about human history. It will also aid the search for mutations that contribute to genetic diseases, some of which affect only certain populations. Finally, it will help pinpoint unique changes in specific populations that signal subtle adaptations to the local

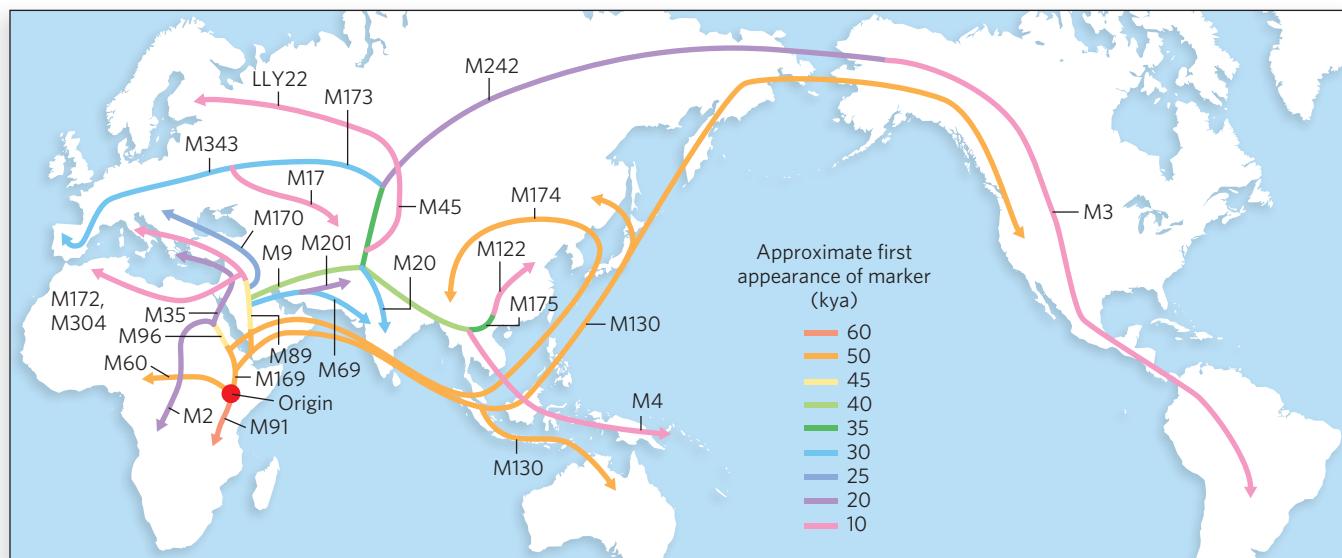


FIGURE 8-19 The paths of human migrations. This map was generated from an analysis of genetic markers (defined haplotypes with M or LLY numbers) on the Y chromosome. Genetic markers that reflect changes

appearing in certain isolated populations (in “founder events”) enable researchers to trace migrations from that point. [Source: Adapted from G. Stix, *Sci. Am.* 299(July):56–63, 2008.]

environment, a hallmark of ongoing evolution and a human journey that continues today.

SECTION 8.3 SUMMARY

- Genome studies have facilitated new approaches to defining the last universal common ancestor, LUCA. Genomics is used to identify genes that are common to all extant life forms and were therefore likely to exist in LUCA.
- Perhaps the ultimate challenge of genomics is to define the tree of life, a detailed description of the evolutionary history and relationships of all the species now living on Earth. An evolutionary tree is referred to as a phylogeny, and phylogenetics is the study of evolutionary relationships.
- The study of mitochondrial DNA, Y chromosome DNA, and genetic haplotypes in the human population enables geneticists to trace human evolution and more recent human migrations.

Unanswered Questions

Genomics, transcriptomics, and proteomics are disciplines designed to obtain large amounts of information about an organism and the systems within it. The list of accomplishments is long, but the list of questions is even longer. There is a rich but largely unpredictable future in these fields.

1. **How many classes of RNAs exist in cells, and how do we find the corresponding genes?** The discovery of new classes of RNA is a rapidly evolving area of research. Among these are RNAs involved in all kinds of processes, most notably regulation. The capacity of some RNAs to affect the expression of protein-coding genes could make them prominent targets of evolution. Some of this research effort is described in Chapter 22.
2. **Are there additional domains of living organisms?** The 1977 discovery of the archaea surprised many researchers. These microorganisms are now firmly established as a distinct branch of the evolutionary

tree, separate from the eukaryotes and bacteria. Is there another domain we have missed, or even more than one? Massive efforts to sequence genomes, including work in metagenomics, have produced persistent rumors that another domain of life may be described in the near future. There are certainly sufficient numbers of unusual life-sustaining niches on Earth to make this plausible.

3. What are the most likely characteristics of LUCA?

Ongoing endeavors to define the complete tree of life will gradually refine our understanding of how biological evolution proceeded. Aided by new sequence information, increased knowledge of mutagenesis and nonlinear events that contribute to evolution (horizontal gene transfer and transposon introduction), and complementary data from other fields (including more precise dating methods), we can expect a detailed, consensus tree of life in the decades ahead. A parallel effort to better define the processes common to living systems and those that must have been present in LUCA may provide us with a look at our deepest biological past.

4. How do we investigate interdependent microbial communities?

The new field of metagenomics is starting to tackle questions about the diversity of unculturable bacterial species in environments such as the digestive tract of termites. New approaches will be needed to generate complete genome sequences of the many members of such communities and to analyze the genomes for clues about why individual species cannot survive without the others present.

5. What evolutionary innovations define humans as a species?

Among the many subtle differences we can see between the human genome and other primate genomes are mutations of many kinds that hold the key to our capacity for problem solving, language development, and other human traits. Understanding these will advance our understanding of medicine and neurochemistry in myriad ways, some of which we cannot predict.

How We Know

Haemophilus influenzae Ushers in the Era of Genome Sequences

Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty, J.M. Merrick, et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269:496–512.

The first genome sequencing projects, in the early 1990s, used this strategy: clone, map carefully, and then sequence. Craig Venter, who had recently established The Institute for Genome Research (TIGR), was eager to test his idea that by using new computational methods one could skip the time-consuming mapping steps. The result was the first complete sequence of a free-living organism—the bacterium *Haemophilus influenzae*.

H. influenzae was first described by Richard Pfeiffer in 1892 during an influenza outbreak. Until 1933, this bacterium was incorrectly thought to be the cause of the common flu. There are six types of *H. influenzae* (designated a through f) that can be immunologically distinguished by differences in their polysaccharide coat or capsule, and many other unencapsulated types. This bacterium is an opportunistic human pathogen that lives in tissues and rarely causes disease. However, *H. influenzae* type B is responsible for acute bacterial meningitis and bacteremia, primarily in children. The organism chosen for analysis was *Haemophilus influenzae* Rd, a well-characterized type d strain often used in laboratory studies. For the purposes of Venter and his associates at TIGR, the bacterium had several advantages. As a human pathogen it was a good target for genome sequencing. Its genome, known to be about 1.8×10^6 bp, is large, yet smaller than the genomes of other sequencing targets in use at the time. Its genomic 35% G + C content is close to that of humans, making it a good subject for developing methods for the Human Genome Project. Most important, no physical clone map existed for the *H. influenzae* genome. If the work succeeded, there would be no question that it was a victory for the overall strategy Venter had in mind.

The DNA isolated from *H. influenzae* was sheared mechanically and size-fractionated to yield random fragments of 1,600 to 2,000 bp. The fragments were cloned into a plasmid, and a library of the clones was constructed. The clones were sequenced at random. Then, 19,346 separate “forward sequence” reactions (entering the clone from the same end relative to the vector in which it was cloned) were carried out, with an 84% success rate. Just over half of the same set of clones were also sequenced in the reverse direction.

The average length of DNA in each sequencing read was about 460 bp. The end result was 11,631,485 bp of DNA sequence in the random assembly.

Next, the computer algorithm tackled the immense job of assembling the genome by building a table of all 10 bp oligonucleotide subsequences and using the table to generate a list of potential fragment overlaps. With a single DNA fragment beginning the assembly of a contig, candidate overlap fragments were chosen and tested for more extended matches by strict criteria. Gradually, overlapping fragments were pieced together to generate a genome sequence. Assembling the 24,304 fragments required 30 hours of computer time. When the assembly was complete, the fragments had been ordered into 42 contigs, with 42 gaps in the genome and little information about how to order the contigs. However, many of the gaps were short. Sometimes, a contig end fell within the same single gene as another contig end, both being identified by virtue of existing peptide sequences for the gene in question. Additional libraries containing long clones of *H. influenza* DNA were probed with sequences near contig ends, to identify ends that were near each other. The gaps were closed by this method and by other targeted sequencing efforts.

The genomic sequences of the bacterium were sequenced, by the end, with more than sixfold redundancy. The final error rate was estimated at 1 in 5,000 to 10,000 bp. At \$0.48 per finished base pair, the total cost was just under \$900,000. Newer sequencing technologies, such as those described in Highlight 7-2, have lowered the cost of sequencing a typical bacterial genome by almost three orders of magnitude.

The result of Venter and colleagues’ endeavor was a complete genome sequence with 1,830,137 bp, published in July 1995. The genome included 736 predicted genes, over half of which had no known function at that time. More importantly, the effort inspired a new generation of genome analysts. TIGR is now the J. Craig Venter Institute, and it remains a major force in genome sequencing. The shotgun sequencing approach successfully pioneered in *H. influenzae* is now routinely paired with the new sequencing technologies to provide rapid and inexpensive genome assemblies.

Key Terms

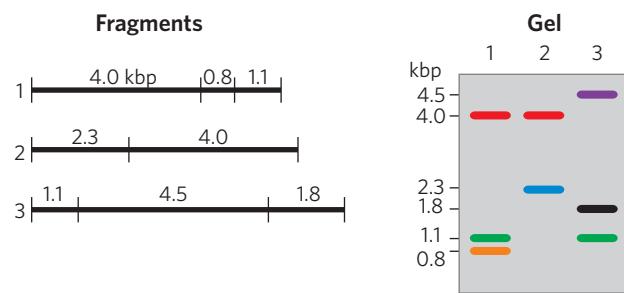
genome, p. 260
 genomics, p. 260
 contig, p. 261
 sequence tagged site (STS), p. 261
 expressed sequence tag (EST), p. 261
 whole-genome shotgun sequencing, p. 262
 genome annotation, p. 263
 homolog, p. 265
 ortholog, p. 265
 paralog, p. 265

synteny, p. 266
 metagenomics, p. 267
 intervening sequence (intron), p. 269
 exon, p. 269
 simple-sequence repeat (SSR), p. 271
 single nucleotide polymorphism (SNP), p. 271
 haplotype, p. 271
 outgroup, p. 273
 linkage analysis, p. 274

systems biology, p. 277
 transcriptome, p. 277
 transcriptomics, p. 277
 proteome, p. 278
 proteomics, p. 278
 last universal common ancestor (LUCA), p. 283
 phylogeny, p. 283
 phylogenetics, p. 283
 taxon, p. 285
 horizontal gene transfer, p. 286
 allopatric speciation, p. 288

Problems

- 1.** Three different but overlapping BAC clones (see Chapter 7) were digested with the restriction enzyme EcoRI and the fragments separated on an agarose gel, as shown in the figure below. Only the cloned DNA (not the plasmid vector) is shown. Order these three clones into a contig, and label the contig with the location of the EcoRI restriction sites and the distances between them.



- 2.** A researcher compares the amino acid sequences of cytochrome *c* from four vertebrates: sheep, dog, rabbit, and kangaroo. The amino acid differences among the four species are shown in the difference matrix below. From this information, build a simple evolutionary tree expressing the apparent relationships among these organisms. The branch lengths do not have to represent any measure of time.

Difference Matrix

	Sheep	Dog	Rabbit	Kangaroo
Sheep	0			
Dog	3	0		
Rabbit	4	5	0	
Kangaroo	6	7	6	0

- 3.** In random shotgun sequencing, cloned genomic DNA from an organism is sequenced at random. Sequencing requires the use of a primer targeted to a known se-

quence, which can then be extended to reveal the entire sequence by the traditional Sanger method (see Chapter 7). If the researcher has no sequence information, how can any genome sequences be targeted by primers to initiate the sequencing reactions?

- 4.** A hypothetical protein is found in orangutans, chimpanzees, and humans that has the following sequences (red indicates the amino acid residue differences; dashes indicate a deletion—the residues are missing in that sequence):

Human: ATSAAGYDEWEGGKVLIHL--KLQNRGALLELDIGAV
Orangutan: ATSAAGWDEWEGGKVLIHLDGKLQNRGALLELDIGAV
Chimpanzee: ATSAAGWDEWEGGKILIHLDGKLQNRGALLELDIGAV

What is the most likely sequence of the protein present in the last common ancestor of chimpanzee and human?

- 5.** For defining a cell's transcriptome, RNA-Seq provides an alternative to microarrays. In this method, cellular RNA is isolated, transcribed to complementary DNA, and sequenced.

- (a) How does the sequencing yield information about the levels of specific RNAs in a cell?
 (b) Why must rRNA be removed from most samples before conversion of the cellular RNA to cDNA?

- 6.** A comparison between two homologous chromosomes in two closely related mammals reveals synteny over most of the length of the chromosomes. However, researchers encounter a segment of about 2,300 bp in mammal Y that is not present in mammal X. What evolutionary processes could account for this difference?

- 7.** In large genome sequencing projects, the initial data usually reveal gaps where no sequence information has been obtained. To close the gaps, DNA primers complementary to the 5'-ending strand (i.e., identical to the sequence of the 3'-ending strand) at the end of each contig are especially useful. Explain how these primers might be used.

- 8.** In proteomics work, two-dimensional gel electrophoresis is often used to separate the thousands of proteins in a cell on a single gel. Proteins are separated by their charge (*pI*) in one dimension, and then by size in the second dimension. Why do researchers use two different electrophoretic procedures, instead of simply separating proteins by size in both dimensions?
- 9.** You are a gene hunter, trying to find the genetic basis for a rare inherited disease. Examination of six pedigrees of families affected by the disease provides inconsistent results. For two of the families, the disease is co-inherited with markers on chromosome 7. For the other four families, the disease is co-inherited with markers on chromosome 12. Explain how this might occur.
- 10.** In two closely related bacterial species, a cluster of five genes is found on the chromosome, with the genes arranged in the same order. However, in species B, gene X, not present in species A, is found between genes 2 and 3 of the five-gene sequence. If gene X has no sequence homology to any other gene in species B, how might it have arisen? If the nucleotide sequence of gene X is

72% identical to gene 2 in species B, how might it have arisen?

- 11.** Native American populations in North and South America have mitochondrial DNA haplotypes that can be traced to populations in northeast Asia. The Aleut and Eskimo populations in the far northern parts of North America possess a subset of the same haplotypes that link other native Americans to Asia, and have several additional haplotypes that can also be traced to Asian origins but are not found in native populations in other parts of the Americas. Provide a possible explanation.
- 12.** In the evolutionary line leading to modern humans, bottleneck periods occurred in which relatively few individuals survived. All humans living today carry mitochondrial DNA markers derived from a single female, sometimes referred to as mitochondrial Eve. All male humans living today have Y chromosome markers from an ancestor called Y chromosome Adam. Did Y chromosome Adam carry mitochondrial DNA from mitochondrial Eve?

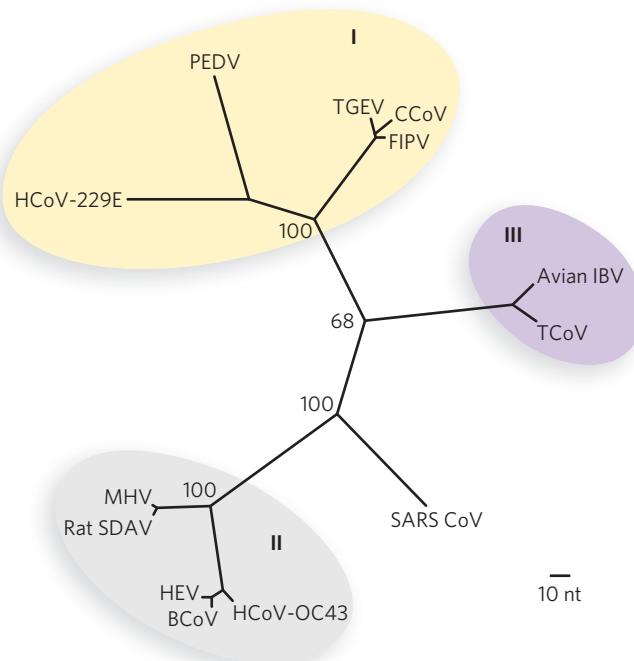
Data Analysis Problem

Ksiazek, T.G., D.V.M. Ksiazek, D. Erdman, C.S. Goldsmith, S.R. Zaki, T. Peret, S. Emery, S. Tong, C. Urbani, J.A. Comer, et al. 2003. A novel coronavirus associated with severe acute respiratory syndrome. *N. Engl. J. Med.* 348:1953–1966.

13. T. G. Ksiazek and other members of the SARS Working Group described their discovery of the SARS virus and its identification as a novel coronavirus. They identified the virus as a coronavirus through electron microscopy, and confirmed it with a sequence analysis of PCR-amplified genomic segments from SARS patient samples.

- (a) The PCR method for amplifying the coronavirus genomic sample requires the use of reverse transcriptase. Why?
- (b) To amplify sequences from a new virus with an unsequenced genome, what considerations would go into the design of appropriate PCR primers?
- (c) A sequence alignment involving a 405 bp segment of the DNA polymerase gene in four coronaviruses, BCoV, HEV, SARS CoV, and TGEV, is given on the facing page—the alignment that Ksiazek et al. used to generate the phylogenetic tree shown here (this includes coronaviruses not discussed in this problem; nt = nucleotides). In the alignment, nucleotide polymorphisms relative to the bovine coronavirus sequence (bcov) are shown in red. How many sequence differences exist between the genome segments from HEV (hev) and BCoV? How many between BCoV and SARS CoV (sars)? How many between TGEV (tgev)

and SARS CoV? Which two coronaviruses are most closely related? Are your counts in general agreement with the phylogenetic tree?



	1	60
bcov.pol.seq	TCGTGCTATGCCAACATACTACGTATTGTTAGTAGTCTGGTTGGCTCGAAAACATGA	
hev.pol.seq	TCGTGCTATGCCAACATACTACGTATTGTTAGTAGTCTGGTATTGGCCCGAAAACATGA	
sars.pol.seq	CAGAGCCATGCCTAACATGCTTAGGATAATGGCCTCTCTGGCTTGCTCGCAAACATAA	
tgev.pol.seq	CCGTGCTTACCTAATATGATTAGAATGGCTCTGCCATGATATTAGTTCTAAGCATGT	
	61	120
bcov.pol.seq	GGCATGTTGTCGCAAAGCGATAGGTTTATCGACTTGCATGAATGCGCACAAAGTTCT	
hev.pol.seq	GGCATGTTGTCGCAAAGCGATAGGTTTATCGACTTGCATGAATGCGCACAAAGTTCT	
sars.pol.seq	CACTTGCTGTAACTTATCACACCGTTCTACAGGTTAGCTAACGAGTGTGCGCAAGTATT	
tgev.pol.seq	TGGTTGTTGTACACATAATGATAGGTTCTACCGCCTCTCCAATGAGTTAGCTCAAGTACT	
	121	180
bcov.pol.seq	GAGTGAAATTGTTATGTGTTGGCTGTTATTATGTTAACGCTGGTGGCACTAGTAGTGG	
hev.pol.seq	TAGTGAAATTGTTATGTGTTGGCTGTTATTATGTTAACGCTGGTGGCACTAGTAGTGG	
sars.pol.seq	AAGTGAGATGGTCATGTGTTGGCGGCTCACTATATGTTAACCAAGGTGGAACATCATCCGG	
tgev.pol.seq	CACAGAAAGTTGTGCATTGCACAGGTGGTTTTATTAAACCTGGTGGTACAACTAGCGG	
	182	240
bcov.pol.seq	TGATGCAACTACTGCTTTGCTAACATTGAGTTAACATATGTCAGCTGTTCAGCAA	
hev.pol.seq	TGATGCAACTACTGCTTTGCTAACATTGAGCTTAAACATATGTCAGCTGTTCAGCAA	
sars.pol.seq	TGATGCTACAACTGCTTATGCTAATAGTGTCAAGCTGTTAACATTTGTCAGCTGTTAAC	
tgev.pol.seq	TGATGGTACTACAGCATATGCTAACACTCTGCTTTAACATCTTCAAGCTGTTCTGCTAA	
	241	300
bcov.pol.seq	TGTATGTGCTTTAATGTCATGCAATGTAATAAGATTGAAGATTGAGTATACGTGCTCT	
hev.pol.seq	TGTATGTT CCCTTAATGTCATGCAATGGCAATAAGATTGAAGATTGAGTATACGTGCTCT	
sars.pol.seq	TGTAATGCACCTCTTCAACTGATGTAATAAGATAGCTGACAAAGTATGTCCGCAATCT	
tgev.pol.seq	TGTTAAATAAGCTTTGGGGTTGATTCAAACGCTTGTAAACACGTTACAGTAAAATCCAT	
	302	360
bcov.pol.seq	TCAGAACGCCTTATACTCACATGTTAGTAGAAGTGTATGGTTGATTCAACCTTGTAC	
hev.pol.seq	TCAGAACGCCTTATACTCACATGTTAGTAGAAGTGTATGGTTGATTCAACCTTGTAC	
sars.pol.seq	ACAACACAGGCTCTATGAGTGTCTCTATAGAAATAGGGATGTTGATCATGAAATTCGTGGA	
tgev.pol.seq	ACAACGTTAAATTACGATAATTGTTACGTTAGCTAGTACAGTAAATCCAT	
	361	406
bcov.pol.seq	AGAATATTATGAAATTAAATAAGCATTAGTATGATGATTG	
hev.pol.seq	AGAATATTATGAAATTAAATAAGCATTAGTATGATGATTG	
sars.pol.seq	TGAGTTTACGCTTACCTGCGTAAACATTCTCCATGATGATTCTT	
tgev.pol.seq	TGAGTACTTTAGTTATTGAGAAAAACACTTTCTATGATGATTAA	

[Sequence alignment courtesy of Dr. Ann Palmenberg, Department of Biochemistry and the Institute for Molecular Virology, University of Wisconsin-Madison.]

Additional Reading

Genomes and Genomics

- Chen, K., and L. Pachter.** 2005. Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Comput. Biol.* 1:e24.
- Chimpanzee Sequencing and Analysis Consortium.** 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437:69–87.
- Fitch, W.M.** 1970. Distinguishing homologous from analogous proteins. *Syst. Zool.* 19:99–114. The paper that introduced the concept of orthologs and paralogs.
- Giallourakis, C., C. Henson, M. Reich, X. Xie, and V.K. Mootha.** 2005. Disease gene discovery through integrative genomics. *Annu. Rev. Genomics Hum. Genet.* 6:381–406. A good summary of how these important searches are done.
- Gibbs, J.R., and A. Singleton.** 2006. Application of genome-wide single nucleotide polymorphism typing: Simple association and beyond. *PLoS Genet.* 2:e150.
- Griffiths-Jones, S.** 2007. Annotating noncoding RNA genes. *Annu. Rev. Genomics Hum. Genet.* 8:279–298.
- International Human Genome Sequencing Consortium.** 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431:931–945. Moving from the draft sequences to the completed sequences.
- Koonin, E.V.** 2005. Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.* 39:309–338.
- Mardis, E.R.** 2008. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* 9:387–402.
- Marques-Bonet, T., O.A. Ryder, and E.E. Eichler.** 2009. Sequencing primate genomes: What have we learned? *Annu. Rev. Genomics Hum. Genet.* 10:355–386.
- Noonan, J.P., G. Coop, S. Kudaravalli, D. Smith, J. Krause, J. Alessi, F. Chen, D. Platt, S. Pääbo, J.K. Pritchard, and E.M. Rubin.** 2006. Sequencing and analysis of the Neanderthal genomic DNA. *Science* 314:1113–1118. We are slowly generating a detailed genomic understanding of our closest evolutionary relative.
- Rasmussen, M., Y. Li, S. Lindgreen, J.S. Pedersen, A. Albrechtsen, I. Moltke, M. Metspalu, E. Metspalu, T. Kivisild, R. Gupta, et al.** 2010. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* 463:757–762.
- Rusch, D.B., A.L. Halpern, G. Sutton, K.B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J.A. Eisen, J.M. Hoffman, K. Remington, et al.** 2007. The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biol.* 5:e77. Every molecular biologist would enjoy this kind of work.
- Sikela, J.M.** 2006. The jewels of our genome: The search for the genomic changes underlying the evolutionarily unique capacities of the human brain. *PLoS Genet.* 2:e80. A favorite subject for every human.

Transcriptomes and Proteomes

- Andersen, J.S., and M. Mann.** 2006. Organellar proteomics: Turning inventories into insights. *EMBO Rep.* 7:874–879.
- Kim, T.H., and B. Ren.** 2006. Genome-wide analysis of protein-DNA interactions. *Annu. Rev. Genomics Hum. Genet.* 7:81–102.
- Nie, L., G. Wu, D.E. Culley, J.C. Scholten, and W. Zhang.** 2007. Integrative analysis of transcriptomic and proteomic data: Challenges, solutions and applications. *Crit. Rev. Biotechnol.* 27:63–75.
- Wang, Z., M. Gerstein, and M. Snyder.** 2009. RNA Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10:57–63.

Our Genetic History

- Cavalli-Sforza, L.L.** 2007. Human evolution and its relevance for genetic epidemiology. *Annu. Rev. Genomics Hum. Genet.* 8:1–15.
- Ciccarelli, F.D., T. Doerks, C. von Mering, C.J. Creevey, B. Snel, and P. Bork.** 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science* 311: 1283–1287. Building an evolutionary tree based on sequences of 36 genes in 191 species, with methods that eliminate the effects of horizontal gene transfer.
- Forster, A.C., and G.M. Church.** 2006. Towards synthesis of a minimal cell. *Mol. Syst. Biol.* 2:45. An update on efforts to synthesize a living system from chemical components.
- Galperin, M.Y.** 2006. The minimal genome keeps growing. *Environ. Microbiol.* 8:569–573. A description of a project to define the minimum number of genes required for independent life.
- Koonin, E.V., and W. Martin.** 2005. On the origin of genomes and cells within inorganic compartments. *Trends Genet.* 21:647–654.
- Li, R., Y. Li, H. Zheng, R. Luo, H. Zhu, Q. Li, W. Qian, Y. Ren, G. Tian, J. Li, et al.** 2010. Building the sequence map of the human pan-genome. *Nat. Biotechnol.* 28:57–63.
- Pakendorf, B., and M. Stoneking.** 2005. Mitochondrial DNA and human evolution. *Annu. Rev. Genomics Hum. Genet.* 6:165–183.
- Rannala, B., and Z. Yang.** 2008. Phylogenetic inference using whole genomes. *Annu. Rev. Genomics Hum. Genet.* 9:217–231.
- Serre, D., and T.J. Hudson.** 2006. Resources for genetic variation studies. *Annu. Rev. Genomics Hum. Genet.* 7:443–457.

Topology: Functional Deformations of DNA



Carlos Bustamante [Source: Courtesy of Carlos Bustamante.]

Moment of Discovery

There was an experiment I had wanted to do for many years, but I could never convince anyone to try it. *The idea was to measure the elastic properties of DNA directly using a single molecule of DNA tethered between two opposing pipette tips, such that it could be exquisitely controlled by rotating one pipette tip relative to the other to twist the DNA to different extents.* In solution, DNA can twist on its long axis, and it can also *writhe* by coiling around itself—and it can be very hard

to decouple twisting from writhing by measuring properties of DNA in bulk.

After many students turned me down, eventually two students, Zev Bryant and Michael Stone, were intrigued to try the DNA twisting experiment. These guys worked very hard setting up the technical aspects of the experiment. They figured out how to tether the ends of a nicked fragment of DNA to two opposing pipette tips, and they coupled a readily visualized small bead (the “rotor”) to an internal position in the nicked DNA duplex. They finally got everything working late one night. Zev and Michael started using a hand crank to introduce a specific number of twists into the tethered piece of DNA. But they were so tired that they would get up to, say, 345 turns, and then they weren’t sure if it was 345 or 346! They were determined to do the experiment accurately, so they had to untwist the DNA and start over. But Jan Liphardt, another student in the lab, had an idea. He offered to bring in a small motor from his Lego set at home to rotate the pipette tip, and thus twist the DNA, automatically. So Jan ran home and got the motor, hooked it up to the system, and it worked beautifully. All the data we ultimately published were measured using the Lego motor (we even listed it in the Methods section of the paper). And we learned that DNA is about 50% stiffer than had been previously estimated from bulk solution experiments!

—**Carlos Bustamante**, on discovering the elasticity of DNA

9.1 The Problem: Large DNAs in Small Packages 298

9.2 DNA Supercoiling 305

9.3 The Enzymes That Promote DNA Compaction 312

In all free-living organisms—all bacteria, archaea, and eukaryotes—genomic information comes in the form of DNA; RNA genomes occur only in some classes of viruses. Genomic nucleic acids are large. In fact, they are *much* (orders of magnitude) longer than the biological packages—cells, organelles, or virus particles—that contain them. Most human cells are 7 to 30 μm (micrometers, also called microns) in diameter. The nuclei that house the DNA molecules are rarely more than 10 μm in diameter. The shortest human chromosome (chromosome 21), at just under 47 million base pairs, would be about 16 mm long if stretched out in a line, or more than a thousand times longer than the nucleus. If all of the chromosomes in a diploid human cell were laid end to end, they would stretch for nearly 2 m. It quickly becomes apparent, then, that genomic DNA must be compacted to fit inside the cell.

The compaction (also called condensation) of DNA is extensive, but never random. The genome is restricted to a limited cellular space, but the cell must retain access to the information contained within the DNA. The enzymes that carry out DNA replication, repair, recombination, and transcription must all have ready access to their sites of action. Regulatory proteins must have access to the specific sequences to which they bind. In all cells, a range of DNA-binding proteins contributes to compaction and to regulation of the function of chromosomal DNA. The compaction is both orderly and dynamic.

As we'll see, DNA is compacted primarily by coiling. The process is somewhat analogous to rolling up a length of garden hose or a spool of electrical wire. However, the coiling of DNA occurs in the context of structural constraints peculiar to this nucleic acid, constraints that are dealt with by a unique set of proteins and enzymes.

In this chapter, we shift our focus from the secondary structure of DNA (see Chapter 6) to the extraordinary degree of organization required for the tertiary packaging of DNA into chromosomes. We explore the principles related to the compaction of DNA, beginning with a review of the structural elements that make up viral and cellular chromosomes, and consider chromosomal size and organization in more detail. We then discuss DNA topology for a quantitative description of the coiling and supercoiling of DNA molecules. We conclude with a discussion of the key enzymes, found in all cells, that are involved in creating and maintaining a very high order of compaction. As in all other areas of molecular biology, this information is not merely of academic interest. Many of these enzymes are important targets of antibiotics and other medicines. In Chapter 10, we expand on this discussion of tertiary

packaging by examining the complete structure of chromosomes in the context of the structural DNA-binding proteins unique to eukaryotes and bacteria.

9.1 The Problem: Large DNAs in Small Packages

A great deal of information is packed in the chromosomes: they contain the blueprints for an organism. The genes in each chromosome constitute only part of that information. The chromosomes themselves are macromolecular entities that must be synthesized, packaged, protected, and properly distributed to daughter cells at cell division. Significant segments of every chromosome are dedicated to these functions. All aspects of chromosome function are affected by the reality of chromosome size.

Chromosome Function Relies on Specialized Genomic Sequences

The chromosomes of cells and viruses come in several forms. Bacterial chromosomes are often circular (in the sense of an endless loop rather than a perfect circle). Most eukaryotic chromosomes are linear. In viruses we find additional variations, including both single-stranded and double-stranded forms, as well as RNA genomes. Each type of chromosome structure imposes a unique set of demands on the mechanisms for replicating and transmitting the genome from one generation to the next.

Genes provide the information to specify all the RNAs and proteins produced in a given cell, but other sequences are dedicated to the maintenance of the chromosome itself: initiation and termination of replication, segregation during cell division, and, where necessary, protection and maintenance of the chromosome ends. In bacteria, origins of replication provide start sites for chromosomal replication (see Figure 8-1). Specialized replication-termination regions also exist in most known bacterial species. Within or near these regions, additional sequences serve as binding sites for proteins that ensure the faithful segregation of replicated chromosomes to daughter cells. Eukaryotic chromosomes, too, contain sequences that are critical to chromosome maintenance. Unlike bacteria, eukaryotic chromosomes often have many replication origins. (The structure and function of replication origins are described in Chapter 11.) Eukaryotic chromosomes also have specialized sequences called centromeres and telomeres.

The **centromere** is a segment of each eukaryotic chromosome that functions during cell division as an

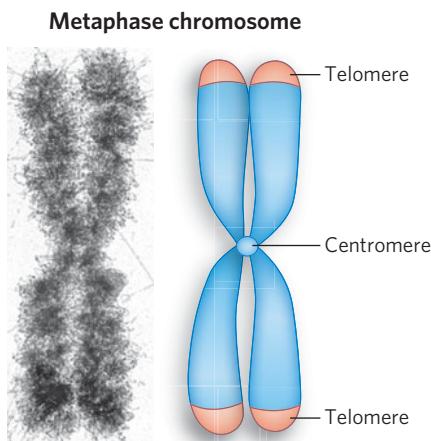
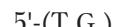


FIGURE 9-1 Linked and condensed sister chromatids of a human chromosome. The products of chromosomal replication in eukaryotes are linked sister chromatids. These are fully condensed at metaphase, during mitosis. The point where they are joined is the centromere. Telomeres are sequences at the ends of the chromatids. [Source: Photo from G. F. Bahr/Biological Photo Service.]

attachment point for proteins that link the chromosome to the mitotic spindle at metaphase (Figure 9-1). This attachment is essential for the equal and orderly distribution of chromosome sets to daughter cells. (See Chapter 2 for a review of the events of mitosis.) The centromeres of *Saccharomyces cerevisiae* have been isolated and studied. The sequences essential to centromere function are about 130 bp long and very rich in A=T pairs. The centromere sequences of higher eukaryotes are much longer and, unlike those of yeast, generally contain regions of simple-sequence DNA consisting of thousands of tandem copies of one or a few 5 to 10 bp sequences. This DNA serves as a binding site for centromere-binding proteins, or cen proteins. The centromere is also the site of kinetochore assembly. Built up on each centromere, the kinetochore anchors

the spindle fibers as chromosomes are segregated into daughter cells during mitosis. Centromeres thus play a key role in stable chromosome segregation during cell division.

Telomeres are sequences at the ends of eukaryotic chromosomes that add stability by protecting the ends from nucleases and providing unique mechanisms for the faithful replication of linear DNA molecules. DNA polymerases cannot synthesize DNA to the very ends of a linear chromosome (see Chapter 11). Solving the end-replication problem is one key function of telomeres, which are replicated by the enzyme telomerase. Telomeres end with multiple repeated sequences of the form



where x and y are generally between 1 and 4 (Table 9-1) and the number of telomere repeats, n , is in the range of 20 to 100 for most single-celled eukaryotes, and generally exceeds 1,500 in mammals. As in centromeres, the telomere repeats serve as binding sites for specialized proteins that are part of telomere function. These proteins package the telomeres and help maintain them in actively dividing cells (see Chapter 11).

Artificial chromosomes have been constructed as a means of better understanding the functional significance of many structural features of eukaryotic chromosomes. A reasonably stable artificial linear chromosome requires only three components: a centromere, a telomere at each end, and an appropriate number of replication origins. Yeast artificial chromosomes (YACs) have been developed as a research tool in biotechnology (see Figure 7-7). YACs have also been useful in confirming the critical functions of centromeres and telomeres. Building on this foundation, human artificial chromosomes (HACs) are now being developed. HACs are reasonably stable when introduced into a human tissue culture cell line, if they include human

Table 9-1 Telomere Sequences

Species	Telomere Repeat Sequence	n^*
<i>Homo sapiens</i> (human)	(TTAGGG) $_n$	800-2,500
<i>Tetrahymena thermophila</i> (ciliated protozoan)	(TTGGGG) $_n$	40
<i>Saccharomyces cerevisiae</i> (yeast)	((TG) $_{1-3}$ (TG) $_{2-3}$) $_n$	50-75
<i>Arabidopsis thaliana</i> (plant)	(TTTAGGG) $_n$	300-1,200

*Number of telomere repeats. Telomere length is longer and fluctuates over a wider range in multicellular eukaryotes. In vertebrates, including humans, telomere length declines with the age of the organism in most cells, but not in germ-line cells.

centromere and telomere sequences in addition to active replication origins.

Continued development of HACs, particularly in their efficient introduction into human cells, may eventually provide new avenues for the treatment of genetic diseases. Most genetic diseases can be traced to an alteration in a particular gene that changes or eliminates that gene's function. The process of correcting disease-causing genetic errors in somatic cells is termed somatic gene therapy. Efforts to directly remove such genes and replace them with normal, functional versions at the correct chromosomal locus have met with limited success in human cells. A simpler technique is to introduce the functional genes into random locations on chromosomes through recombination mechanisms (see Chapters 13 and 14). However, this technique has its own set of problems. The inserted gene can run afoul of regulatory mechanisms that suppress gene expression over large segments of a chromosome, effectively silencing any new gene that is inserted there. Random integration can also result in insertion into the coding sequence of another gene, inactivating that gene. If the inactivated gene has a role in the regulation of cell division, uncontrolled cell division and tumor development can result. The introduction of functional gene copies on stable HACs may eventually circumvent these problems. Success will depend on further advances in clarifying the mechanisms by which chromosomes are stably maintained in cells, and on the development of more efficient procedures for introducing large DNAs into the nuclei of a large number of cells in a living human being.

Chromosomes Are Longer Than the Cellular or Viral Packages Containing Them

The observation that genomic DNAs are orders of magnitude longer than the cells or viruses that contain them applies to every class of organism and viral parasite. Lengths of double-stranded nucleic acids are often described in terms of contour length, or the length measured along the axis of the double-helical DNA. For a closed-loop DNA, contour length is the circumference it would have if laid out in a perfect circle. Lengths are more difficult to describe for a single-stranded nucleic acid, particularly when segments of that nucleic acid fold up into secondary structures. These lengths are sometimes approximated by assuming that the single strand is arrayed in the helical path that would be described by one strand of a double helix, then measuring along the resulting axis.

Given the magnitude of the one-dimensional length of a typical chromosome, how can it be accommodated

within the three-dimensional volume of a virus particle, cell, or nucleus? The compaction mechanisms required for this are highly conserved across the spectrum of living systems. Compaction entails the coiling and structural organization of the chromosome resulting from the action of enzymes; the structural organization is maintained by DNA-binding proteins, including the histones of eukaryotic chromosomes (see Chapter 10), the DNA-binding proteins of bacteria, and the coat proteins of viral particles.

KEY CONVENTION

Molecular biology involves structures with dimensions that are small fractions of a meter. One-thousandth of a meter is 1 millimeter (mm); 1 mm = 1,000 μm (micrometer, or micron) = 1,000,000 nm (nanometer) = 10,000,000 Å (angstrom). Nucleotides, segments of chromosomes, and cells are most often discussed in terms of angstroms, nanometers, and micrometers, respectively.

Viruses Viruses are not free-living organisms; they are infectious parasites that use the resources of a host cell to carry out many of the processes they require to propagate. Many viral particles consist of no more than a genome (usually a single RNA or DNA molecule) surrounded by a protein coat.

Almost all plant viruses and some bacterial and animal viruses have RNA genomes, and they are quite small. For example, the genomes of mammalian retroviruses, such as HIV, have about 9,000 nucleotides, and the genome of the bacteriophage Q β has 4,220 nucleotides. However, even these small nucleic acids have total lengths of about 3 and 1.4 μm , respectively. In comparison, the protein coat of HIV is about 100 nm in diameter, and that of Q β is about 26 nm, so the RNAs are 30 to 50 times longer than their viral protein coats. Both types of virus have linear, single-stranded RNA genomes. Some of the viral coat proteins are effectively RNA-binding proteins, and they enforce a highly compacted folding arrangement of the RNA within the viral particle. An example can be seen in the tobacco mosaic virus (TMV), a pathogen of tobacco plants. The single-stranded RNA genome of TMV, 6,400 nucleotides long, is wound into a tight left-handed helix by its packaging within the rodlike helical protein coat ([Figure 9-2a](#)).

The genomes of DNA viruses vary greatly in size (see Table 8-1). Many viral DNAs are circular for at least part of their life cycle. During viral replication inside a host cell, specific types of viral DNA, called replicative forms, may appear; for example, many linear DNAs

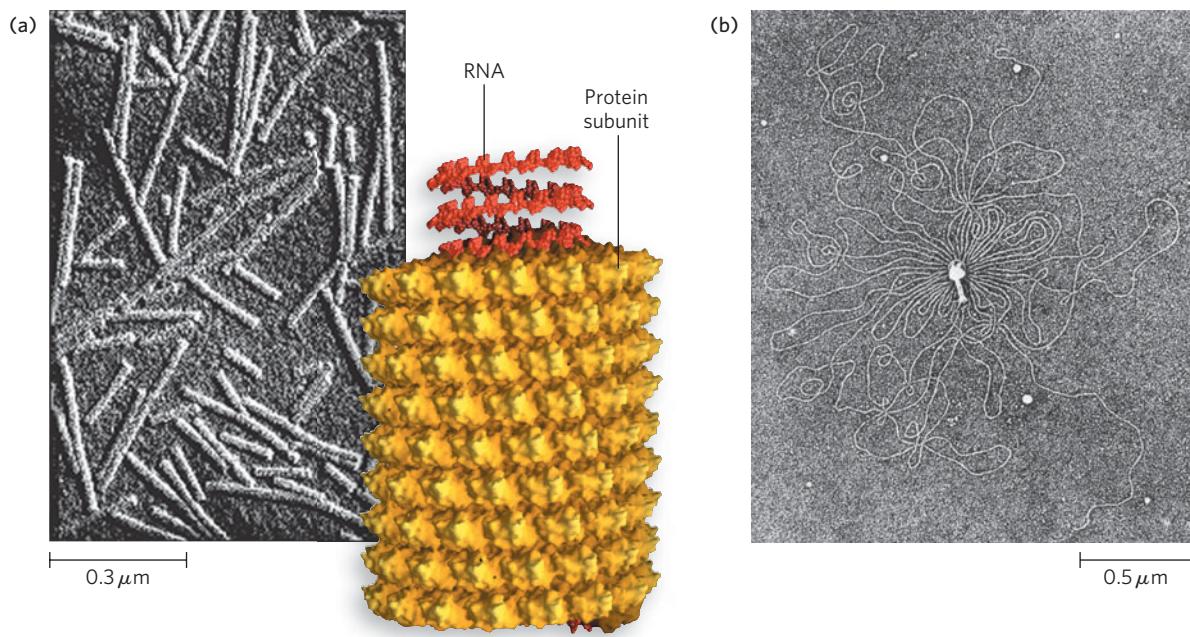


FIGURE 9-2 Genome packaging in a virus. (a) The tobacco mosaic virus has an RNA genome coiled inside a rod-shaped viral coat by RNA-binding proteins, as shown in an electron micrograph and molecular model. (b) A bacteriophage T2 particle was lysed and its DNA allowed to spread on the

surface of distilled water in this electron micrograph. All the DNA shown here is normally packaged inside the phage head. [Sources: (a) Science Source/Photo Researchers; PDB ID 1VTM. (b) From A. K. Kleinschmidt et al., *Biochim. Biophys. Acta* 61:857–864, 1962.]

become circular, and all single-stranded DNAs become double-stranded. Bacteriophage T2 has a double-stranded DNA genome of 160,000 bp, a molecule more than 50 μm long that must be packaged into a virus head about 100 nm across in its longest dimension (Figure 9-2b).

Table 9-2 summarizes the genome and particle dimensions for several DNA viruses. A typical medium-sized DNA virus is bacteriophage λ (lambda), which infects *Escherichia coli*. In its replicative form inside cells, λ DNA is a circular double helix. This double-

stranded DNA contains 48,502 bp and has a contour length of 17.5 μm . Bacteriophage φ X174 is a much smaller DNA virus; the DNA in the viral particle is a single-stranded circle, and the double-stranded replicative form, in the host cell, has 5,386 bp.

Bacteria A single *E. coli* cell contains almost 100 times more DNA than a bacteriophage λ particle. The chromosome of the most common *E. coli* strain studied in laboratories worldwide (K-12 MG1655) is a single double-stranded, circular DNA molecule (Table 9-3).

Table 9-2 The Sizes of DNA and Viral Particles for Some Bacterial Viruses (Bacteriophages)

Virus	Number of Base Pairs in Viral DNA*	Length of Viral DNA (nm)	Long Dimension of Viral Particle (nm) [†]	Chromosome Form
φ X174	5,386	1,939	25	Circular
T7	39,936	14,377	78	Linear
λ (lambda)	48,502	17,460	190	Linear
T4	168,889	60,800	210	Linear

*Data are for the replicative form (double-stranded). The φ X174 chromosome is single-stranded inside the viral particle. The bacteriophage λ chromosome is circularized after it enters a host cell. Calculation of contour length assumes 3.4 Å per base pair (see Figure 6-14).

[†]This measurement includes the head and the tail where relevant.

Table 9-3 DNA, Gene, and Chromosome Content in Some Genomes

Species	Total DNA (bp)	Chromosomes*	Genes
<i>Escherichia coli</i> (bacterium)	4,600,000	1	~4,300
<i>Drosophila melanogaster</i> (fruit fly)	180,000,000	18	~13,600
<i>Arabidopsis thaliana</i> (plant)	125,000,000	10	~25,500
<i>Oryza sativa</i> (plant)	480,000,000	24	~57,000
<i>Mus musculus</i> (mouse)	2,500,000,000	40	~26,000-29,000
<i>Saccharomyces cerevisiae</i> (yeast)	12,068,000	16 [†]	~5,800
<i>Caenorhabditis elegans</i> (nematode)	97,000,000	12 [‡]	~19,000
<i>Homo sapiens</i> (human)	~3,200,000,000	46	~25,000

*Diploid chromosome number for all eukaryotes except yeast.

[†]Haploid chromosome number; wild yeast strains generally have eight (octoploid) or more sets of chromosomes.

[‡]Number for females, with two X chromosomes; males have an X but no Y, for 11 total.

Its 4,639,221 bp have a contour length of about 1.7 mm, some 850 times the length of the *E. coli* cell, 2 μm. In addition to the very large, circular DNA chromosome, many bacteria contain one or more **plasmids**, much smaller circular DNA molecules that are free in the cytosol (Figure 9-3; see also Chapter 7). Most plasmids are only a few thousand base pairs long, but some have more than 100,000 bp. They carry genetic information and undergo replication to yield daughter plasmids, which pass into the daughter cells at cell division. The spread of bacterial plasmids (and transposons) that

confer antibiotic resistance among pathogenic bacteria has reduced the utility of standard antibiotics in medicine and agriculture (Highlight 9-1).

Eukaryotes A yeast cell, one of the simplest eukaryotes, has 2.6 times more DNA in its genome than an *E. coli* cell (see Table 9-3). Cells of *Drosophila melanogaster*, the fruit fly used in classical genetics studies, contain more than 35 times as much DNA as *E. coli* cells, and human cells have almost 700 times as much. The cells of many plants and amphibians contain even

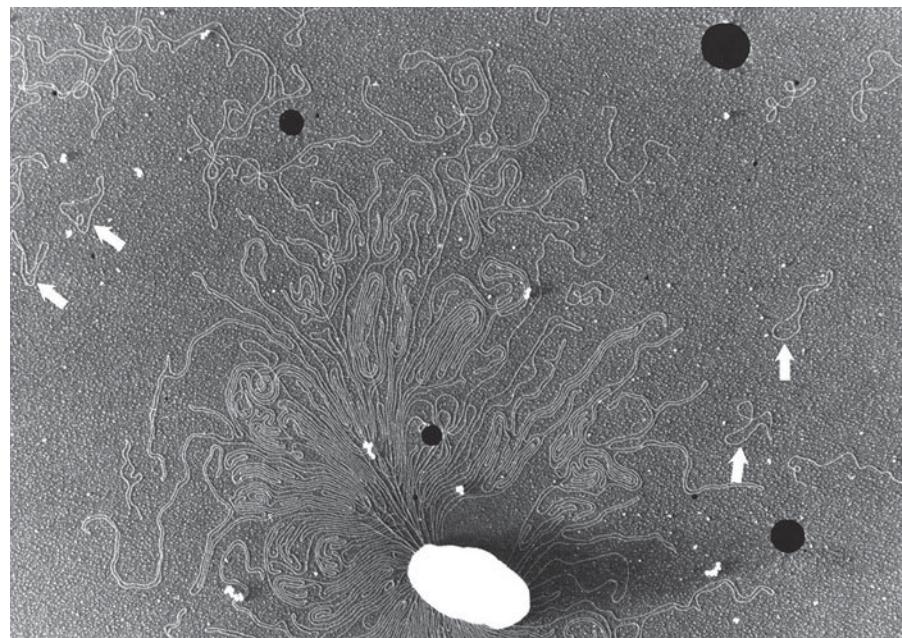


FIGURE 9-3 DNA from a lysed *E. coli* cell. In this electron micrograph, white arrows indicate several small, circular plasmid DNAs. The black spots and white specks are artifacts of the preparation. [Source: Huntington Potter and David Dressler, Department of Neurobiology, Harvard Medical School.]

HIGHLIGHT 9-1 MEDICINE

The Dark Side of Antibiotics

Over the course of the twentieth century, the average life expectancy for people in the developed countries increased 10 years, and the development of antibiotics for the treatment of infectious diseases was a major contributor to this improved longevity. Ironically, an overuse of antibiotics is now leading to their demise as useful therapeutics, as bacterial pathogens evolve resistance to them.

The most common vehicles for transmitting antibiotic-resistance elements between bacterial populations are plasmids, and large numbers are present in the environment. Some plasmids confer no obvious advantage on their host, and their sole function seems to be self-propagation. However, many plasmids carry genes that are useful to the host bacterium. These may include genes that extend the range of environments that can be exploited by the host, such as conferring resistance to naturally occurring antibiotics, new metabolic properties, or the ability to synthesize toxins or agents that facilitate tissue colonization—and thus make the host bacterium pathogenic to other organisms. Given that most antibiotics are natural products (e.g., penicillin is derived from the mold *Penicillium notatum*), it is not surprising that genes conferring antibiotic resistance occur in natural bacterial populations.

Plasmids include a range of sequences involved in their own propagation. These sequences often function in several related bacterial species, and the host range can increase with the aid of small numbers of mutations. When genes conferring antibiotic

resistance are integrated into a plasmid, the resistance element is more readily transferred from one bacterium to another, and even between species. For example, plasmids carrying the gene for the enzyme β -lactamase confer resistance to β -lactam antibiotics, such as penicillin, ampicillin, and amoxicillin. Horizontal gene transfer also can occur, in which plasmids pass from an antibiotic-resistant cell to an antibiotic-sensitive cell of the same or another bacterial species (see Figure 1-11). This can occur when cells of a resistant strain die and rupture, releasing DNA into the environment. If an antibiotic-sensitive strain or species takes up the DNA, it may acquire the antibiotic resistance. When an antibiotic-resistance gene occurs on a conjugational plasmid, its transfer between bacteria becomes particularly efficient. Many antibiotic-resistance elements are further harbored in transposons, which can move from cellular chromosomes to plasmids and back again, further facilitating the dispersal of these elements.

Under the strong selective pressure brought about by widespread antibiotic treatments, bacterial pathogens can acquire antibiotic resistance rapidly. The extensive use of antibiotics in some human populations has encouraged the spread of antibiotic resistance-coding plasmids (as well as transposable elements that harbor similar genes) in disease-causing bacteria. Physicians are becoming increasingly reluctant to prescribe antibiotics unless a clear medical need is confirmed. For similar reasons, the widespread use of antibiotics in animal feeds is being curbed.

more. All of this DNA must fit into a eukaryotic cell that is typically 10 to 20 μm across (although size can vary greatly even within a single organism). The genetic material of eukaryotic cells is apportioned into multiple chromosomes, the diploid ($2n$) number depending on the species. A human somatic cell, for example, has 46 chromosomes (Figure 9-4). Each chromosome of a eukaryotic cell contains a single, very large, duplex DNA molecule. The DNA molecules in the 24 different types of human chromosomes (22 matching pairs plus the X and Y sex chromosomes) vary in length over a 25-fold range. Each type of chromosome in eukaryotes carries a characteristic set of genes.

Eukaryotic cells also have organelles that contain DNA. Mitochondria and chloroplasts carry their own ge-

nomic DNAs (Figure 9-5). The evolutionary origin of mitochondrial and chloroplast DNAs has been the sub-

ject of much speculation. A widely accepted hypothesis, proposed by Lynn Margulis, is that they are vestiges of the chromosomes of ancient bacteria that gained access to the cytoplasm of host cells and became the precursors of these organelles (see Figure 8-17a). Mitochondrial DNA (mtDNA) codes for mitochondrial tRNAs and rRNAs and a



Lynn Margulis [Source: Courtesy of Lynn Margulis.]



FIGURE 9-4 Eukaryotic chromosomes. This is a complete set of chromosomes from a leukocyte of one of the authors. There are 46 chromosomes in every normal human somatic cell. [Source: G. F. Bahr/Biological Photo Service.]

few mitochondrial proteins; more than 95% of mitochondrial proteins are encoded by nuclear DNA. Mitochondria and chloroplasts divide when the cell divides. Their DNA is replicated before and during cell division, and the daughter DNA molecules pass into the daughter organelles.

Mitochondrial DNA molecules are much smaller than nuclear chromosomes. In animal cells, mtDNA contains fewer than 20 kbp (16,569 bp in human

mtDNA) and is a circular duplex. Each mitochondrion typically has 2 to 10 copies of the mtDNA, but the number can be much higher: hundreds in muscle cells, and 100,000 in a mature oocyte. In a few organisms (e.g., trypanosomes), each mitochondrion contains thousands of copies of mtDNA organized into a complex interlinked matrix known as a kinetoplast. Plant cell mtDNA is much larger than that in animal cells, ranging from 200 to 2,500 kbp. Chloroplast DNA (cpDNA) exists as circular duplexes of 120 to 160 kbp. Organelle DNA, like nuclear DNA, undergoes considerable compaction: DNA molecules 5 to 500 μm long must be accommodated in organelles about 1 to 5 μm in diameter.

Some eukaryotes also contain plasmids; they have been found in yeast and some other fungi.

SECTION 9.1 SUMMARY

- All cellular chromosomes contain sequences required for chromosome function, including replication origins. Bacterial chromosomes also contain termination sequences and other sequences necessary for chromosomal segregation during mitosis.
- Eukaryotic chromosomes contain centromeres, which are attachment points for the mitotic spindle, and telomeres, specialized sequences at the ends of a chromosome that protect and stabilize the entire chromosome.
- All genomic DNA and RNA molecules are longer—often orders of magnitude longer—than the viral coats, organelles, and cells in which they are packaged. Special mechanisms for the compaction/condensation of the nucleic acids are employed in chromosomal packaging.
- Viral genomes vary in nucleic acid (DNA or RNA), structure (single-stranded or double-stranded), and length.
- Bacterial cells contain both genomic DNA (usually circular) and plasmids; both types are compacted in the cell and replicated and segregated into daughter cells at cell division.
- Eukaryotic chromosomes are linear and vary in number, depending on the species. Humans have 46 chromosomes, varying in length and condensed to fit into the cell nucleus. Mitochondria and chloroplasts contain their own circular genomes, in numbers ranging from several copies to hundreds of thousands of copies per organelle, depending on cell type.

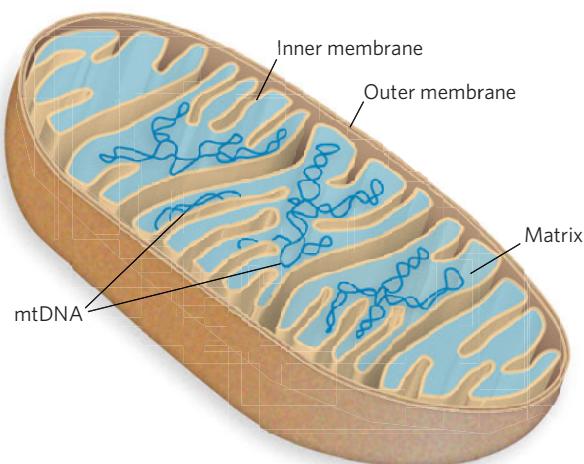


FIGURE 9-5 Mitochondrial DNA. Some mitochondrial proteins and RNAs are encoded by the multiple copies of mtDNA in the mitochondrial matrix. The mtDNA is replicated each time the organelle divides, before cell division.

9.2 DNA Supercoiling

Cellular DNA, as we have seen, is a very long molecule that must somehow be made to fit inside the cell, implying a high degree of structural organization. The folding mechanism must not only pack the DNA but permit access to the information in the DNA, in processes such as replication and transcription. Any consideration of how this is accomplished requires an understanding of an important property of DNA structure known as supercoiling.

Supercoiling simply means the coiling of a coil. An old-fashioned telephone cord, for example, is typically a coiled wire. The path taken by the wire between the base of the phone and the receiver often includes one or more supercoils (Figure 9-6). DNA is coiled in the form of a double helix, with both strands coiling around an axis. **DNA supercoiling** is the further coiling of that axis upon itself (Figure 9-7). As described below, **supercoiled DNA** is generally a manifestation of structural strain. When there is no net coiling or bending of the DNA axis upon itself, the molecule is referred to as **relaxed DNA** (see How We Know). Supercoiling occurs in all chromosomal DNAs in all cells, as well as in viruses that have a double-stranded DNA genome or

generate double-stranded DNA as a replication intermediate.

We might have predicted that DNA compaction involved some form of supercoiling. Perhaps less predictable is that replication and transcription of DNA also affect, and are affected by, supercoiling. Both processes require a separation of DNA strands, which is complicated by the helical interwinding of the strands (Figure 9-8). As we'll see in Section 9.3, specialized enzymes found in all organisms alleviate the stress of replication and transcription by introducing or relaxing supercoils.

Inside the cell, a DNA molecule must bend, and it inevitably becomes supercoiled. However, even when extracted and purified, many circular DNA molecules remain highly supercoiled, even in the absence of protein and other cellular components. This indicates that

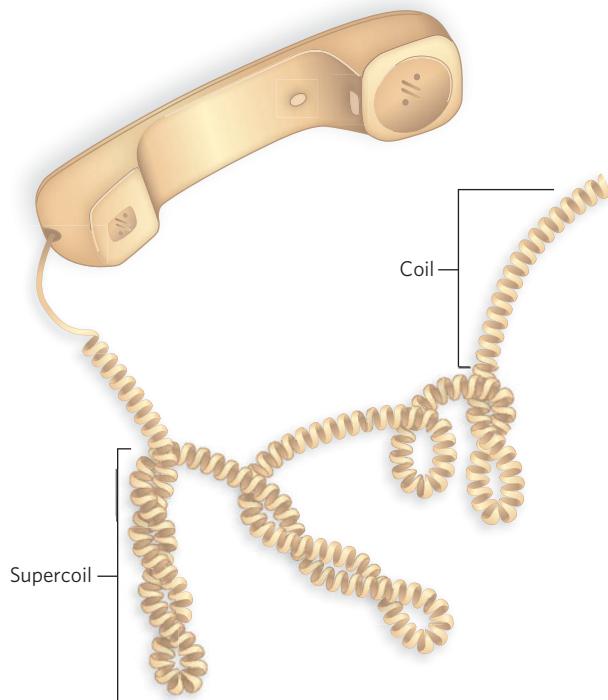


FIGURE 9-6 Supercoils. An old-fashioned phone cord is coiled like a DNA helix, and the coiled cord can supercoil.

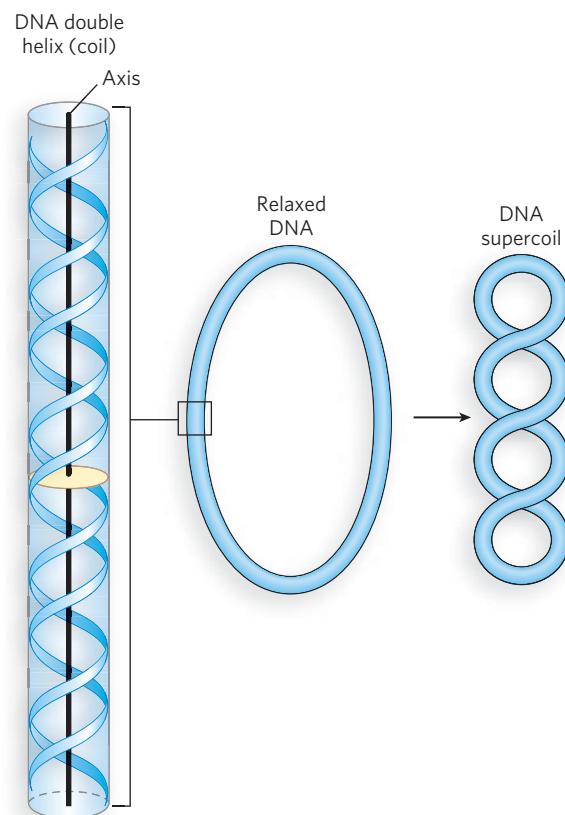


FIGURE 9-7 DNA supercoiling. When the DNA double helix is coiled on itself, it forms a new helix, or superhelix. The DNA superhelix is usually referred to as a supercoil. [Source: Adapted from N. R. Cozzarelli, T. C. Boles, and J. H. White, in *DNA Topology and Its Biological Effect* (N. R. Cozzarelli and J. C. Wang, eds), Cold Spring Harbor Laboratory Press, 1990, pp. 139–184.]

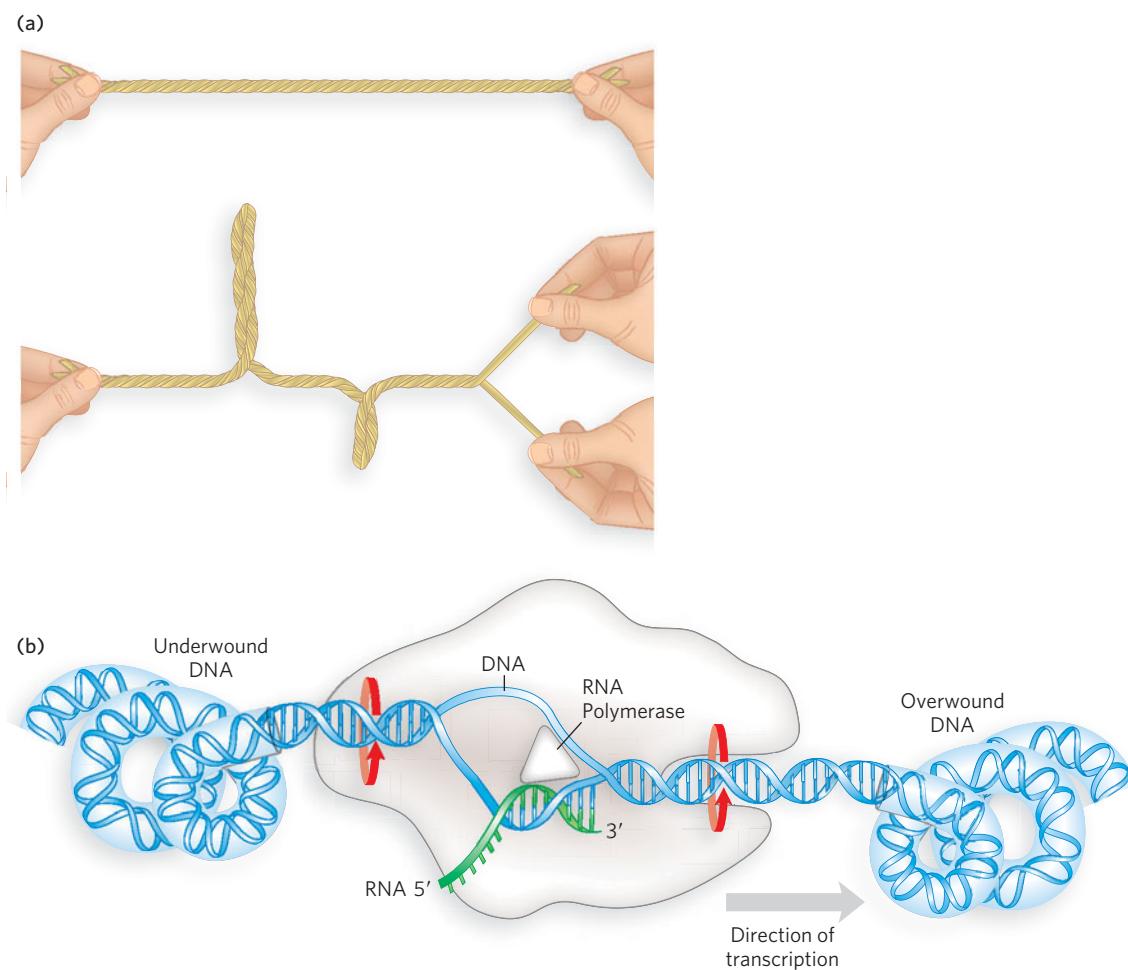


FIGURE 9-8 The effects of replication and transcription on DNA supercoiling. Because DNA is a double-helical structure, strand separation leads to added stress and supercoiling if the DNA is constrained (not free to rotate) ahead of the strand separation. (a) The general effect can be illustrated by twisting two strands of a rubber band about each other to form a double helix. If one end is constrained, separating the two strands at the other end will lead to

twisting. (b) In a DNA molecule, the progress of a DNA polymerase or RNA polymerase (as shown here) along the DNA involves separation of the strands. As a result, the DNA becomes overwound ahead of the enzyme (upstream) and underwound behind it (downstream). Red arrows indicate the direction of winding. [Sources: (a) Adapted from W. Saenger, *Principles of Nucleic Acid Structure*, Springer-Verlag, 1984, p. 452.]

supercoiling is an intrinsic property of DNA tertiary structure, as opposed to an incidental result of spatial constriction. Supercoiling is the direct result of structural strain caused by the underwinding of the DNA—that is, the removal of helical turns relative to the most stable structure of B-form DNA. DNA underwinding is catalyzed by enzymes called topoisomerases, and the degree of DNA underwinding is highly regulated in every cell.

Several measurable properties of supercoiling have been established, and the study of supercoiling has provided many insights into DNA structure and function. This work has drawn heavily on concepts

derived from topology, a branch of mathematics that studies the properties of an object that do not change under continuous deformations. In the context of **DNA topology**, continuous deformations include conformational changes due to stretching, thermal motion, or interaction with proteins or other molecules. The twisting experiments in the Bustamante lab (see Moment of Discovery) provide an example of continuous deformation. Discontinuous deformations involve DNA strand breakage. Topological properties of DNA can be changed only by the breakage and rejoining (ligation) of the backbone of one or both DNA strands.

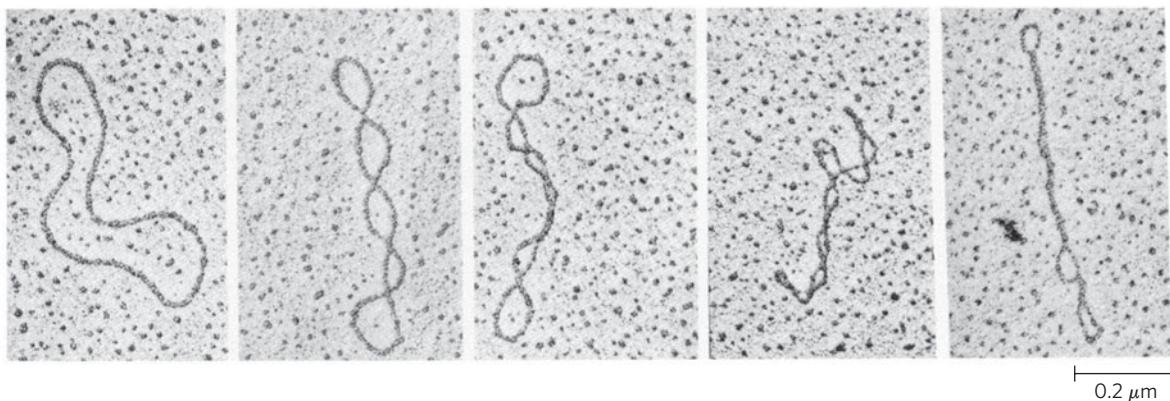


FIGURE 9-9 Relaxed and supercoiled closed-circular plasmid DNAs. The scanning electron micrograph on the far left shows relaxed DNA. Increased supercoiling is shown, from left to right. [Source: Laurien Polder, from A. Kornberg, *DNA Replication*, W. H. Freeman, 1980, p. 29.]

Most Cellular DNA Is Underwound

To understand supercoiling, we must first focus on the properties of small circular DNAs, such as plasmids and small, double-stranded viral DNAs. When these DNA molecules have no breaks in either strand, they are referred to as **closed-circular DNAs**. If the DNA of a closed-circular molecule conforms closely to the B-form structure (see Figure 6-17), with one turn of the double helix per 10.5 bp, it is relaxed rather than supercoiled (Figure 9-9). Supercoiling results when DNA is subject to some form of structural strain. Purified closed-circular DNA is rarely relaxed, regardless of its biological origin. Furthermore, all DNA derived from a given cellular source has a characteristic degree of supercoiling. DNA structure is therefore strained in a manner that is regulated by the cell to induce the supercoiling.

In almost every instance, the strain is a result of underwinding of the DNA double helix in the closed circle. In **DNA underwinding**, the molecule has *fewer* helical turns than would be expected for the B-form structure. Consider, for example, an 84 bp segment of a circular DNA in the relaxed state: it would contain eight double-helical turns, one for every 10.5 bp (Figure 9-10a). If one of these turns were removed, there would be $(84 \text{ bp})/7 = 12.0 \text{ bp}$ per turn, rather than the 10.5 found in B-DNA (Figure 9-10b). This is a deviation from the most stable DNA form, and the molecule would be thermodynamically strained as a result. Generally, much of this strain would be accommodated by coiling the axis on itself, forming a supercoil (Figure 9-10c). Some of the strain in this 84 bp segment would simply become dispersed in the untwisted structure of the larger DNA molecule. In principle, the strain could also be accommodated by separating the two DNA strands over a distance of about

10 bp (Figure 9-10d). In isolated closed-circular DNA, strain introduced by underwinding is generally accommodated by supercoiling rather than strand separation, because coiling the axis of the DNA usually requires less energy than breaking the hydrogen bonds that stabilize paired bases. Note, however, that the underwinding of DNA *in vivo* eases the separation of the DNA strands and thus access to the information they contain.

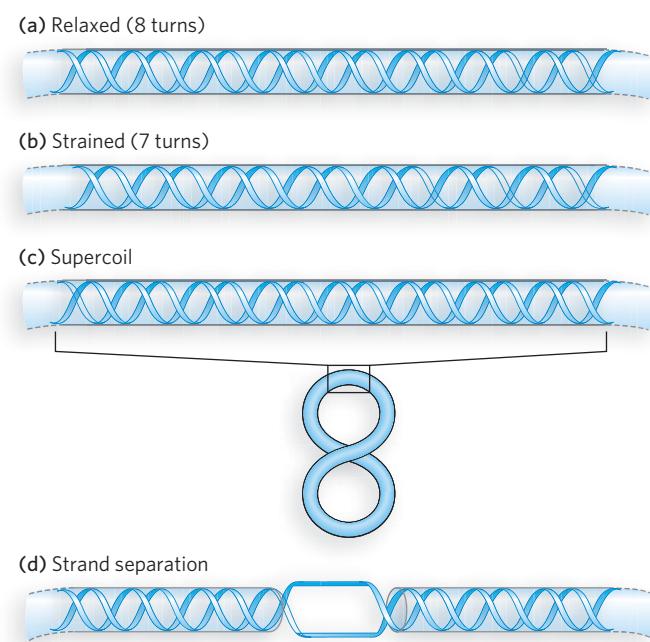


FIGURE 9-10 The effects of DNA underwinding. See text for details. [Source: Adapted from N. R. Cozzarelli, T. C. Boles, and J. H. White, in *DNA Topology and Its Biological Effect* (N. R. Cozzarelli and J. C. Wang, eds), Cold Spring Harbor Laboratory Press, 1990, pp. 139–184.]

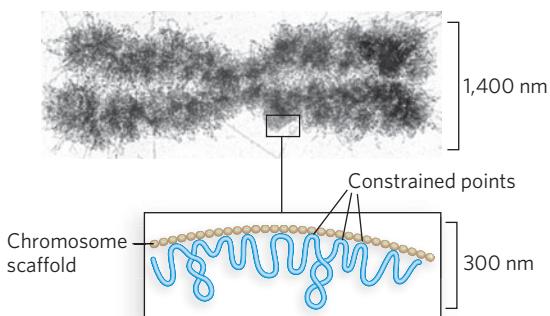


FIGURE 9-11 Loops in a eukaryotic chromosome constrained by scaffold proteins. The chromatin scaffold is attached to the chromosome at intervals, with the DNA between the attachment points defining loops that are topologically constrained. [Source: Photo from G. F. Bahr/Biological Photo Service.]

Every cell actively underwinds its DNA with the aid of enzymes (see Section 9.3), and the resulting strained state represents a form of stored energy. Underwinding thus accomplishes two things. First, cells maintain DNA in an underwound state to promote its compaction by coiling. Second, underwinding facilitates strand separation and enzymatic access to the encoded information. DNA underwinding is thus important to the enzymes of DNA replication and transcription, which must bring about strand separation as part of their function.

The underwound state can be maintained only if the DNA is a closed circle or, if linear, is bound and stabilized by proteins so that the strands are not free to rotate about each other. If there is a break in one strand of an isolated, protein-free circular DNA, free rotation at that point will cause the underwound DNA to revert spontaneously to the relaxed state. In a closed-circular DNA molecule, however, the number of helical turns cannot be changed without at least transiently breaking one of the DNA strands. The number of helical turns in a DNA molecule therefore provides a precise description of supercoiling. In the linear chromosomes of eukaryotic cells, DNA underwinding is maintained by bound proteins that constrain the DNA in an elaborate structure called chromatin. In chromatin, large loops of DNA are constrained at their base, such that each loop is topologically fixed as if it were circular (Figure 9-11; we discuss this further in Chapter 10).

DNA Underwinding Is Defined by the Topological Linking Number

The **linking number (Lk)** is a topological property of double-stranded DNA—that is, it does not vary when

the DNA is bent or deformed. To define linking number, imagine the separation of the two strands of a double-stranded circular DNA. If the two strands are joined as shown in Figure 9-12a, they are effectively linked by a topological bond. Even if all hydrogen bonds and base-stacking interactions were removed so that the strands were not in physical contact, the two strands would still be linked. If we think of one of the circular strands as the boundary of a surface (such as the soap film on the loop of a bubble wand before you blow a bubble), the linking number can be defined as the number of times the second strand pierces this surface (Figure 9-12b). The linking number for a closed-circular DNA is always an integer. By convention, if the DNA strands are interwound in a right-handed helix, the linking number is positive (+); for strands interwound in a left-handed helix, the linking number is negative (-). Negative linking numbers are, for all practical purposes, not encountered in DNA.

Consider a closed-circular DNA with 2,100 bp (Figure 9-13a). When the molecule is relaxed, the linking

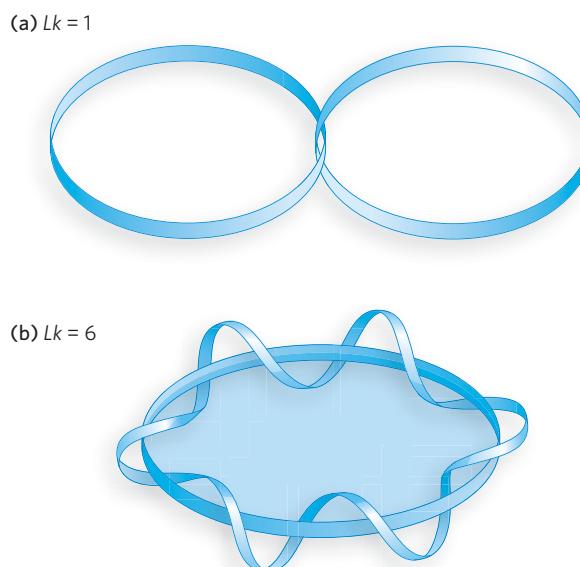


FIGURE 9-12 Linking number (Lk). Each blue ribbon represents one strand of a double-stranded DNA molecule. (a) The linking number of the molecule is 1. (b) The linking number is 6. One of the strands is kept untwisted for illustrative purposes, to define the border of an imaginary surface (solid blue oval). The number of times the twisting strand (its relative length exaggerated here) penetrates this surface is the linking number. [Source: Adapted from N. R. Cozzarelli, T. C. Boles, and J. H. White, in *DNA Topology and Its Biological Effect* (N. R. Cozzarelli and J. C. Wang, eds), Cold Spring Harbor Laboratory Press, 1990, pp. 139–184.]

number is equal to the number of base pairs divided by the number of base pairs per turn, 10.5; in this case, $Lk = 200$. As noted above, DNA can have a topological property such as a linking number only if both strands are intact. If there is a break in either strand, the strands can, in principle, be unraveled and separated; in this case, no topological bond exists and Lk is undefined ([Figure 9-13b](#)).

We can now describe DNA underwinding in terms of changes in the linking number. The linking number in relaxed DNA, Lk_0 , is used as a reference. For the molecule in Figure 9-13a, $Lk_0 = 200$; if two turns are removed from this molecule by breaking a strand, unwinding, and joining the ends back together, $Lk = 198$ (Figure 9-13c). The change can be described by the equation:

$$\Delta Lk = Lk - Lk_0 \quad (9-1)$$

For our example, $\Delta Lk = 198 - 200 = -2$.

We can also express the change in linking number independent of the length of the DNA molecule. This quantity, usually called the **superhelical density** (σ), is a measure of the number of turns removed relative to the number present in relaxed DNA:

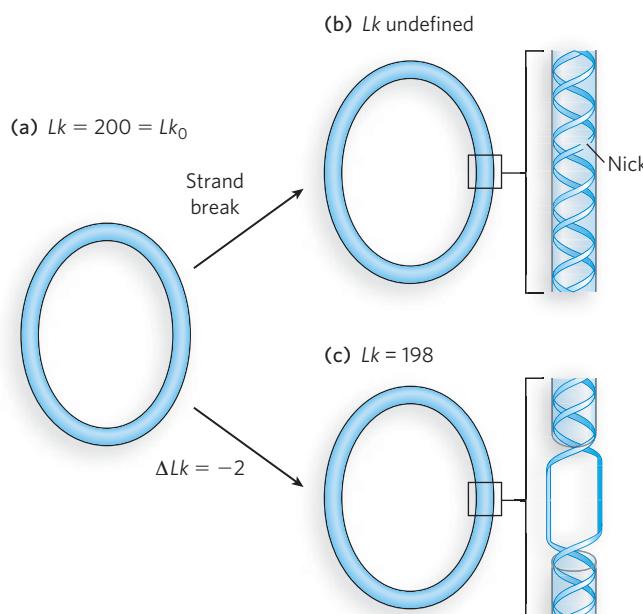


FIGURE 9-13 Linking number of closed-circular DNAs. A 2,100 bp molecule is shown in three forms: (a) relaxed, $Lk = 200$; (b) relaxed with a nick in one strand, Lk undefined; (c) underwound by two turns, $Lk = 198$. The underwound molecule is generally supercoiled, but underwinding also facilitates the separation of DNA strands.

$$\sigma = \frac{\Delta Lk}{Lk_0} \quad (9-2)$$

In the Figure 9-13c example, $\sigma = -0.01$, which means that 1% (2 of 200) of the helical turns present in the DNA (when it is relaxed, in its B form) have been removed. The degree of underwinding in cellular DNAs generally falls in the range of 5% to 7%; that is, $\sigma = -0.05$ to -0.07 . The negative sign indicates that the change in the linking number is due to underwinding of the DNA. The supercoiling induced by underwinding is therefore defined as **negative supercoiling**. Conversely, under some conditions DNA can be overwound, resulting in **positive supercoiling**.

Notice that negative supercoiling results in a twisting of the axis of the DNA to form a right-handed superhelix, and positive supercoiling results in a left-handed superhelix ([Figure 9-14](#)). Supercoiling is not a random process; it is largely prescribed by the torsional strain imparted to the DNA by decreasing or increasing the linking number relative to that of B-DNA.

The linking number can be changed by ± 1 by breaking one DNA strand, rotating one of the ends 360° about the unbroken strand, and rejoining the broken ends. This reaction is catalyzed by topoisomerases (see Section 9.3). The change in linking number has no effect on the number of base pairs or the number

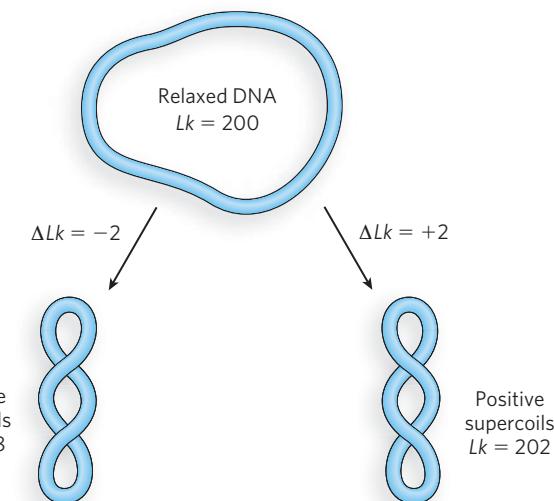


FIGURE 9-14 Negative and positive supercoils. For the relaxed DNA in Figure 9-13a, underwinding or overwinding by two helical turns ($\Delta Lk = \pm 2$) produces negative or positive supercoiling, respectively. Notice that the DNA axis in the two forms twists in opposite directions.

of atoms in the circular DNA molecule. Two forms of a circular DNA that differ only in a topological property such as linking number are referred to as **topoisomers**.

We can break down the linking number into two structural components, writhe (Wr) and twist (Tw) (Figure 9-15). **Writhe (Wr)** is a measure of the coiling of the helical axis, and **twist (Tw)** describes the local twisting or spatial relationship of neighboring base pairs. When the linking number changes, some of the resulting strain is usually compensated for by writhe (supercoiling) and some by changes in twist, giving rise to the equation:

$$Lk = Tw + Wr \quad (9-3)$$

Tw and Wr need not be integers. Twist and writhe are geometric rather than topological properties, because they may be changed by deformation of a closed-circular DNA molecule. Tw and Wr may change in a reciprocal manner without altering Lk .

In addition to causing supercoiling and making strand separation somewhat easier, the underwinding of DNA facilitates structural changes in the molecule. Although these changes are of less physiological importance, they help illustrate the effects of underwinding. Recall that a cruciform generally contains a few unpaired bases (see Figure 6-20b); DNA underwinding helps maintain the required strand separation in regions where palindromic sequences allow cruciform formation (Figure 9-16). Underwinding of a right-handed DNA helix also facilitates the formation of short stretches of left-handed Z-DNA in regions where

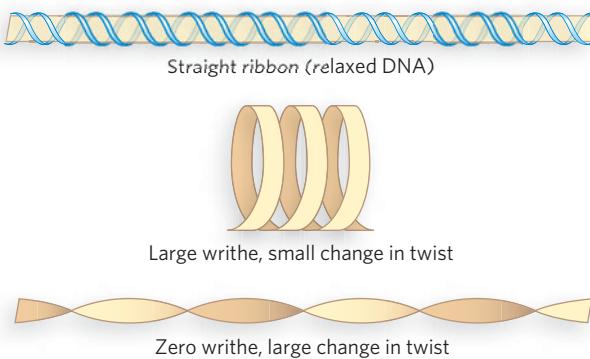


FIGURE 9-15 Writhe and twist. The beige ribbon represents the axis of a relaxed DNA molecule. Strain from underwinding of the DNA can manifest as writhe or twist. Topological changes in the linking number are usually accompanied by geometric changes in both writhe and twist.

the base sequence is consistent with the Z form (see Chapter 6).

DNA Compaction Requires a Special Form of Supercoiling

All supercoiled DNA molecules are similar in several respects. The supercoils are right-handed in a negatively supercoiled DNA molecule, and they tend to be extended and narrow rather than compacted, often with multiple branches (Figure 9-17a). At the superhelical densities normally encountered in cells, the length of the supercoil axis (the axis about which the supercoils turn), including branches, is about 40% of the length of the DNA. This type of supercoiling is referred to as **plectonemic supercoiling** (Figure 9-17b). This term can be applied to any structure with strands intertwined in some simple and regular way, and it is a good

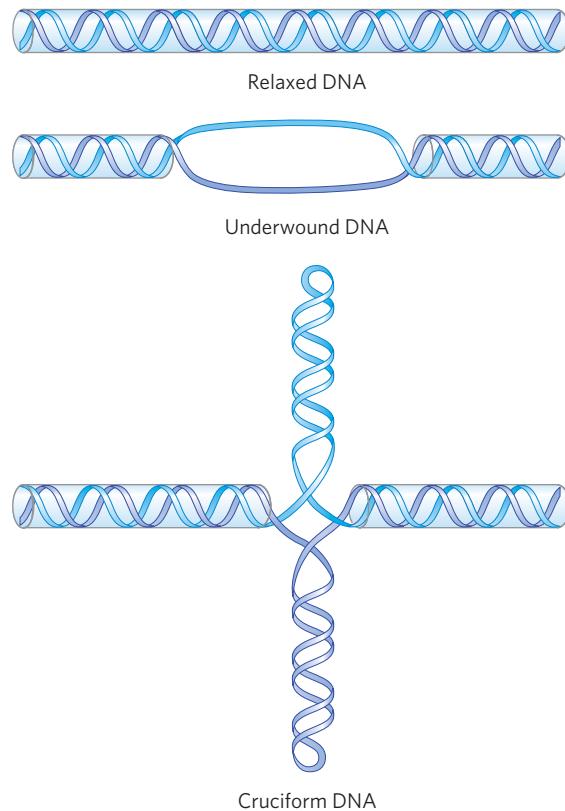


FIGURE 9-16 Promoting cruciform structures by DNA underwinding. Cruciforms can form at palindromic sequences, but they seldom occur in relaxed DNA because linear DNA accommodates more paired bases than the cruciform structure. DNA underwinding facilitates the partial strand separation required for promoting cruciform formation at appropriate sequences.

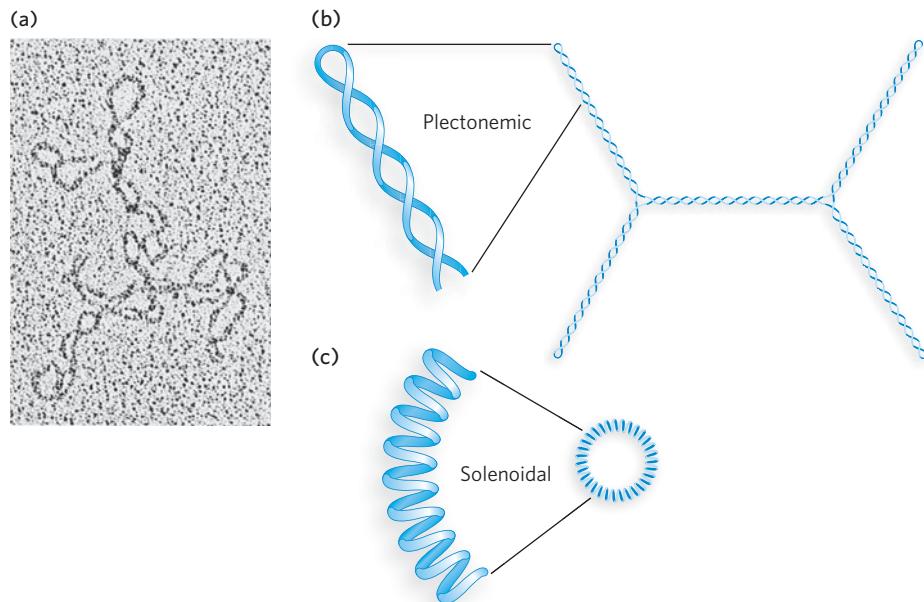


FIGURE 9-17 Plectonemic and solenoidal supercoiling.

(a) Electron micrograph of plectonemically supercoiled plasmid DNA. (b) Plectonemic supercoiling consists of extended right-handed coils. (c) Solenoidal supercoiling of the same DNA molecule depicted in (b), drawn to scale. Solenoidal negative

supercoiling consists of tight left-handed turns about an imaginary tubelike structure. Solenoidal supercoiling provides a much greater degree of compaction. [Source: (a) James H. White, T. Christian Boles, and N. R. Cozzarelli, Department of Molecular and Cell Biology, University of California, Berkeley.]

description of the general structure of supercoiled DNA in solution.

Plectonemic supercoiling, the form observed in isolated DNAs in the laboratory, does not produce sufficient compaction to package DNA in the cell. A second form, **solenoidal supercoiling**, can be adopted by an underwound DNA (Figure 9-17c). Instead of the extended right-handed supercoils characteristic of the plectonemic form, solenoidal supercoiling involves tight left-handed turns, much like a garden hose neatly wrapped on a reel. Although their structures are very different, plectonemic and solenoidal supercoiling are two forms of negative supercoiling that can be adopted by the *same* segment of underwound DNA. The two forms are readily interconvertible. The plectonemic form is more stable in solution, but the solenoidal form can be stabilized by protein binding and provides a much greater degree of compaction. Solenoidal supercoils are formed when DNA is wrapped around the nucleosomes that make up eukaryotic chromatin (see Chapter 10). Similarly, in bacteria, the tight wrapping of DNA around a variety of DNA-binding proteins gives rise to solenoidal supercoils. Solenoidal supercoiling is the primary mechanism by which underwinding contributes to DNA compaction.

SECTION 9.2 SUMMARY

- Most cellular DNAs are supercoiled. Underwinding decreases the total number of helical turns in the DNA relative to the relaxed, B form. To maintain an underwound state, DNA must be either a closed circle or, if linear, bound to protein.
- Underwinding is quantified by the linking number (Lk), a topological parameter that describes the number of times two DNA strands are intertwined.
- Underwinding is measured in terms of σ , the superhelical density; $\sigma = (Lk - Lk_0)/Lk_0$. For cellular DNAs, σ is typically -0.05 to -0.07 , which means that about 5% to 7% of the helical turns in the DNA have been removed. DNA underwinding facilitates strand separation by enzymes of DNA metabolism.
- Plectonemic supercoiling includes right-handed branches and is the most common type of supercoiling in isolated DNA. Solenoidal supercoiling, an alternative form that produces a greater degree of compaction, is characterized by tight left-handed turns that are stabilized by wrapping the DNA around proteins; this occurs in eukaryotic and bacterial chromosomes.

9.3 The Enzymes That Promote DNA Compaction

DNA supercoiling—or, more specifically, DNA underwinding—is a precisely regulated process that influences many aspects of DNA metabolism. As we've seen, it allows access to DNA during replication and transcription, and it contributes to DNA condensation during mitosis. The underwinding and relaxation of DNA are catalyzed by DNA topoisomerases, enzymes that break one or both DNA strands to allow a topological change, and then religate them. Additional condensation of cellular DNA is facilitated by SMC proteins, a class of enzymes that reversibly form protein loops that link DNA segments, affecting both condensation/compaction of chromosomes and cohesion of daughter DNA molecules for periods following replication. The maintenance of the underwound and condensed state of chromosomes by structural DNA-binding proteins such as histones is discussed in Chapter 10.

Topoisomerases Catalyze Changes in the Linking Number of DNA

All cells, from bacteria to eukaryotes, have enzymes with the sole function of underwinding and relaxing DNA. **Topoisomerases** increase or decrease the extent of DNA underwinding by changing the linking number. They play an especially important role in the complex changes in DNA topology during replication and DNA packaging.

There are two classes of topoisomerases (Table 9-4). **Type I topoisomerases** break one of the two DNA strands, pass the unbroken strand through the break,

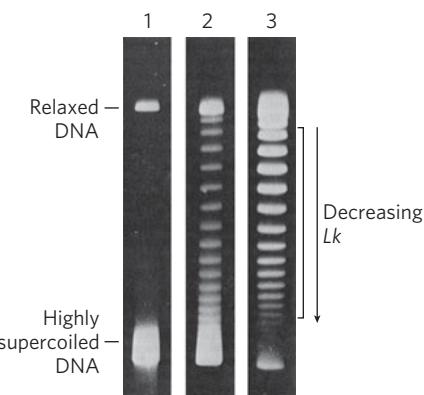


FIGURE 9-18 Visualizing topoisomers. In this experiment, DNA molecules (plasmids) have an identical number of base pairs but differ in the degree of supercoiling. In lane 1, highly supercoiled DNA migrates as a single band. Lanes 2 and 3 show the effect of treating supercoiled DNA with a type I topoisomerase; the DNA in lane 3 was treated for a longer time than the DNA in lane 2. Each individual band in the bracketed region of lane 3 contains DNA plasmids with the same linking number; Lk changes by 1 from one band to the next. [Source: W. Keller, Proc. Natl. Acad. Sci. USA 72:2553, 1975.]

and ligate the broken ends; they change Lk in increments of 1. **Type II topoisomerases** break both DNA strands and change Lk in increments of 2. The DNA is never released from the enzyme during these topological transactions, so uncontrolled relaxation of the DNA does not occur.

The activity of these enzymes can be observed with agarose gel electrophoresis, which separates DNA species according to their topoisomeric form (Figure 9-18).

Table 9-4 Topoisomerases in Bacteria and Eukaryotes

	Class	Function
Bacteria		
Topoisomerase I	Type I	Relaxes negative supercoils
Topoisomerase II (DNA gyrase)	Type II	Introduces negative supercoils
Topoisomerase III	Type I	Specialized functions in DNA repair and replication
Topoisomerase IV	Type II	Decatenation of replicated chromosomes
Eukaryotes		
Topoisomerase I	Type I	Relaxes negative supercoils, especially during DNA replication
Topoisomerase II α	Type II	Relaxes positive or negative supercoils; functions in chromatin condensation, replication, transcription
Topoisomerase II β	Type II	Relaxes positive or negative supercoils; functions in chromatin condensation, replication, transcription
Topoisomerase III	Type I	Specialized functions in DNA repair and replication

A population of identical plasmid DNAs with the same linking number migrates as a discrete band during electrophoresis. DNA topoisomers that are more supercoiled are more compact and migrate faster in the gel. Topoisomers with Lk values differing by as little as 1 can be separated by this method, so the changes in linking number induced by topoisomerases are readily detected.

E. coli has at least four individual topoisomerases, I through IV. Topoisomerases I and III are of type I, and they generally relax DNA by introducing transient single-strand breaks to remove negative supercoils (increasing Lk). Figure 9-19 shows the steps in the reaction catalyzed by bacterial type I topoisomerases (also see How We Know). A DNA molecule binds to the topoisomerase, and one DNA strand is cleaved (step 1). The enzyme changes conformation (step 2), and the unbroken DNA strand moves through the break in the first strand (step 3). Finally, the DNA strand is ligated and released (step 4). ATP is not used in this reaction. The enzyme promotes the formation of a less strained, more relaxed state by removing supercoils.

The topoisomerase must both cleave a DNA strand and ligate it again after the topological change is complete. The phosphodiester bond is not simply hydrolyzed, because this would entail loss of a high-energy bond, and an activation step would then be required to promote the subsequent ligation. Instead, a nucleophile on the enzyme (usually a Tyr residue, as in the case of *E. coli* topoisomerase I) attacks the phosphodiester bond, displacing the 3' hydroxyl and forming a covalent 5'-phosphotyrosine linkage with the DNA strand at the break. Strand passage brings about the topological change. The broken strand is then ligated by means of a direct attack of the free 3'-hydroxyl group on the phosphotyrosyl linkage. In this scheme, one high-energy bond is replaced by another at each chemical step. The resulting conservation of energy allows strand ligation without an activation step that would otherwise consume ATP.

The Two Bacterial Type II Topoisomerases Have Distinct Functions

Bacterial topoisomerase II, also known as DNA gyrase, can introduce negative supercoils (decrease Lk). This enzyme cleaves both strands of a DNA molecule (thus is a type II topoisomerase) and passes another duplex through the break (see How We Know). It uses the energy of ATP to drive key conformational changes that counteract the thermodynamically unfavorable introduction of negative supercoils that the gyrase activity

brings about. Bacterial DNA gyrases are the only topoisomerases known to actively introduce negative supercoils.

Gyrase is composed of two types of subunits, GyrA and GyrB, functioning as a GyrA₂GyrB₂ heterotetramer (Figure 9-20a). GyrB interacts with DNA and ATP, and catalyzes ATP binding and hydrolysis. Parts of GyrB form the entry point for DNA, called the N-gate. The DNA exits through a domain in GyrA called the C-gate. A separate domain of GyrA binds DNA and promotes DNA wrapping. Reaction steps are detailed in Figure 9-20b. To introduce negative supercoils, a gyrase complex first binds to a DNA segment via the N-gate (step 1), and wraps the DNA around itself (step 2). ATP is bound and both strands of the DNA are cleaved by active-site Tyr residues (step 3), forming two 5'-phosphotyrosine intermediates. ATP hydrolysis is coupled to the passage of a second region of DNA through the cleaved DNA strands, entering at the N-gate and exiting at the C-gate. To complete the reaction (step 4), the DNA strands are ligated by attack of the free 3'-hydroxyl groups on the phosphotyrosine intermediates. The complex is then poised to initiate another reaction cycle. The degree of supercoiling of bacterial DNA is maintained by regulation of the net activity of topoisomerase I, which increases Lk , and DNA gyrase, which decreases Lk .

Immediately following replication, the circular daughter chromosomes of bacteria are topologically intertwined. Circular DNAs that are intertwined in this way are called **catenanes** (Figure 9-21). The second bacterial type II topoisomerase, DNA topoisomerase IV, has a specialized function in unlinking the catenated daughter chromosomes, allowing their proper segregation at cell division. Unlike DNA gyrase, this enzyme does not use ATP and does not introduce negative supercoils.

Eukaryotic Topoisomerases Have Specialized Functions in DNA Metabolism

Eukaryotic cells also have type I and type II topoisomerases. The type I enzymes are called topoisomerases I and III. They function primarily in relieving tension and resolving topological problems in DNA during replication and repair. The type II enzymes are topoisomerases II α and II β (see Table 9-4). Eukaryotic type II topoisomerases cannot underwind DNA (introduce negative supercoils), but they can relax both positive and negative supercoils. They function in all aspects of eukaryotic DNA metabolism, resolving a range of topological problems that occur during replication, transcription, and repair. They play an especially important

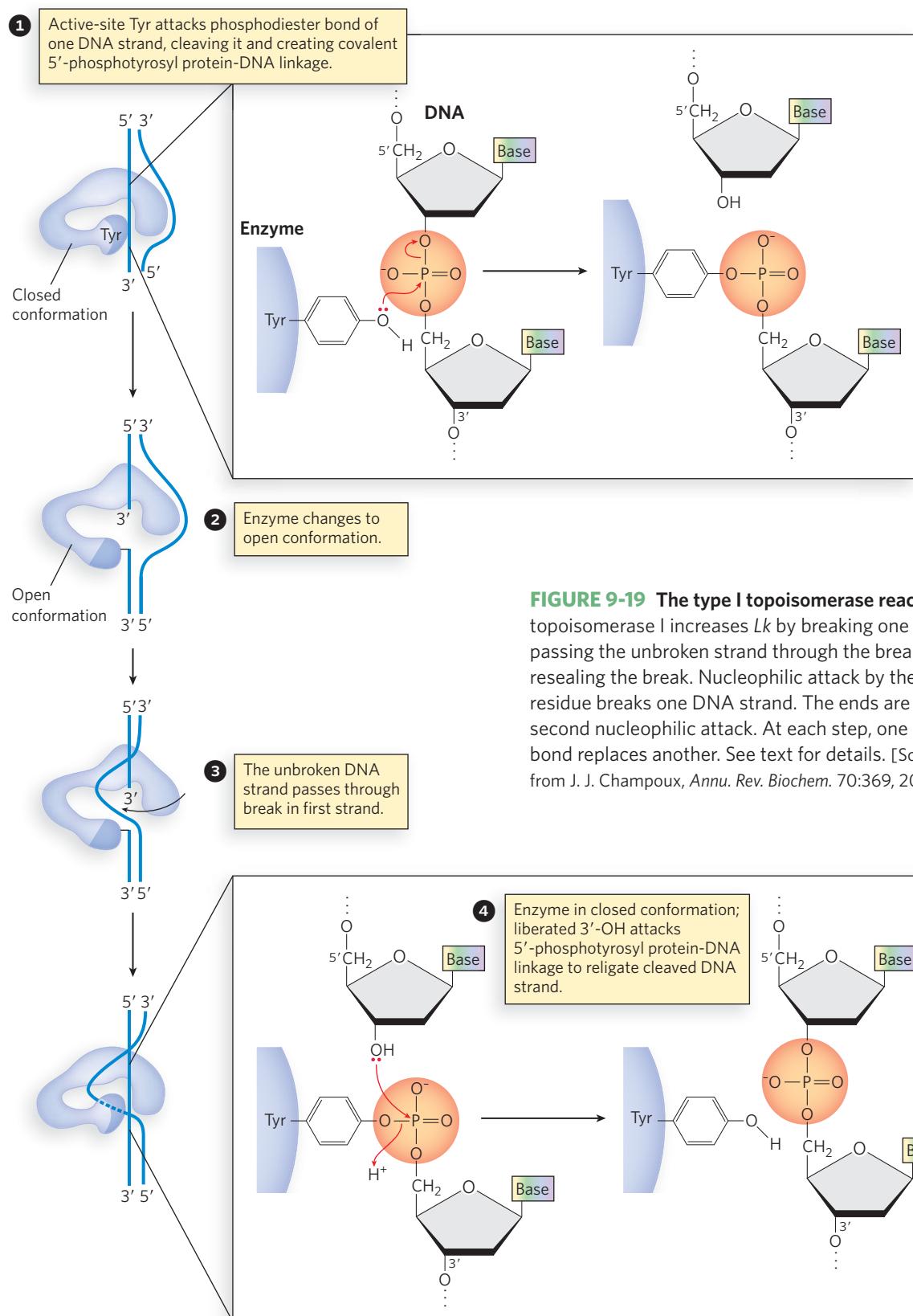


FIGURE 9-19 The type I topoisomerase reaction. Bacterial topoisomerase I increases Lk by breaking one DNA strand, passing the unbroken strand through the break, then resealing the break. Nucleophilic attack by the active-site Tyr residue breaks one DNA strand. The ends are ligated by a second nucleophilic attack. At each step, one high-energy bond replaces another. See text for details. [Source: Adapted from J. J. Champoux, *Annu. Rev. Biochem.* 70:369, 2001, Fig. 3.]

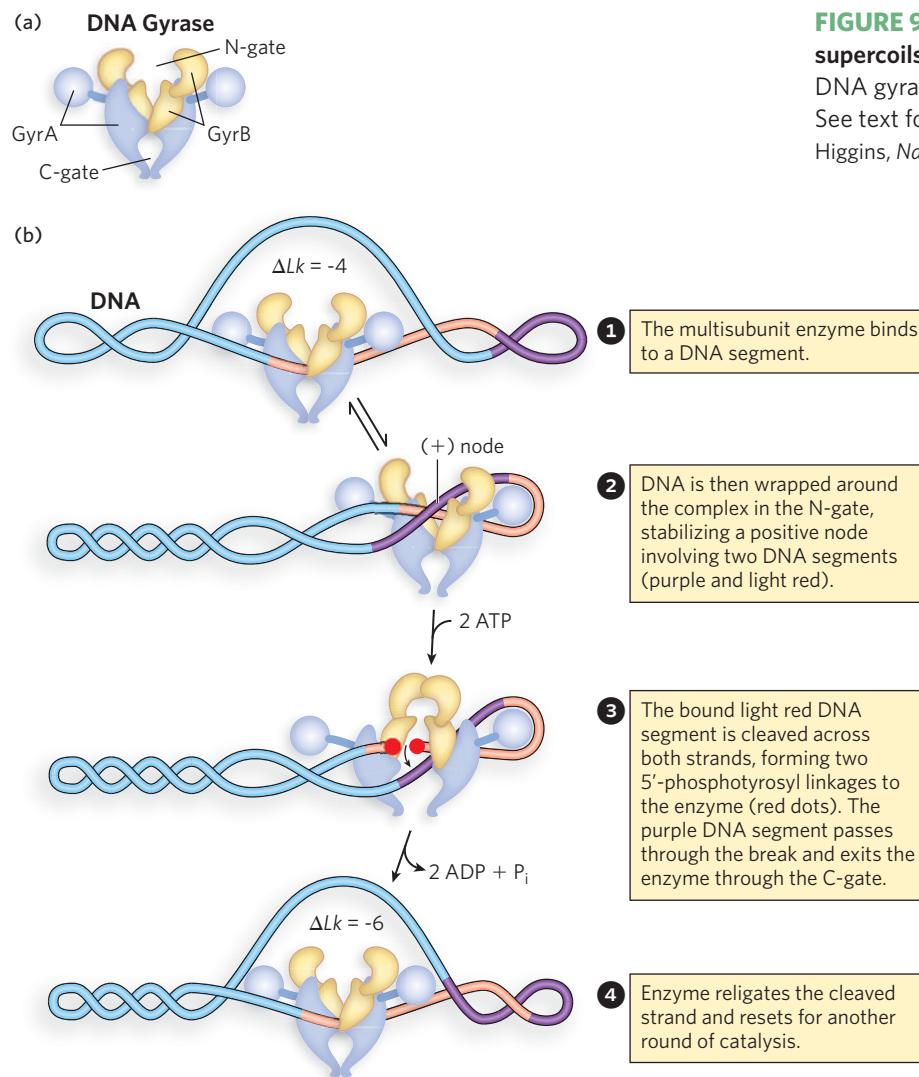


FIGURE 9-20 Introduction of negative DNA supercoils by DNA gyrase. (a) The structure of DNA gyrase. (b) The mechanism of gyrase action. See text for details. [Source: Adapted from N. P. Higgins, *Nat. Struct. Mol. Biol.* 14:264–271, 2007.]

role in the condensation of chromosomes into highly structured chromatin.

Although eukaryotes do not have an enzyme that can introduce negative supercoils into DNA, when a circular DNA is isolated from a eukaryotic cell (e.g., a plasmid from yeast), it is negatively supercoiled. This reflects the generally underwound state of cellular DNA in eukaryotic cells. One probable origin of negative supercoils in eukaryotic DNA is the tight wrapping of the DNA around a nucleosome in chromatin, which introduces a negative solenoidal supercoil (see Chapter 10). In the absence of any change in Lk , a positive supercoil must form elsewhere in the DNA to compensate (Figure 9-22). The type II topoisomerases can relax the unbound positive supercoils that arise in this way. The bound and stabilized negative supercoils are left behind,

conferring a net negative superhelicity on the DNA. Next to the histones that make up the nucleosomes, type II topoisomerases are the most abundant proteins in chromatin.

Figure 9-23 shows the reaction catalyzed by eukaryotic type II topoisomerases. The multisubunit enzyme binds a DNA molecule (step 1). The gated cavities above and below the bound DNA are the N-gate and C-gate, respectively. The second segment of the same DNA is bound at the N-gate (step 2). Both strands of the first DNA are now cleaved (step 3; the chemistry is similar to that in Figure 9-19), forming phosphotyrosine intermediates. The second DNA segment passes through the break in the first segment (step 4), and the broken DNA is ligated and the second segment released through the C-gate (step 5). Two ATPs are bound and

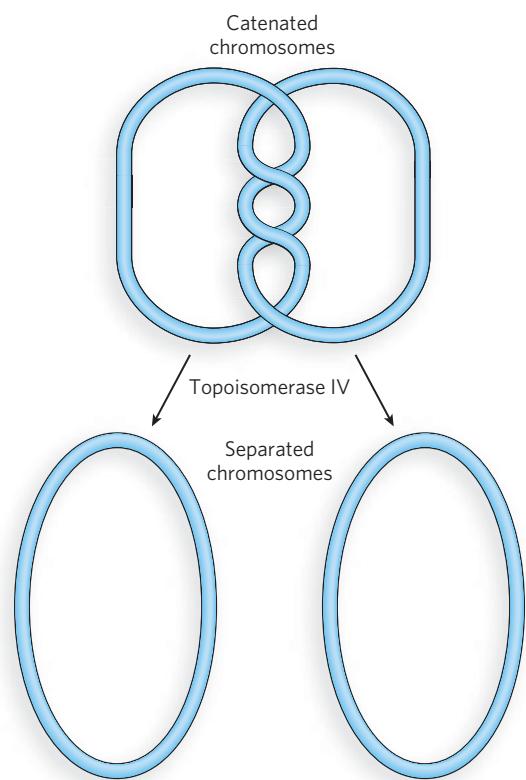


FIGURE 9-21 Solving topological problems with type II topoisomerases. Type II topoisomerases resolve knots and catenanes that arise in DNA by passing one duplex through a transient double-strand break in another duplex.

hydrolyzed during this cycle; it is likely that one is hydrolyzed in the step leading to the complex in step 4. Additional details of the ATP hydrolysis have yet to be worked out.

As we'll show in later chapters, topoisomerases are crucial to every aspect of DNA metabolism. As a consequence, they are important drug targets for the treatment of bacterial infections and cancer (Highlight 9-2).

SMC Proteins Facilitate the Condensation of Chromatin

Whereas topoisomerases influence supercoiling by changing the linking number of chromosomes, **SMC proteins** (structural maintenance of chromosomes) promote chromosome condensation by creating physical contact between segments of DNA that may otherwise be quite distant in the chromosome, or even on different chromosomes. These enzymes have integral roles in DNA condensation and chromosome segregation during mitosis, as well as in DNA repair. They perform their tasks by lining up

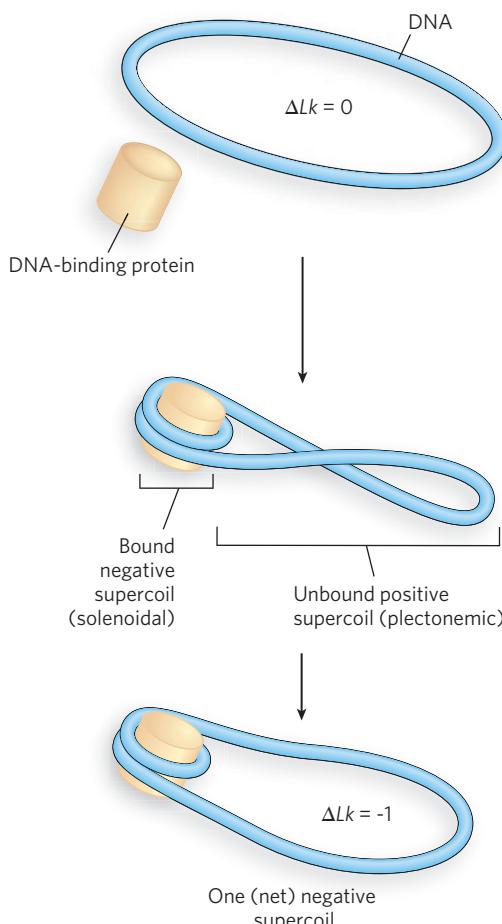


FIGURE 9-22 The origin of negative supercoiling in eukaryotic DNA. When DNA is wrapped tightly around a DNA-binding protein or protein complex, a solenoidal negative supercoil is fixed in the DNA. In a constrained DNA molecule, positive supercoils must develop elsewhere to compensate for the resulting strain. Relaxation of unbound positive supercoils by topoisomerases leads to development of a net negative superhelicity in the DNA.

along the DNA and binding to each other, providing a link between distant parts of the chromosome.

SMC proteins have five distinct domains (Figure 9-24a). The amino-terminal (N) and carboxyl-terminal (C) domains each contain part of an ATP-hydrolytic site, and they are connected by two regions of α -helical coiled-coil motifs (see Figure 4-16b) joined by a hinge domain. With bending at the hinge, the N and C domains come together to form a head structure at one end with a complete ATP-binding site. SMC proteins are generally dimeric, forming a V-shaped complex linked through the hinge domains (Figure 9-24b). Thus the dimeric SMC complex contains two head domains and two ATP-binding sites. ATP is not hydrolyzed until the two heads come together. Although

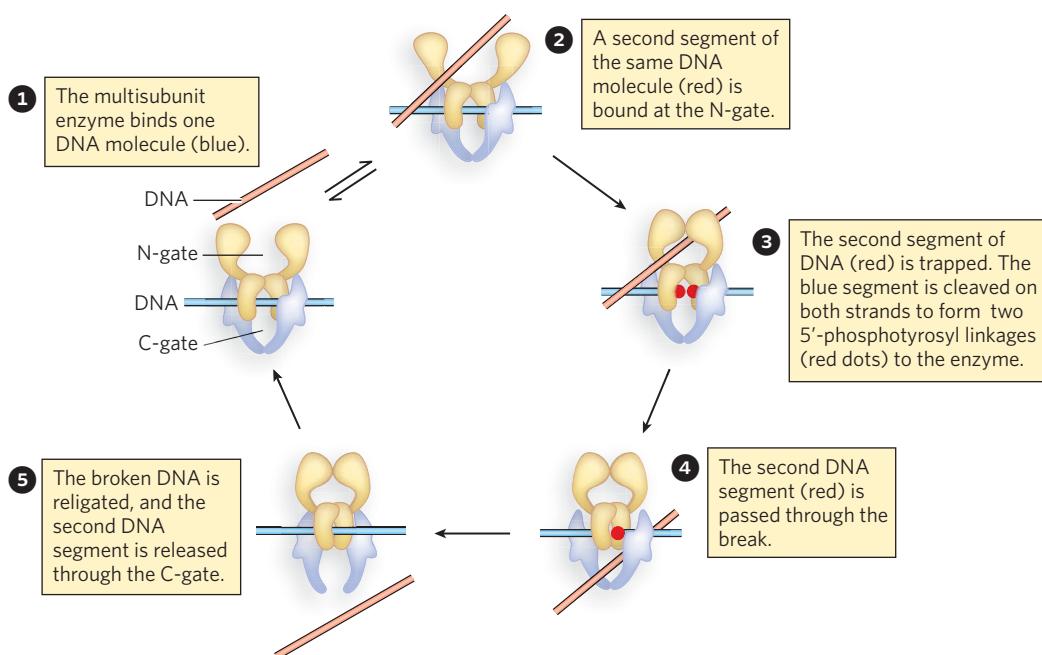


FIGURE 9-23 Alteration of the linking number by eukaryotic type II topoisomerases. The general mechanism is similar to that of the bacterial DNA gyrase (see Figure 9-20b), with one intact duplex DNA segment passed through a transient double-

strand break in another segment. The enzyme structure and use of ATP are distinct to this reaction. See text for details. [Source: Adapted from J. J. Champoux, *Annu. Rev. Biochem.* 70:369, 2001, Fig. 11.]

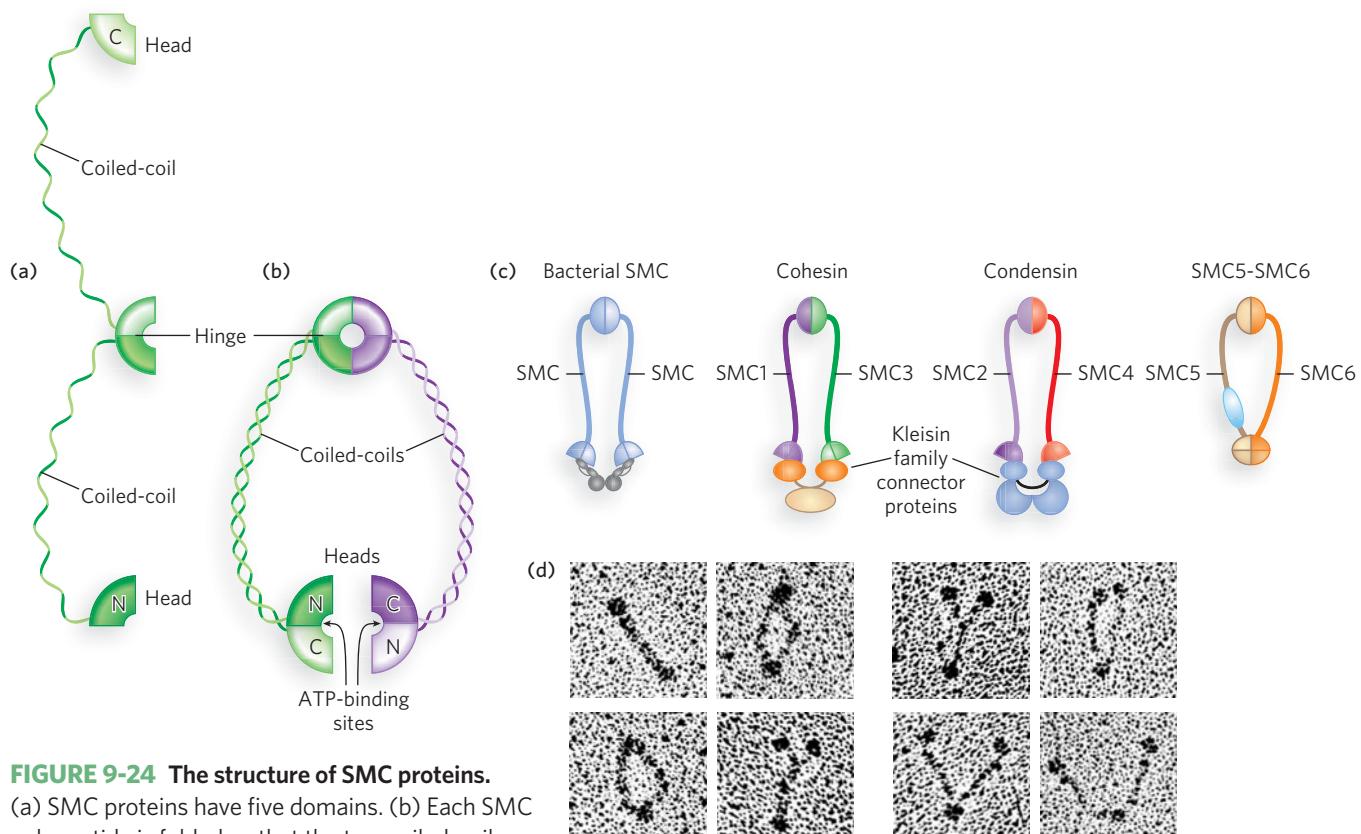


FIGURE 9-24 The structure of SMC proteins. (a) SMC proteins have five domains. (b) Each SMC polypeptide is folded so that the two coiled-coil domains wrap around each other and the N and C domains come together to form a complete ATP-binding site. Two polypeptides are linked at the hinge region to form the dimeric V-shaped SMC molecule. (c) Bacterial SMC proteins form a homodimer. The three different eukaryotic SMC proteins form heterodimers. Cohesins are made up of SMC1-SMC3 pairs, and condensins consist of SMC2-SMC4 pairs. (d) Electron micrographs of SMC dimers. [Source: (d) Courtesy of Harold Erickson.]

HIGHLIGHT 9-2 MEDICINE

Curing Disease by Inhibiting Topoisomerases

The topological state of cellular DNA is intimately connected with its function. Without topoisomerases, cells cannot replicate or package their DNA, or express their genes—and they die. Inhibitors of topoisomerases have therefore become important pharmaceutical agents, targeted at infectious organisms and malignant cells.

Two classes of bacterial topoisomerase inhibitors have been developed as antibiotics. The coumarins, including novobiocin and coumermycin A1, are natural products derived from *Streptomyces* species. They inhibit the ATP binding of the bacterial type II topoisomerases, DNA gyrase and topoisomerase IV. These antibiotics are not used to treat infections in humans, but research continues to identify clinically effective variants.

The quinolone antibiotics, also inhibitors of bacterial DNA gyrase and topoisomerase IV, first appeared in 1962 with the introduction of nalidixic acid (Figure 1). This compound had limited effectiveness and is no longer used clinically in the United States, but the continued development of this class of drugs eventually led to the introduction of the fluoroquinolones, exemplified by ciprofloxacin (Cipro). The quinolones act by blocking the last step of the topoisomerase reaction in bacteria, the resealing of the DNA strand breaks. Ciprofloxacin is a broad-spectrum antibiotic that works on a wide range of disease-causing bacteria. It is one of the few antibiotics reliably effective in treating anthrax in-

fections and is considered a valuable agent in protection against possible bioterrorism. Quinolones are selective for the bacterial topoisomerases, inhibiting the eukaryotic enzymes only at concentrations several orders of magnitude greater than the therapeutic doses.

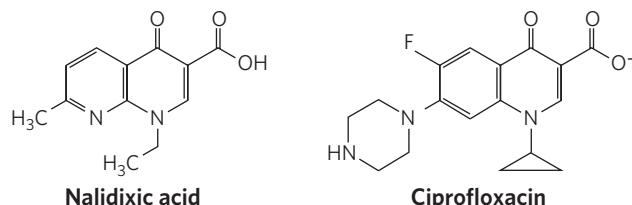


FIGURE 1 Inhibitors of bacterial type II topoisomerases.

Some of the most important chemotherapeutic agents used in cancer treatment are inhibitors of human topoisomerases. Tumor cells generally contain elevated levels of topoisomerases, and agents targeted to these enzymes are much more toxic to the tumors than to most other tissue types. Inhibitors of both type I and type II topoisomerases have been developed as anticancer drugs.

Camptothecin, isolated from a Chinese ornamental tree and first tested clinically in the 1970s, is an inhibitor of eukaryotic type I topoisomerases. Clinical trials indicated limited effectiveness, however, despite its early promise in preclinical work on mice. Two effective derivatives were developed in the 1990s: irinotecan (Campto) and topotecan (Hycamtin), used to treat colorectal cancer and ovarian cancer, respectively (Figure 2). Additional derivatives are likely to be approved for clinical use in the

many details of SMC protein function have yet to be elucidated, the head-head association between the two subunits seems to be critical.

Proteins in the SMC family are found in all types of organisms. All bacteria have at least one SMC protein that functions as a homodimer to assist in compacting the genome, whereas eukaryotes generally have six SMC proteins, functioning in defined pairs as heterodimers with different functions (Figure 9-24c). The SMC1-SMC3 and SMC2-SMC4 pairs have roles in mitosis, and the SMC5-SMC6 pair is involved in DNA repair, but its molecular role is not well understood. All these complexes are bound by regulatory

and accessory proteins. The interactions with DNA involve patches of basic amino acid residues near the hinge regions of the SMC proteins. Electron microscopy reveals the flexible V shape of these proteins (Figure 9-24d).

The SMC1-SMC3 pair forms a functional unit called a **cohesin**. During mitosis, cohesins link two sister chromatids immediately after chromosomal replication and keep them together as the chromosomes condense to metaphase (Figure 9-25). Additional proteins, particularly proteins in the kleisin family such as SCC1, bridge the cohesin head units to form a ring (see Figure 9-24c). The ring wraps around the sister chromatids, tying them

coming years. All of these drugs act by trapping the topoisomerase-DNA complex in which the DNA is cleaved, inhibiting religation.

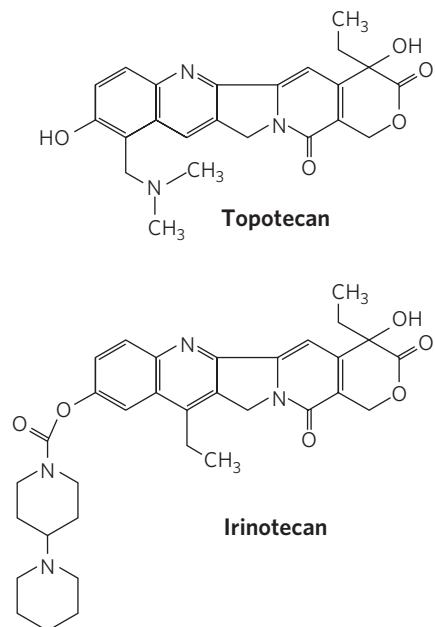


FIGURE 2 Inhibitors of eukaryotic topoisomerase I that are used in cancer chemotherapy.

The human type II topoisomerases are targeted by a variety of antitumor drugs, including doxorubicin (Adriamycin), etoposide (Etopophos), and ellipticine (Figure 3). Doxorubicin, effective against several kinds of human tumors, is in clinical use. Most of these drugs stabilize the covalent topoisomerase-cleaved DNA complex.

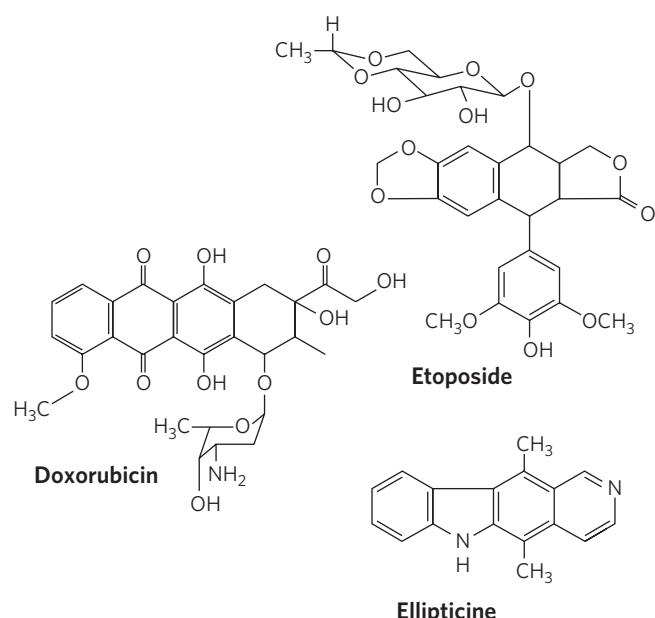


FIGURE 3 Inhibitors of human topoisomerase II that are used in cancer chemotherapy.

All of these anticancer agents generally increase the levels of DNA damage in targeted, rapidly growing tumor cells, but noncancerous tissues can also be affected, leading to a more general toxicity and unpleasant side effects that must be managed during therapy. As cancer therapies become more effective and survival statistics for cancer patients improve, the independent appearance of new tumors is becoming a greater problem. In the continuing search for new cancer therapies, the topoisomerases are likely to remain prominent targets.

together until separation is required at cell division. The ring may expand and contract in response to ATP hydrolysis. As chromosome segregation begins, the cohesin tethers are removed by enzymes known as separases.

Head-to-head engagement of SMC proteins has the potential to produce several different architectures, such as rings, rosettes, and filaments (Figure 9-26). It is not yet clear whether the ringed cohesin tethers around sister chromatids are intra- or intermolecular. The associated proteins may modulate intermolecular interactions, or, for intramolecular rings, they may perform a gatekeeping function in bringing DNA molecules into the ring.

The SMC2-SMC4 complex is called a **condensin**. The bacterial SMC proteins are most closely related to condensins. The condensins are critical to chromosome condensation as cells enter mitosis (see Figure 9-25). In the laboratory, condensins bind DNA to create positive supercoils; that is, condensin binding causes the DNA to become overwound, in contrast to the underwinding induced by the binding of nucleosomes. Figure 9-27 shows a current model of how condensins may interact with DNA to promote chromosome condensation. The condensin complexes (SMC2-SMC4 plus associated proteins) first bind to the DNA in a closed form. ATP hydrolysis then opens the intramolecular ring and

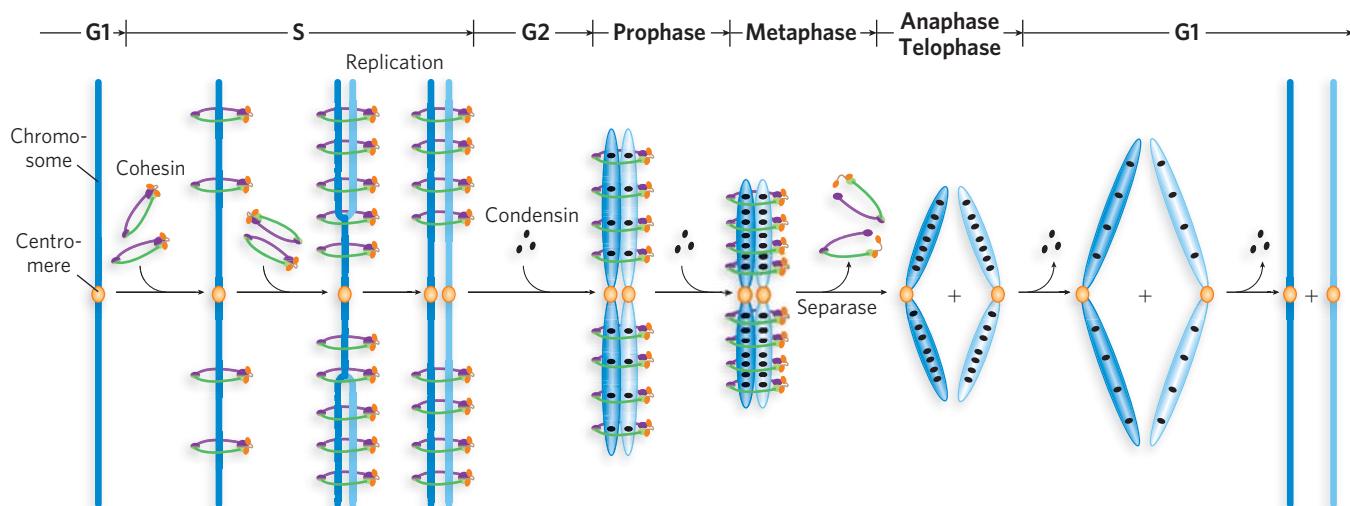


FIGURE 9-25 The roles of cohesins and condensins in the eukaryotic cell cycle. Cohesins are loaded onto the chromosomes during G1 (see Section 2.2), tying the sister chromatids together during replication. At the onset of mitosis, condensins bind and maintain the chromatids in a

condensed state. During anaphase, the enzyme separase removes the cohesin links. Once the chromatids separate, condensins begin to unload and the daughter chromosomes return to the uncondensed state. [Source: Adapted from D. P. Bazett-Jones, K. Kimura, and T. Hirano, *Mol. Cell* 9:1183, Fig. 5.]

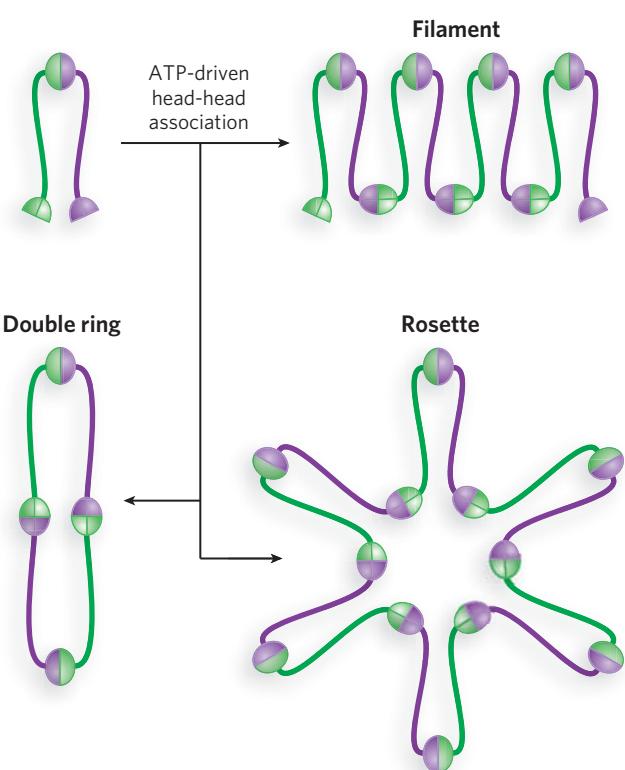


FIGURE 9-26 Potential architectural arrangements of SMC proteins. Head-to-head association results in the formation of ring structures, rosettes, or filaments. [Source: Adapted from T. Hirano, *Nat. Rev. Mol. Cell. Biol.* 7:311–322, 2006.]

brings the DNA inside. Head-to-head association creates a structure that traps DNA with a positive superhelical tension. Finally, aggregation of the condensins into rosettes forms a condensed chromatid with a defined architecture.

The topoisomerases and SMC proteins enable cells to deal with the complex topological changes occurring as DNA strands separate during replication, repair, and transcription, and the extraordinary degree of DNA compaction required in every cell. The compaction is maintained by additional specialized DNA-binding proteins, and we turn to these proteins, and their organization and function, in Chapter 10.

SECTION 9.3 SUMMARY

- Topoisomerases catalyze the underwinding and relaxation of DNA. On a molecular level, topoisomerases catalyze changes in the linking number.
- The two classes of topoisomerases, type I and type II, change Lk in increments of 1 or 2, respectively, per catalytic event.
- The reactions catalyzed by DNA topoisomerases involve the formation of transient covalent DNA-enzyme intermediates, usually in the form of a phosphotyrosyl linkage.
- Bacterial DNA gyrases introduce negative supercoils.