



CRC Press
Taylor & Francis Group

Practical Bioinformatics

Michael Agostino

Practical Bioinformatics

Dedication

This book is dedicated to my mother, Ruth Agostino, who tolerated my smelly biology and chemistry experiments in the basement of our house, and the endless number of muddy clothes and shoes from my frequent explorations of the woods near my home. I owe my love of exploration and discovery to you.

Practical Bioinformatics

Michael Agostino



CRC Press

Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

Front cover image:

Chapter 8 of this book focuses on protein analysis. One example in this chapter is the superimposition of the black swan and Atlantic cod fish lysozyme structures (see Section 8.6). This allows the viewer to see the impact, or lack thereof, of the amino acid differences on the structures of these distantly related proteins. The book cover shows the amino acid sequence of this swan lysozyme (UniProt accession number P00717), repeated many times to fill the page, and is combined with the structure of the lysozyme protein (PDB identifier 1gbs).

About the author:

Michael Agostino received his PhD in Molecular Biology from Roswell Park Memorial Institute, a division of SUNY at Buffalo, New York. His thesis characterized the unusual structure and evolution of sea urchin histone genes. Postdoctoral work included the development of a molecular assay for DNA strand scission agents used in chemotherapy. In 1984, he moved to the University of North Carolina at Chapel Hill where he co-developed a vector trap for gene enhancers. Other work included the creation of a synthetic gene and an *E. coli* blue-white reporter gene assay for HIV protease activity. In 1991 he formally switched careers to bioinformatics by joining GlaxoSmithKline. There, he provided sequence analysis, consulting, user-support, and training for the Glaxo scientists. In 1996 he moved to Genetics Institute, where he was appointed manager of a bioinformatics department. This group was responsible for the sequence analysis and database of a high-throughput effort to identify, express, and patent the human genes that encode secreted proteins. Presently, he provides bioinformatics analysis and end-user support for multiple sites of the Pfizer Research organization. He is also an adjunct professor in the Biology Department at Merrimack College, North Andover, Massachusetts (USA).

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2013 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

ISBN 978-0-8153-4456-8

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
<http://www.taylorandfrancis.com>

and the CRC Press Web site at
<http://www.crcpress.com>

Library of Congress Cataloging-in-Publication Data

Agostino, Michael J.

Practical bioinformatics / Michael Agostino.

p. cm.

ISBN 978-0-8153-4456-8 (alk. paper)

1. Nucleotide sequence--Data processing. 2. Bioinformatics. I. Title.

QP625.N89A39 2013

572.8'6330285--dc23

2012017992

Preface

Although bioinformatics is a relatively new scientific discipline, it has become quite broad in definition. It is often described as including diverse topics such as the analysis of microarrays and the accompanying statistics, protein structure prediction, and pathway and protein interaction analysis. Of course, computer programming, database development, and even hardware design are included in the field. *Practical Bioinformatics* is focused on the fundamental skills of bioinformatics: the analysis of DNA, RNA, and protein sequences. The chapters take the reader through a commonly asked question, “What can I learn about this sequence?” The only requirement is access to the Internet and a web browser; no other software is required.

This book is designed as an introduction to bioinformatics sequence analysis for biology and biochemistry majors. There are many published books that teach about detailed algorithms, sophisticated programs, and advanced interpretation of data. Although these are excellent sources of information, many biologists and biochemists are not prepared for, nor do they need, the depth and detail of these texts. Instead, they need the practical knowledge and skills to analyze sequences. They are asking questions such as “Which tools do I use?” “What settings should I use?” “What database should I search?” “What do these results mean?” “How do I export this information?” *Practical Bioinformatics* addresses these questions, and many more, in 12 easily read chapters.

Concepts will be introduced within each chapter and then demonstrated through the analysis of problems using selected gene/protein examples. Adequate background, details, illustrations, and references will be provided to insure that readers understand the fundamentals and can do further reading if desired. Along the way, interesting genes, phenotypes, mutations, and biology will be introduced but not discussed extensively or analyzed. These topics are purposefully left open so they may easily be turned into literature searches, analysis problems, or senior projects for the ambitious student. Just thinking about these problems and how to analyze them will instill the habit of identifying topics needing exploration.

The best way to learn this material is by “doing.” Readers of this book will learn the concepts by performing many analysis problems. To get the most out of this book, readers should perform most, if not all, of the analysis steps and recreate the figures for themselves. By the time readers finish the book, they will have significant experience in sequence analysis problems, approaches, and solutions. They should then be ready to perform many analysis steps on their own, and tackle more advanced books on the subject.

A common error when approaching a sequence analysis problem is to use powerful analysis software with little understanding of how it works or how to interpret the output. Web forms and software can completely hide the details. This text will emphasize the proper use of established analysis software and the need to evaluate new tools. There are literally hundreds of bioinformatics tools available and no book could possibly contain or instruct on all the tools that are available. However, the repeated experience of performing guided analysis problems will teach the reader to be critical of bioinformatics software and to use proper positive and negative controls when testing unfamiliar tools. When this book is finished, readers will have both the practical knowledge and experience to address their own problems, and take advantage of the mountains of genetic data being generated today.

I would like to thank the staff and associates of Garland Science for their tremendous support during the process of writing this book. Thanks to Gina Almond who believed in the project from the very beginning and was never

short on enthusiasm, David Borrowdale for guiding the book through the many steps, and Mary Purton for her infinite patience during the editing process. My thanks go to Ioana Moldovan, Georgina Lucas, Jo Clayton, and Sally Huish for their tremendous attention to detail and style during the final editing. Special thanks go to Oxford Designers & Illustrators for numerous illustrations. Thanks to Josephine Modica-Napolitano who gave me my first job in teaching, and the students at Merrimack College; together they put me on the path of writing this book. My special thanks to Donald J. Mulcare, my undergraduate advisor, for advice, encouragement, and my first real taste of what it is like to be a scientist. My years in industry would not have been the same without knowing the members of the “Dream Team:” Yuchen Bai, Sreekumar Kodangattil, Ellen Murphy, Padma Reddy, and Wenyan Zhong. They are the best sequence analysts I know. Additional thanks go to Maryann Whitley and Steve Howes for providing a calm and steady leadership at Pfizer. Many thanks to my daughter Becky, who inspires me to be better every day. Finally, this book would not have been possible without the years of support, encouragement, and love from my wife, Nan. Dreams can come true.

Instructor Resources Website

Accessible from www.garlandscience.com, the Instructor Resource Site requires registration and access is available only to qualified instructors. To access the Instructor Resource Site, please contact your local sales representative or email science@garland.com.

The images in *Practical Bioinformatics* are available on the Instructor Resource Site in two convenient formats: PowerPoint® and JPEG, which have been optimized for display. The resources may be browsed by individual chapter or a search engine. Figures are searchable by figure number, figure name, or by keywords used in the figure legend from the book.

Answers to end of chapter questions/exercises are available on the Instructor Resource Site.

Resources available for other Garland Science titles can be accessed via the Garland Science Website.

PowerPoint is a registered trademark of Microsoft Corporation in the United States and/or other countries.

Acknowledgments

The author and publisher of *Practical Bioinformatics* gratefully acknowledge the contributions of the following reviewers in the development of this book:

Enrique Blanco	University of Barcelona, Spain
Ron Croy	Durham University, UK
John Ferguson	Bard College, USA
Laurie Heyer	Davidson College, USA
Torgeir Hvidsten	Umeå University, Sweden
Ian Kerr	University of Nottingham, UK
Daisuke Kihara	Purdue University, USA
Peter Kos	Biological Research Centre of the Hungarian Academy of Sciences, Hungary
Jean-Christophe Nebel	Kingston University London, UK
Samuel Rebelsky	Grinnell College, USA
Rebecca Roberts	Ursinus College, USA
Hugh Shanahan	Royal Holloway, University of London, UK
Shin-Han Shiu	Michigan State University, USA
Shaneen Singh	Brooklyn College, USA
Alan Ward	Newcastle University, UK

Contents

Chapter 1 Introduction to Bioinformatics and Sequence Analysis 1

- 1.1 Introduction 1
- 1.2 The Growth of GenBank 2
- 1.3 Data, Data, Everywhere 2

Further examples of human genome sequencing 4

Personal genome sequencing 4

Paleogenetics 4

Focused medical genomic studies 5

- 1.4 The Size of a Genome 5

- 1.5 Annotation 6

- 1.6 Witnessing Evolution Through Bioinformatics 7

Recent evolutionary changes to plants and animals 7

- 1.7 Large Sources of Human Sequence Variation 7

- 1.8 Recent Evolutionary Changes to Human Populations 8

- 1.9 DNA Sequence in Databases 9

Genomic DNA assembly 10

cDNA in databases—where does it come from? 12

- 1.10 Sequence Analysis and Data Display 14

- 1.11 Summary 20

Further Reading 20

Internet resources 21

Chapter 2 Introduction to Internet Resources 23

- 2.1 Introduction 23

- 2.2 The NCBI Website and ENTREZ 23

- 2.3 PubMed 25

- 2.4 Gene Name Evolution 27

- 2.5 OMIM 29

- 2.6 Retrieving Nucleotide Sequences 30

- 2.7 Searching Patents 31

- 2.8 Public Grants Database: NIH RePORTER 33

- 2.9 Gene Ontology 34

- 2.10 The Gene Database 36

- 2.11 UniGene 38

- 2.12 The UniGene Library Browser 43

- 2.13 Summary 44

Exercises 44

Williams syndrome and oxytocin: research with Internet tools 44

Further Reading 45

Chapter 3 Introduction to the BLAST Suite and BLASTN 47

- 3.1 Introduction 47

Why search a database? 47

- 3.2 What is BLAST? 48

How does BLAST work? 48

- 3.3 Your First BLAST Search 49

Find the query sequence in GenBank 49

Convert the file to another format 51

Performing BLASTN searches 52

- 3.4 BLAST Results 54

Graphic 54

Interpretation of the graphic 55

Results table 55

Interpretation of the table 57

The alignments 57

Other BLASTN hits from this query 60

Simultaneous review of the graphic, table, and alignments 63

- 3.5 BLASTN Across Species 64

BLASTN of the reference sequence for human beta hemoglobin against nonhuman transcripts 64

Paralogs, orthologs, and homologs 66

- 3.6 BLAST Output Format 68

- 3.7 Summary 68

Exercises 68

Exercise 1: Biofilm analysis 68

Exercise 2: RuBisCO 70

Further Reading 71

Internet resources 71

Chapter 4 Protein BLAST: BLASTP 73

- 4.1 Introduction 73

- 4.2 Codons and the Genetic Code 73

Memorizing the genetic code 76

- 4.3 Amino Acids 76

Amino acid properties 77

4.4	BLASTP and the Scoring Matrix	78
	Building a matrix	78
4.5	An Example BLASTP Search	80
	Retrieving protein records	81
	Running BLASTP	81
	The results	82
	The alignments	84
	Distant homologies	84
4.6	Pairwise BLAST	85
4.7	Running BLASTP at the ExPASy Website	86
	Searching for pro-opiomelanocortin using a protein sequence fragment	87
	Searching for repeated domains in alpha-1 collagen	91
4.8	Summary	94
	Exercises	94
	Exercise 1: Typing contest	94
	Exercise 2: How mammoths adapted to cold	95
	Exercise 3: Longevity genes?	96
	Further Reading	97

Chapter 5 Cross-Molecular Searches: BLASTX and TBLASTN 99

5.1	Introduction	99
5.2	Messenger RNA Structure	100
5.3	cDNA	101
	Synthesis	101
	cDNA in databases	102
	ESTs	103
	Normalized cDNA libraries	104
	An EST record	106
5.4	BLASTX	107
	Reading frames in nucleic acids	107
	A simple BLASTX search	108
	A more complex BLASTX	109
	Using the annotation of sequence records	115
	BLASTX alignments with the reverse strand	117
5.5	TBLASTN	117
	A TBLASTN search	118
	Metagenomics and TBLASTN	120
5.6	Summary	122
	Exercises	122
	Exercise 1: Analyzing an unknown sequence	122
	Exercise 2: Snake venom proteins	123
	Exercise 3: Metagenomics	124
	Further Reading	125

Chapter 6 Advanced Topics in BLAST 127

6.1	Introduction	127
6.2	Reciprocal BLAST: Confirming Identities	127
	Demonstration of a reciprocal BLASTP	128
6.3	Adjusting BLAST Parameters	131
	Gap cost	131
	Compositional adjustments	133
6.4	Exon Detection	134
	Exon detection with BLASTN	135
	Look at the coordinates	138
	Exon detection with TBLASTN	138
	Orthologous exon searching with TBLASTN	141
6.5	Repetitive DNA	144
	Simple sequences	145
	Satellite DNA	145
	Mini-satellites	145
	LINEs and SINEs	145
	Tandemly arrayed genes	146
6.6	Interpreting Distant Relationships	147
	Name of the protein	147
	Percentage identity	148
	Alignment length and length similarity between query and hit	148
	E value	149
	Gaps	149
	Conserved amino acids	150
6.7	Summary	152
	Exercises	152
	Exercise 1: Simple sequences	152
	Exercise 2: Reciprocal BLAST	153
	Exercise 3: Exon identification with TBLASTN	153
	Exercise 4: Identification of orthologous exons with TBLASTN	154
	Further Reading	155

Chapter 7 Bioinformatics Tools for the Laboratory 157

7.1	Introduction	157
7.2	Restriction Mapping and Genetic Engineering	158
	Restriction enzymes	158
	Restriction enzyme mapping: the polylinker site	160
	NEBcutter	160
	Generating reverse strand sequences: Reverse Complement	162
	DNA translation: the ExPASy Translate tool	162

7.3	Finding Open Reading Frames	163
	The NCBI ORF Finder	163
7.4	PCR and Primer Design Tools	165
	Primer3	166
	Primer-BLAST	169
7.5	Measuring DNA and Protein Composition	170
	DNA Stats	170
	Composition/Molecular Weight Calculation Form	171
7.6	Asking Very Specific Questions: The Sequence Retrieval System (SRS)	172
7.7	DotPlot	174
	DotPlot of alternative transcripts	175
	DotPlots of orthologous genes	176
7.8	Summary	179
Exercises		179
	Spider silk: a workflow of analysis	179
Further Reading		181

Chapter 8 Protein Analysis 183

8.1	Introduction	183
8.2	Finding Functional Patterns	183
	A repeating pattern within a zinc finger	184
8.3	Annotating an Unknown Sequence	187
	A zinc protease pattern	188
	The ADAM_MEPRO profile	188
8.4	Looking at Three-dimensional Protein Structures	190
	Jmol: a protein structure viewer	192
	Exploring and understanding a structure	193
	Jmol scripting	194
8.5	ProPhyIER	195
	The Interface view	196
	The CrystalPainter view	198
8.6	The Impact of Sequence on Structure	201
8.7	Building Blocks: A Multiple Domain Protein	204
8.8	Post-translational Modification	204
	Secretion signals	206
	Prediction of protein glycosylation sites	208
8.9	Transmembrane Domain Detection	208
8.10	Summary	211
Exercises		211
	Aquaporin-5	211
Further Reading		213
Internet resources		214

Chapter 9 Explorations of Short Nucleotide Sequences 215

9.1	Introduction	215
9.2	Transcription Factor Binding Sites	216
	Transfac	216
	Identifying other binding sites for the estrogen receptor	219
	Predicting transcription factor binding sites	220
	An experiment with MATCH	221
	An experiment with PATCH	224
9.3	Translation Initiation: The Kozak Sequence	226
9.4	Viewing Whole Genes	228
9.5	Exon Splicing	231
	Renin: a striking example of a small exon	234
	Another striking splice: human <i>ISG15</i> ubiquitin-like modifier	235
	Alternative splicing	236
	Human plectin: alternative splicing at the 5P end	237
	Consensus splice junctions, translated	238
9.6	Polyadenylation Signals	239
9.7	Summary	240
Exercises		242
	Inhibitor of Kappa light polypeptide gene enhancer in B-cells (<i>IKBKAP</i>)	242
Further Reading		243

Chapter 10 MicroRNAs and Pathway Analysis 245

10.1	Introduction	245
10.2	miRNA Function	245
10.3	miRNA Nomenclature	247
10.4	miRNA Families and Conservation	247
10.5	Structure and Processing of miRNAs	248
10.6	miRBase: The Repository for miRNAs	250
10.7	Numbers and Locations	251
10.8	Linking miRNA Analysis to a Biochemical Pathway: Gastric Cancer	251
10.9	KEGG: Biological Networks at Your Fingertips	253
	miRNAs in the cell cycle pathway	255
10.10	TarBase: Experimentally Verified miRNA Inhibition	256
	Verified miRNA-driven translation repression	256

10.11 TargetScan: miRNA Target Site Prediction	258	12.3 Synteny	303
TargetScan predictions for cell cycle transcripts	260	Synteny of the sex chromosomes	304
10.12 Expanding miRNA Regulation of the Cell Cycle Using TarBase and TargetScan	263	12.4 The UCSC Genome Browser	304
10.13 Making Sense of miRNAs and Their Many Predicted Targets	265	<i>OPN5</i> : a sample gene to browse	305
10.14 miRNAs Associated With Diseases	266	Simple view changes in the UCSC Genome Browser	308
10.15 Summary	267	Configuring the UCSC Genome Browser window	310
Exercises	267	Searching genomes and adding tracks through BLAT	312
GDF8	267	Viewing the Multiz alignments	314
Further Reading	269	Zooming out: seeing the big picture	316
		Very large genes: dystrophin and titin	318
		Gene density	320
		Interspecies comparison of genomes	323
		The beta globin locus	324
Chapter 11 Multiple Sequence Alignments	271	12.5 Summary	325
11.1 Introduction	271	Exercises	325
11.2 Multiple Sequence Alignments Through NCBI BLAST	271	Olfactory genes	325
11.3 ClustalW from the ExPASy Website	274	Further Reading	327
11.4 ClustalW at the EMBL-EBI Server	276		
MARK1 kinase	277	Appendix 1 Formatting Your Report	329
MAPK15 kinase	280	A1.1 Introduction	329
DNA versus protein identities	282	A1.2 Font Choice and Pasting Issues	329
11.5 Modifying ClustalW Parameters	282	A1.3 Find and Replace	331
Gap-opening penalty	282	Changing file format	332
The clustering method	283	A1.4 Hypertext	333
11.6 Comparing ClustalW, MUSCLE, and COBALT	286	Creating hypertext	334
11.7 Isoform Alignment Problem: Internal Splicing	288	Selecting a column of text	334
11.8 Aligning Paralog Domains	292	A1.5 Summary	334
11.9 Manually Editing a Multiple Sequence Alignment	294	Appendix 2 Running NCBI BLAST in “batch” Mode	337
Jalview	294		
Editing with a word processor	296	Abbreviations	340
11.10 Summary	296	Glossary	341
Exercises	296	Web Resources	344
FOXP2	296	Index	347
Further Reading	297		
Chapter 12 Browsing the Genome	299		
12.1 Introduction	299		
12.2 Chromosomes	299		
Human chromosome statistics	300		
Chromosome details and comparisons	302		



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

ACAAGGGACTAGAGAAACCAAAA

AGAAACCAAAACGAAAGGTGCAGAA

AACGAAAGGTGCAGAAGGGGAAACAGATGCAGA

GAAGGGGAAACAGATGCAGAAAGCATC

AGAAAGCATC

ACAAGGGACTAGAGAAACCAAAACGAAAGGTGCAGAAGGGGAAACAGATGCAGAAAGCATC

Introduction to Bioinformatics and Sequence Analysis

AGAAACCAAAACGAAAGGTGCAGAA

AACGAAAGGTGCAGAAGGGG

GAAGGGG

Key concepts

- The scope of bioinformatics
- The origins and growth of DNA databases
- Evidence of evolution from bioinformatics
- Example sequence analysis and displays using human Factor IX

1.1 INTRODUCTION

We are witnessing a revolution in biomedical research. Although it has been clear for decades that exploring the genetics of biological systems was crucial to understanding them, it was far too expensive and complex to consider obtaining genetic sequences for that exploration. But now, acquiring genetic sequences is affordable and simple, and data are being generated at unprecedented rates. The heart of understanding all this sequence lies in bioinformatics sequence analysis, and this book serves as an introduction to this powerful study of DNA, RNA, and protein sequence.

Bioinformatics concerns the generation, visualization, analysis, storage, and retrieval of large quantities of biological information. The generation of biomedical data, including DNA sequence, in its raw form does not involve bioinformatics skills. But in order for that sequence to be usable, it must be analyzed, annotated, and reformatted to be suitable for databases. These are all bioinformatics activities. Many of these activities can be automated, but their development and support come from someone with skills or experience in bioinformatics.

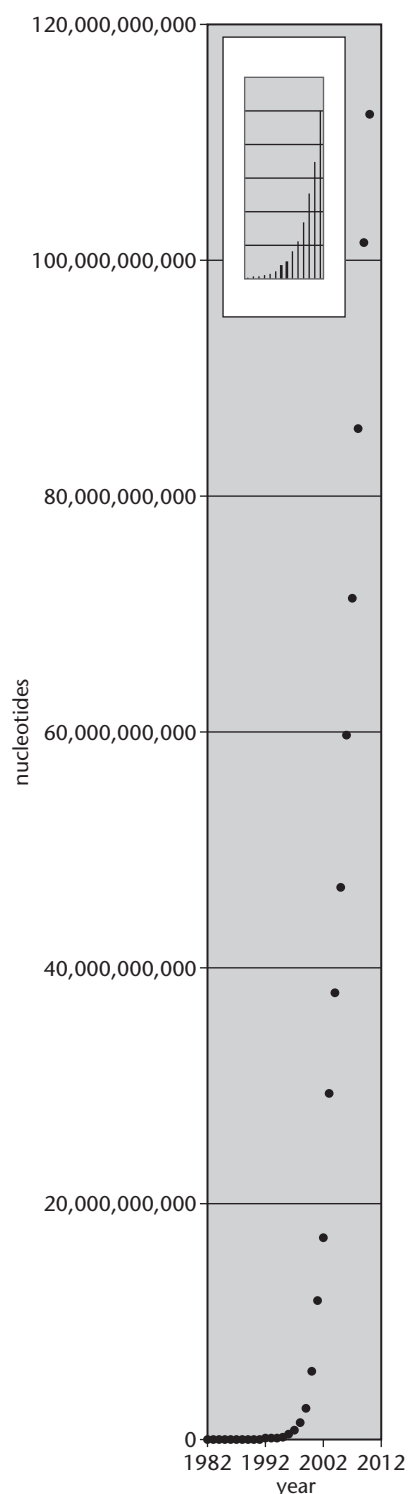
Once the data have been made available, how do you analyze the data? Is there text like DNA and protein sequence files? If yes, it should be presented in a way to allow interpretation or easy input into programs for analysis. Or is there so much information that data are represented graphically? This form of data reduction is quite powerful and without it we would be staring at pages and pages of sequence without, literally, seeing the big picture.

Some analysis is manual, ranging from looking at the individual nucleotides or amino acids, to submitting sequence to a program that transforms the sequence into another form. This could include the location of features such as functional domains, modification sites, and coding regions. Often, analysis includes the searching of databases for the purposes of comparison or discovery, and this will be the primary activity for a number of chapters. Much of the content of this book is concerned with analysis.



Floppy disk databases

In the early years of GenBank, if you wanted access to the database you ordered a handful of floppy disks that were delivered in the mail.



Storage is usually not a responsibility of those who will analyze the sequences. However, the creation of properly structured databases or storage forms so data can be queried and retrieved is essential for the analysts to do their work. Sequence files and other forms of data can be decades old or just created yesterday. But unless you can retrieve them easily, the value decreases quickly. “Easily” is not just describing the speed of the computers and connections delivering the information to you, although this can be extremely important. It also includes the steps to access and query the stored data. The ideal approach is often a Web form with easily understood options, online help, and results pages rich with hypertext. Bioinformatics was one of the first areas of science to embrace the Web as a vehicle for disseminating information and we’ll be using many Web pages in this book.

Finally, bioinformatics activities often involve large quantities of data. Even if you are focusing on a single gene, you still may have mountains of data that are connected to this single sequence. With a good database or software tool, you may only be aware of the quantities yet not overwhelmed with details that don’t interest you. Still, it can’t be emphasized enough that one of the biggest challenges facing the field of bioinformatics is the absolute deluge of information and how to generate, visualize, analyze, store, and retrieve these data.

1.2 THE GROWTH OF GenBank

How much data are we talking about? One way to answer this is to describe the amount of DNA sequence data in public databases. GenBank is a huge repository run by the US National Center for Biotechnology Information (NCBI). The inset in **Figure 1.1** shows the steady growth in the early years of GenBank but the rate of growth has been rapid since then. As of early 2011, there are over 126 billion nucleotides in this standard division of GenBank from over 380,000 organisms. If this were not impressive enough, there are an additional 91 billion nucleotides in the whole genome shotgun (a type of sequencing) division, the section of GenBank dedicated to unfinished large sequencing efforts. If a DNA sequence is considered “public information,” it is deposited in GenBank, the DNA Data Bank of Japan (DDBJ), or the database of the European Bioinformatics Institute (EBI). The contents of these three databases are synchronized. In terms of disk space, the database is over 500 Gigabytes in size.

1.3 DATA, DATA, EVERYWHERE

Where are all the data coming from? The quick answer is everywhere! In recent years there has been a dramatic drop in prices and rapid advances in both sequencing technology and computing power. What was once too time-consuming and expensive is now very possible and affordable; biological sequence generation is now commonplace. A major driver for the advances being realized today is the **Human Genome Project**. Even though the completion of the sequencing of the human genome was announced in 2001, the analysis of the data is ongoing and will take many years. These advances had to be coupled with dramatic improvements in computers and the drop in cost for processing power, memory, and storage. Of course, the Human Genome Project and all the spin-offs are only possible because of simultaneous advances in bioinformatics.

This intersection of sequencing technologies, computational power, and advances in bioinformatics has made DNA sequencing quite routine and paved the way for many bold and ambitious projects. Projects now come from scientists

Figure 1.1 GenBank growth. Plotted is the size of GenBank in nucleotides versus the years from 1982 to the first three months of 2011. The inset shows data for years 1982–1994, not visible on the larger plot. From the GenBank Release Notes of Release 184, ncbi.nlm.nih.gov.

in numerous fields of biology, medicine, agriculture, ecology, history, energy, and forensics, just to name a few. Here are some prominent examples.

- The 1000 Genomes Project (www.1000genomes.org). An effort to sequence the genomes of 1000 people to identify genetic variants that affect 1% of the human population. In addition to providing insights to genetic disorders and health risks, the history of human migrations is being revealed. In recent years, people have proposed that the number of human genomes to sequence for this project grow to be 10,000 or higher.
- The 1001 *Arabidopsis thaliana* Genomes Project (www.1001genomes.org). *Arabidopsis* is a widely used plant model due to its habitat diversity, genetics, and ease of manipulation. This genome project aims to study the genomes of 1001 strains that differ in phenotype including adaptation to growth in a wide variety of conditions. Project scientists and those in the *Arabidopsis* community are able to grow huge numbers of genetically identical plants and can vary the environment at will to challenge and observe the underlying genetic elements which define these strains.
- The Genome 10K Project (genome10k.soe.ucsc.edu). An effort to sequence the genomes of 10,000 vertebrate species, one from every genus. Along with all the other genomes sequenced, this project will make a tremendous impact on understanding the relationship between organisms. We can only guess what will be discovered from these animals, having so much in common with us but with such diverse physiologies and phenotypes, and occupying such a wide range of habitats.
- The i5k Initiative (www.arthropodgenomes.org/wiki/i5K). An effort to sequence the genomes of 5000 insects and arthropods. Many insects are either pests, carriers of disease, or beneficial to agriculture and man. More knowledge of their biochemical pathways will surely result in new avenues of control, utilization, and fascination.
- **Metagenomics.** This is a broad term covering the sequencing of DNA samples from the environment as well as from biomedical sources. For example, sequencing has led to the identification of the hundreds of bacterial species inhabiting our skin, mouth, and digestive system. The populations that live on and within us vary with our health state and are clearly linked to our physiology (as we are to theirs). The NCBI lists almost 350 metagenomics projects (www.ncbi.nlm.nih.gov/genomes/lenvs.cgi) that are either at the beginning stages or completed. These projects each generate anywhere from thousands to millions of sequences.
- Cancer Genome Atlas. This is a massive project (cancergenome.nih.gov) where thousands of specimens from all the major cancer types and their matched normal controls will have their RNAs and many of their genes sequenced.
- EST generation. **ESTs** (expressed sequence tags) are small samples of transcribed genes and a quick avenue for discovering the genes expressed in tissues or organisms. Clones are generated and sequenced by the thousands. There are at least 72 million EST sequences in GenBank.
- The Barcode of Life (www.barcodeoflife.org). Distinguishing closely related species is often difficult, even for taxonomy experts. For example, there are approximately 11,000 species of ants. How can you easily tell them apart? The Barcode of Life project aims to identify a DNA “signature” for each species in the world using a 648 base pair sequence of the cytochrome c oxidase 1 gene. The five-year goal is to have sequences from 500,000 species. Nice examples of consumer use of this information include the identification of illegal fishing of endangered species and illegal logging activities.
- The NCBI lists over 1700 eukaryotic genome sequencing projects (www.ncbi.nlm.nih.gov/genomes/leuks.cgi), over 11,000 microbial genome projects (www.ncbi.nlm.nih.gov/genomes/lproks.cgi), and over 3100 viral genomes.



Tumbling costs

According to Eric Lander, director of the Genome Biology Program at the Broad Institute, it now costs about \$20 to sequence the *Escherichia coli* genome, sequencing each of the 4.7 million nucleotides twenty times to ensure accuracy.



Keep flossing

Human microbiome studies have sampled bacteria from skin from all over your body, the gut flora (of course), even your navel (nicknamed “bellybutton biodiversity”). According to a *Nature Reviews Microbiology* Editorial, dental plaque is very dense with bacteria. The number of bacteria in a single gram is equivalent to the number of people who have ever lived.

There are also private sequencing efforts where the data are not always released to the public yet the parties acquiring the data still have to cope with the huge amount of sequence generated by these projects.

- Firms such as pharmaceutical and biotechnology companies are contracting other companies to generate sequence from patients, animals, important crops, plants, cell lines, tumors, and pathogens. They are also doing **deep sequencing** of complementary DNA (cDNA) libraries to identify rarely expressed genes. These efforts are being used to develop products such as new drugs, crops, and diagnostic kits.
- In response to an infectious disease, the genomes of suspected pathogens are being sequenced. For example, in 2011 there was a major pathogenic *Escherichia coli* outbreak in Europe that eventually killed several dozen people. In 2010 there was a cholera outbreak in Haiti following the devastating earthquake there. In both cases, the genomes of the causative bacteria were sequenced to better understand the pathogens and learn how to treat the diseases. Literally tens of thousands of human immune deficiency virus (HIV) genomes have been sequenced. As the price drops, medical sequencing will probably become more commonplace for diagnosis of individuals in the general population.

There are many “smaller” projects that are contributing to the public data growth. There is a division of the NCBI Website (www.ncbi.nlm.nih.gov/popset) that only contains population studies (PopSet): collections of sequences from many members of the same species. For example, there are PopSets for spiders (102), rabbits (179), squirrels (83), skunks (94), robins (114), and ants (94). Within these records you can find the sequence of a single gene from hundreds or thousands of individuals.

Smaller still in size, but not importance, are the efforts to understand a single gene or gene family. This analysis, often originating in an individual laboratory or academic department, is often very detailed and associated with publications. These analysis studies are at the heart of understanding how genes function. Many automated annotation efforts absolutely depend on these manual and long-term projects to serve as reference sequences.

Further examples of human genome sequencing

Personal genome sequencing

Families with a common last name have often cooperated to establish links by common ancestors. Now some are using sequencing from the Y chromosome (inherited from father to son), mitochondria (passed from mothers to their children), or both with the specific purpose of establishing or verifying these links. Some have already uncovered unknown connections between families that would not have been possible to identify without the DNA sequence. Companies have formed that specialize in these kinds of sequence analysis services. They can provide partial family histories for adoptees, provide information concerning paternity, and even identify the presence of the so-called “warrior gene” (*MAOA*), a gene variant associated with aggressive responses to threats.

There are companies that offer the sequencing of your entire genome and the accompanying analysis as a service. As the cost comes down (estimated to drop as low as \$1000 per genome) and the predictive value of genes goes up, you can expect more people to have their genomes sequenced.

Paleogenetics

This is a relatively new field, made possible by vast improvements in the isolation and amplification of DNA from ancient biological specimens. Scientists are now able to ask genetic questions of ancient times in history. For example, Schuenemann and colleagues sequenced DNA from the remains of people who were fourteenth century victims of the Black Death that swept through Europe.



A long journey

In 2011, there were surprising reports of a mountain lion being seen in Connecticut, not the current habitat of these large cats. Shortly after these reports, a 140-pound male mountain lion was struck and killed by a car on a Connecticut highway. For the preceding several years, scientists using the DNA found in scat and hair samples had been tracing its movement from South Dakota, Minnesota, and Wisconsin, making the journey to Connecticut of at least 1500 miles.

Their work shows that *Yersinia pestis*, the agent most probably causing the Black Death, was present but is a different strain to the one found today.

A spectacular display of **paleogenetics** is the sequencing of the Neanderthal genome. The DNA was obtained from bones thousands of years old and carefully sequenced by Richard E. Green and colleagues in the laboratory of Svante Pääbo. The analysis of these data has just begun but has already yielded interesting findings about our ancient relatives. Early work examined the language gene, *FOXP2*, investigating their ability to speak. It was also discovered that Neanderthals had a *MCR1* gene variant that leads to red hair. Comparisons between our genome and that of the Neanderthal reveal that approximately 2.5% of Neanderthal DNA sequence is in our genome, indicating that our ancestors interbred. Very recently, a 41,000-year-old bone from a new human ancestor (**hominin**) was discovered in Siberia and their genome indicates that they contributed a small amount of sequence to present-day Melanesians (people of the islands northeast of Australia).

Focused medical genomic studies

Genetic testing is a well-established hospital procedure for carrier or prenatal testing, diagnostics, and newborn screening for common genetic disorders. The formation of companies that specialize in gene sequencing for establishing genetic risks has some parties concerned that testing without reason or access to qualified counseling can lead to fear or poorly informed life decisions. Testing positive does not mean that you definitely have or will develop a disorder and a negative test does not guarantee that you will not develop the disorder. Others are concerned that disclosure of a positive test to an employer or insurance company may lead to negative consequences.

There are a number of studies in which patients had their DNA sequenced and analyzed for the purpose of identifying the molecular basis of genetic disorders. In the laboratory of David Galas, genome sequences were obtained from the parents and their two children who inherited separate and different recessive genetic disorders. One child was born with Miller syndrome, which causes facial and limb development abnormalities, and the other child was born with primary ciliary dyskinesia. The latter is characterized by the malfunction of microscopic cilia in the respiratory tract. Through careful analysis of the sequences, the disorders were narrowed down to four possible genes. Another finding from this study was an accurate measurement of the mutation rate per generation. Each child in this family was born with 70 mutations (sequences different from either parent), which was lower than that estimated for human generations using other methods.

In another study, a scientist, James Lupski, used his own DNA to identify the molecular basis of the disease Charcot-Marie-Tooth neuropathy that affected him and other members of his family. By sequencing his own genomic DNA, candidates for the cause of his disease were identified. More directed sequencing of the DNA of family members confirmed the mutations responsible for his family's disorder.

In a final example, newspaper reporters received a Pulitzer Prize for an article describing a team effort at The Medical College and Children's Hospital of Wisconsin where the genome sequence of a sick child was determined to assist in the diagnosis of his unusual disease. The Medical College has started a program where physicians can nominate medical cases where knowing the patient's genomic sequence may help, and at least six patients are in the queue to have their DNA sequenced.

1.4 THE SIZE OF A GENOME

How much data is generated when a genome is sequenced? The genome size and gene number generally increase with the complexity of the organism, but there are some surprises. *E. coli*, the object of research for decades and resident in our

Table 1.1 The size of genomes

Species	Genome size (10 ⁶ nucleotides)	Number of genes
<i>Escherichia coli</i>	4.7	4300
<i>Saccharomyces cerevisiae</i>	12	6700
<i>Drosophila melanogaster</i>	169	13,900
<i>Danio rerio</i>	1500	26,000
<i>Homo sapiens</i>	3200	21,000
<i>Zea mays</i>	3200	63,000
<i>Oryza sativa</i>	488	57,000

Source: The Ensembl Genome Browser (www.ensembl.org) April 2012.

digestive system, has 4300 genes in 4.7 million nucleotide pairs. *Saccharomyces cerevisiae*, a single-celled yeast used in cooking and fermentation, has an incrementally larger genome and number of genes (Table 1.1).

Multicellular organisms show an increase in these numbers. The common fruit fly, *Drosophila melanogaster*, has almost 14,000 genes in 169 million base pairs. The genome of the vertebrate zebrafish, *Danio rerio*, is almost tenfold larger, yet only contains 26,000 protein-coding genes. The human genome is approximately 3.2 billion nucleotides long and contains approximately 21,000 protein-coding genes and at least 12,000 noncoding genes. Each mammal genome sequenced in the projects listed above will generate approximately the same amount of partially processed data as seen in human genome analysis.

Plants have complex genomes, reflecting a history of genetic duplications that far surpass the number seen in vertebrates. As a result they often have large genomes and gene numbers; maize (*Zea mays*) has 63,000 genes in a genome of size comparable to that of mammals while rice (*Oryza sativa*) has over 57,000 genes in a much smaller genome.

1.5 ANNOTATION

Of course, if all we had were file after file of just DNA sequence, we would learn little about the object of our sequencing efforts. The true value is realized when the DNA or protein sequence is described to tell us about genetic or protein elements, structures, similarities, functions, and predictions associated with these sequences. Collectively, these details are referred to as **gene annotation**. Like bioinformatics, annotation is a broad term and has different meanings to different people. Here, it is used to describe details such as where a gene starts and ends; similarities to other genes and proteins based on database searches; places that are known to vary; translation start and stop sites; places where the protein is predicted to be or is modified; association with a phenotype or disorder; and ties to other analysis or publications.

Annotation efforts are a big part of any genome or gene project and, depending on the size of the project, can be either manual or automated. When a genome sequence is finished, the annotation of the hundreds or thousands of genes has to be automated. Bioinformatics experts join together a “**pipeline**” of software tools that systematically analyzes each region of the genome, identifies genes, and then determines the details of those genes. The fields of bioinformatics and gene analysis would be at a near standstill without these pipelines, and this form of analysis is both powerful and very accurate.

An automated process cannot perform every conceivable analysis, however. The developers of the pipeline choose the questions to be asked and this analysis



Economic impact of the Human Genome Project

The Human Genome Project cost the US government \$3.8 billion yet the return on that investment has been incredible. According to a report by Battelle Technology Partnership Practice, the breakthroughs in technology and information spawned the birth and growth of both companies and academic laboratories, followed by the creation of products and services. In 2010 alone, this generated \$67 billion in US output, supporting 310,000 jobs and \$20 billion in personal income. Since the Human Genome Project started, over \$49 billion in taxes have been paid to the US government from these genomic-related activities.

provides valuable, but basic, information about these newly discovered genetic elements. Automated efforts will miss details that the software is not trained to recognize. Importantly, automated annotation is not always updated. Annotations entered in a database when a gene is newly discovered may never be updated. If other members of this gene family later appear in the database, there may be no link between the older sequence and these new, more fully described sequences. The description on the older file may be frozen in time.

1.6 WITNESSING EVOLUTION THROUGH BIOINFORMATICS

In the history of life, there have been countless times when a gene's sequence has randomly mutated with a concomitant change in the encoded protein structure and function. Some of these new functions imparted advantages to the organism and were retained for future generations. Deleterious mutations were quickly eliminated from the population. Other changes were neutral and, because they caused no harm, may or may not have been retained. Genes have been duplicated again and again, with each copy continuing to evolve, leading to large gene families and new functions. The path from unicellular to multicellular organisms, and the development of tissue, organs, and limbs, also increased genetic complexity, visible today in higher organisms. Throughout this book there will be numerous demonstrations where the fields of genomics and bioinformatics will show these steps in evolution.

Recent evolutionary changes to plants and animals

About 10,000 years ago, humans began to change from a hunter-gatherer lifestyle to practicing agriculture. Seeds were collected and kept from consumption for planting in the ground in the vicinity of their dwellings. By selecting seeds of plants with superior characteristics, ancient varieties of plants grew taller, produced bigger seeds, produced more nuts or fruit, and resisted inclement weather or disease. We can barely recognize the ancestral plants because this selection process has transformed their appearance so dramatically. However, their DNA sequence reveals the evolution.

The same applies to domesticated animals. Recent sequencing of the dog genome reveals their origins from wolves and places the time of domestication much earlier than plant domestication. Over time, we have transformed dogs into breeds with strikingly different phenotypes. They are all clearly dogs, but the size range, accomplished by careful breeding by humans, is astonishing. An adult Chihuahua weighs no more than 6 pounds and can be as small as six inches high, while a Saint Bernard can reach 180 pounds. Just based on weight, this variation is equivalent to a small human newborn and an adult man.

Other animals have also been bred for specific traits: cows (increased milk production), horses (speed or strength), sheep (wool quantity and quality), poultry (more breast meat), and fish (speed of maturation). Through the sequencing and study of their genes, genetic screening and manipulation may prove to be a more direct route to desired phenotypes. These studies are taking place now.

1.7 LARGE SOURCES OF HUMAN SEQUENCE VARIATION

One of the contributing reasons for the sharp decline in the cost of sequencing the human genome is that the first sequence to be obtained stands as a template to guide the assembly and analysis of subsequent genomes. These newer, so-called **re-sequencing** efforts do not require the many weeks of assembly and problem solving seen during the first genome sequencing. However, there are still considerable differences seen between individual people.



The Japanese Warrior Crab

The Japanese Warrior Crab (*Heikea japonica*) has on its back an uncanny resemblance to an artistic portrait of a Samurai warrior. Over hundreds of years, any captured crabs not looking like a warrior were kept for the market (and no longer reproduced) while those resembling the warriors were returned to the sea.

First, there are **single nucleotide polymorphisms (SNPs)**. The entire human genome is approximately 3.2 billion nucleotide pairs long, and there are approximately 3 million nucleotides that differ when you compare the genomes of two people. These common differences are found in about 1% of the population. Many of these differences have no apparent impact on the function of the genome, while others disrupt gene regulation, or change coding regions resulting in altered amino acid sequences. People studying the genomes of tumors find many SNPs arise within the tumor. Some of these may be responsible for the cancer state, while others accumulate independently of the biochemical changes necessary to become a cancer cell.

There are also tremendous differences between genomes due to **copy number variations (CNVs)**. Comparing your DNA sequence to that of the human “standard” genome, there are thousands of DNA segments which range from 1000 to several million nucleotides in length, and they are either present, present in multiple copies, or absent from your genome.

Kimberly Pelak and colleagues did a fascinating study published in 2010 where they sequenced the genomes of 20 people, 10 of which had hemophilia A. Although they were faced with the many differences between individuals, they were able to identify the mutations causing hemophilia in 6 of the 10 patients. Surprisingly, they found that “on average, each genome carries 165 homozygous protein-truncating or stop loss variants in genes representing a diverse set of pathways.” Of the 21,000 protein-coding genes, almost 0.8% of our genes are unable to be translated to full-length proteins, essentially “knocking out” many of these genes.

1.8 RECENT EVOLUTIONARY CHANGES TO HUMAN POPULATIONS

Since the emergence from Africa, humans have migrated to all continents except Antarctica (**Box 1.1**).

Box 1.1 The author's DNA

My mother's ancestors walked out of Africa perhaps 50,000 years ago. I don't know their exact path, or how they got across rivers or mountains, or survived winters. For countless generations they fanned out from the Middle East across Europe, displacing, and eventually driving to extinction, the Neanderthals who had inhabited these lands for several hundred thousand years. But not before the Neanderthals contributed a small amount to their gene pool, shown recently by careful analysis of both modern human and Neanderthal genomic DNA sequences. My ancestors eventually crossed Eastern Europe and settled in what is now called Lithuania. The evidence for this narrative is contained in my DNA.

A partnership between National Geographic and IBM (The Genographic Project) aims to establish the migratory paths of modern humans through the collection and sequencing of DNA from many tens of thousands of volunteers. I am one of those volunteers and had a small section of my mitochondrial DNA sequenced. This snippet of DNA tracks with my mother's side of the family as the mitochondrial DNA is only contributed through females. My DNA sequence is

seen in **Figure 1**. The handful of nucleotides that vary from a reference sequence, shown in a lighter shade, indicate that I am in the “T haplogroup,” which places my ancestors on the migratory path described above. With time and more analysis, perhaps more details will be filled in.

YOUR MITOCHONDRIAL HVR I SEQUENCE
16126C, 16147T, 16183C, 16189C, 16294T, 16296T, 16297C,
16304C, 16519C

```

ATTCTAATTTAACTATTCTCTGTTCTTTTCATGGGGAAGCAGATTTGGGTACCA
CCCAAGTATTGACTCACCCATCAACAACCGCTATGATTTTCGTACATTACTGCC
AGCCACCATGAATATTGCACGGTACCATAAATACCTTGATCACCTGTAGTACATAA
AAACCAATCCACATCAAACCCCCCCCCCATGCTTACAAGCAAGTACAGCAAT
CAACGCTCAACTATCACACATCAACTGCAACTCCAAGCCACCCTCACCCAC
TAGGATACCAACAACTACCCATTCTTAACAGCACATAGTACATAAAGCCATT
ACCGTACATAGCACATTACAGTCAAATCCCTTCTCGTCCCCATGGATGACCCC
CCTCAGATAGGGGTCCTTGACCACCATCCTCGTGAAATCAATATCCGCAC
AAGAGTGCTACTCTCCTCGCTCGGGGCCATAACACTTGGGGGTAGCTAAAGT
GAAGTGTATCCGACATCTGGTTCCTACTTCAGGGCATAAAGCCTAAATAGCCC
ACACGTTCCCTTAAATAAGACATCAGGAT

```

Figure 1 The author's DNA. The sequence shown is from the author's mitochondrial DNA. The sequence was provided by The Genographic Project, www.nationalgeographic.com/genographic.

Along the way, in response to the environment, they have changed their diet and lifestyle. Here are a few examples along with the genetic changes associated with these adaptations. Many may have occurred during the last 40,000 years, or since the more recent start of agriculture. Included are the official gene names or symbols when known.

- **Skin color.** Humans near the equator have retained a darker skin color to block damaging ultraviolet light. However, people closer to the Earth's poles need and have paler skin allowing them to make enough light-induced vitamin D. Sequence variation in a number of genes, such as *SLC24A5*, appears to be responsible for this skin pigmentation.
- **Lactose tolerance.** It is estimated that as recently as 8000 years ago, goats and cattle were domesticated and their milk was consumed by humans, especially at times of poor crop yields. This practice was probably a contributing factor to preventing starvation in the historically frequent famines. Normally, the ability to digest lactose rapidly decreases after early childhood, resulting in considerable intestinal discomfort after consuming milk. But a mutation arose which resulted in the persistence of lactase expression into adulthood, allowing milk consumption without side effects. Lactose tolerance quickly spread through the European population and a sequence variation near the promoter of the lactase gene, *LCT*, appears to be responsible. Interestingly, a different set of promoter sequence variations arose independently in pastoral African populations, reaching 90% of the Tutsi population.
- **Digestion of starch.** Like the digestion of lactose, there have been selective pressures for the increased ability to survive on high-starch diets. Amylase is an enzyme found in saliva and provides the first steps in the digestion of starch. It has been found that populations that consume a lot of starch have high copy numbers of the amylase (*AMY1A*) gene while populations with low-starch diets have fewer amylase genes. More amylase gene copies results in higher amylase expression, especially in saliva, conferring an advantage for digesting food.
- **Malaria resistance and sickle cell anemia.** Malaria is a tropical disease caused by a *Plasmodium* blood infection. It attacks hundreds of millions of people each year and is fatal to tens of thousands. Sickle cell anemia is a disease where mutations in the hemoglobin B gene (*HBB*) lead to misshapen red blood cells. In addition to carrying less oxygen, the crescent-shaped cells cause problems of poor circulation such as pain and organ damage. However, you are more likely to survive malaria if you carry one copy of the sickle cell disease gene and, over multiple generations, mutations of the sickle cell trait have spread rapidly through the populations most at risk for malaria.
- **Life at high altitude.** There are several human populations that live and thrive at extremely high altitudes and have high red blood cell counts in response to the low oxygen levels. Yi and colleagues identified a transcription factor gene, *EPAS1*, as having a SNP present in most high-altitude Tibetans but mostly absent from Han Chinese living at low elevation. The Tibetan population split from the Han population less than 3000 years ago. Interestingly, *EPAS1* expression rises in response to low oxygen levels.

1.9 DNA SEQUENCE IN DATABASES

In earlier sections, the major drivers of database growth were described. With this growth and wealth of data comes the ability to address long-term questions such as finding molecular evidence of evolution, and examples of this were also described. Genomic and cDNA sequences are chiefly responsible for the flood of information into GenBank and the basics of DNA sequence assembly and cDNA



The black blood of Uro Indians

There is a legend that the Uro Indians of Peru had “black blood” which helped them survive at the cold and high altitudes. Although the legend may not be true it is interesting that the story comes from a time of limited biochemistry knowledge yet blood color, and therefore oxygen-binding hemoglobin, is connected to this legend.

synthesis will be described here and explained in more detail in [Chapter 5](#). The same principles of **genome sequence assembly** apply to both established and next-generation sequencing methods. Importantly, bioinformatics plays a key role in assembling the millions of sequences into contiguous pieces of genome. As we search databases and come across pieces of DNA sequence, it is important to appreciate the origins of those fragments, both from a scientific point of view and source of pride in human ingenuity.

Genomic DNA assembly

Most genomes are millions of nucleotides long, far surpassing the length of sequence generated by current sequencing technologies. So genomic sequencing efforts have all involved breaking chromosomal DNA into pieces and then working with the smaller fragments. Once the fragments are in a suitable format and size, the DNA sequence is determined and bioinformatics software assembles the fragments into long contiguous stretches with the goal of assembling the genome sequence from end to end.

Now, it is important to remember that when chromosomal DNA is isolated, you are not working with just one copy. DNA is obtained from something abundant, for example cells grown in culture, a whole organ, or even a whole organism or flask of organisms. Since you are not working with a single cell, you are isolating the DNA from many millions of cells and therefore have millions of copies of each gene.

It is also important to realize that you are sequencing random pieces of DNA. The approach you took to randomly fragment the chromosomal DNA generated many different beginnings and endings. That is, you are running thousands or millions of sequencing reactions at once and they correspond to regions all over the genome. Furthermore, your sequencing reactions are not generating the entire sequence of a gene; sequencing only generates short stretches. Finally, you are starting at random places in gene A, gene B, gene C, and so on. After you are done with the sequencing reactions, you have DNA sequence from the beginning and end of the genes, and everywhere in between. Because you had multiple copies of each gene in the original sequencing reactions, you have overlapping copies of sequence. But all of this is required to get any level of accuracy of sequence.

To understand the strength in the randomness described above, let's start with an analogy. Imagine taking a piece of paper on which two sentences are written and with scissors, cutting the page into those individual sentences. Consider these two adjacent sentences,

Here comes a fox. The fox jumps over the lazy dog.

and the pieces that you generated with scissors, deliberately put out of order:

The fox jumps over the lazy dog.
Here comes a fox.

If you had no knowledge of the original order of sentences and were asked to assemble these as they were before, you would be at a loss since there is no overlap between sentences. There is some hint of sentence order if you consider them individually because you understand English, but you can't be absolutely sure about the order when you try to reassemble the sentences. But if you had multiple copies of the sentences, and many pieces of sentences cut at random places, and pieces that spanned the two sentences, you could assemble, with confidence, the order of the sentences and place a consensus assembly (built from the agreement between words) underneath the fragments:

```

Here comes
  comes a fox. The f
        fox. The fox jumps over the
                        over the lazy dog.
                                the lazy dog.
Here comes a fox. The fox jumps over the lazy dog.

```

This analogy is close to the approach and solution to sequencing and assembling genomic DNA. The overlapping nature allows us to confidently determine the relationship among fragments. The unique words (*here, comes, jumps, over, lazy, dog*) are analogous to genes, scattered about. Genomic DNA has repetitive elements, much like the repeating words (*fox, the*), also distributed unevenly. These can be a problem unless you have adjacent unique sequences (words).

Now look at a small stretch of DNA sequence taken from a rat gene:

```
ACAAGGGACTAGAGAAACCAAACGAAAGGTGCAGAAGGGGAAACAGATGCAGAAAGCATCTGGAGACAA
```

Let's "sequence" multiple copies, starting from random locations, in overlapping pieces, and build a consensus, shown below it:

```

ACAAGGGACTAGAGAAACCAAAC
      AGAAACCAAACGAAAGGTGCAGAA
                AACGAAAGGTGCAGAAGGGGAAACAGATGCAGA
                        GAAGGGGAAACAGATGCAGAAAGCATCT
                                AGAAAGCATCTGGAGACAA
ACAAGGGACTAGAGAAACCAAACGAAAGGTGCAGAAGGGGAAACAGATGCAGAAAGCATCTGGAGACAA

```

Like the two-sentence analogy, the overlapping fragments allowed you to order the pieces. Note that there are multiple places of "AAA" which could confuse the assembly process but the adjacent sequence allowed the correct assembly.

Many early genome assemblies aimed for approximately sixfold coverage (overlapping regions six sequences deep) but with the next generation of sequencing machines, 20–100-fold coverage is now commonplace. Even so, errors in assembly can and do occur, often because of scattered repetitive regions ranging from hundreds to thousands of nucleotides long. These repeats are nearly identical in sequence and can be indistinguishable from each other.

For reasons that are not always clear, some regions of DNA cannot be isolated very easily, do not clone at high efficiency, and/or cannot be sequenced very accurately. This results in regions that are underrepresented in the multifold coverage or not represented at all. These "holes" can be mapped and addressed through alternative means, but nevertheless represent a hurdle in generating genomic sequence. You will often see these holes as a long string of Ns (NNNNNNNNNNNNNNNNNNNNNNNNNNNNNN) as placeholders where the scientists know the length of a fragment, based on the distance between flanking markers, but not its content.

Remember, DNA is double-stranded. When you sequence DNA randomly, you could be sequencing the complementary strand, which gives you twice the sequence to consider when trying to pull all your fragments together to build the consensus. In practice, sometimes the sequence of a complementary strand is more easily obtained, so having the other strand's sequence is a benefit and simply contributes to the redundancy you need for accuracy.

Finally, you cannot forget the huge bioinformatics contribution for the assembly of genomic DNA sequence. That is, if you have to assemble literally millions of randomly generated DNA sequences, each fragment ranging from 25 to 900 nucleotides long, you must use a computer program to accomplish this difficult

goal. Going back to the two-sentence analogy above, imagine trying to assemble the sentences of a 23-volume encyclopedia, with millions of words in each volume, and 20–100 copies of those 23 volumes. Assembling the original version of the human genome sequence required the assembly of over 27 million fragments (approximately fivefold coverage) and literally weeks of computing time on one of the largest known computers at that time. Complete understanding of the complex computer programs that accomplished this feat is beyond the scope of this book, but it is valuable to appreciate that bioinformatics was key to this historic event. With the above description as a background, be prepared that the genomic DNA sequence in our databases may:

- be very long pieces (contiguous stretches or contigs) but often, many small fragments;
- contain regions of unknown sequence;
- contain mistakes in sequence;
- be assembled incorrectly;
- contain either strand of the double helix in the database unless it is described in more detail (like a gene);
- be represented multiple times in the database.

cDNA in databases—where does it come from?

A huge contribution to the sequence in DNA databases is cDNA. As more thoroughly explained in [Chapter 5](#), messenger RNA (mRNA) is quite unstable and using enzymes to convert this polymer into stable DNA is the preferred approach for cloning and sequencing of these transcripts. cDNA analysis is critical to the understanding of gene expression and function so, as a result, this form of DNA is very prominent in the analysis in this book. The basic steps in cDNA synthesis are described below.

There are approximately 21,000 human protein-encoding genes. Around 8000 are ubiquitously expressed (that is, transcribed and translated) in all tissues and have functions in common with all cells: DNA replication, energy metabolism, regulation of transcription, translation, and so on. The expression of the other 13,000 genes is thought to be somewhat specific between the cell types, tissues, developmental states, or any circumstances that make a cell type or condition unique. For example, genes important for the function of blood should only be expressed in blood cells, and genes important for liver function should only be found in liver. A direct approach to identifying liver-specific genes would be to isolate all the proteins in a sample of liver and identify them by sequencing. This is technically challenging and is beyond the expertise of many laboratories.

Another approach is to isolate and study the mRNAs in liver. These encode the proteins found in liver and you could generate a list of genes that appear to be liver-specific. However, studying mRNA is technically challenging, as mRNA is very labile so only the most meticulous handling will prevent it from breaking down quickly. A technically easier approach to studying mRNA is to make a complementary DNA copy of the mRNA and then sequence this copy. Complementary DNA, or cDNA, is very stable, easily handled, and sequenced with little difficulty. You still have the challenge of isolating and properly handling mRNA from cells, but once you have cDNA, your success is almost guaranteed. A brief explanation of mRNA and cDNA synthesis will help you understand what you are looking at in DNA databases. This will also be covered in different detail in [Chapter 5](#).

Can measuring mRNA allow good estimates of protein levels? Be aware that not all transcripts are translated. We'll learn in [Chapter 10](#) that there are possibly thousands of genes that have RNA as their end product and are never translated. Furthermore, translation rates of transcripts vary considerably so there may not be a direct correlation between protein levels and the abundance of an mRNA.

mRNA is a long polymer with a “cap” at the beginning (5P end, pronounced “five prime”) and tail of AAAs (**poly(A) tail**) at the 3P end ([Figure 1.2A](#)). cDNA



"Tissue-specific" expression

The transcripts of most genes can be found in more than one tissue. Rather than think of genes as tissue-specific, it may be more accurate to think that gene expression is more “selective” for one tissue or cell type. Nevertheless, it is very common to refer to genes as being tissue-specific.

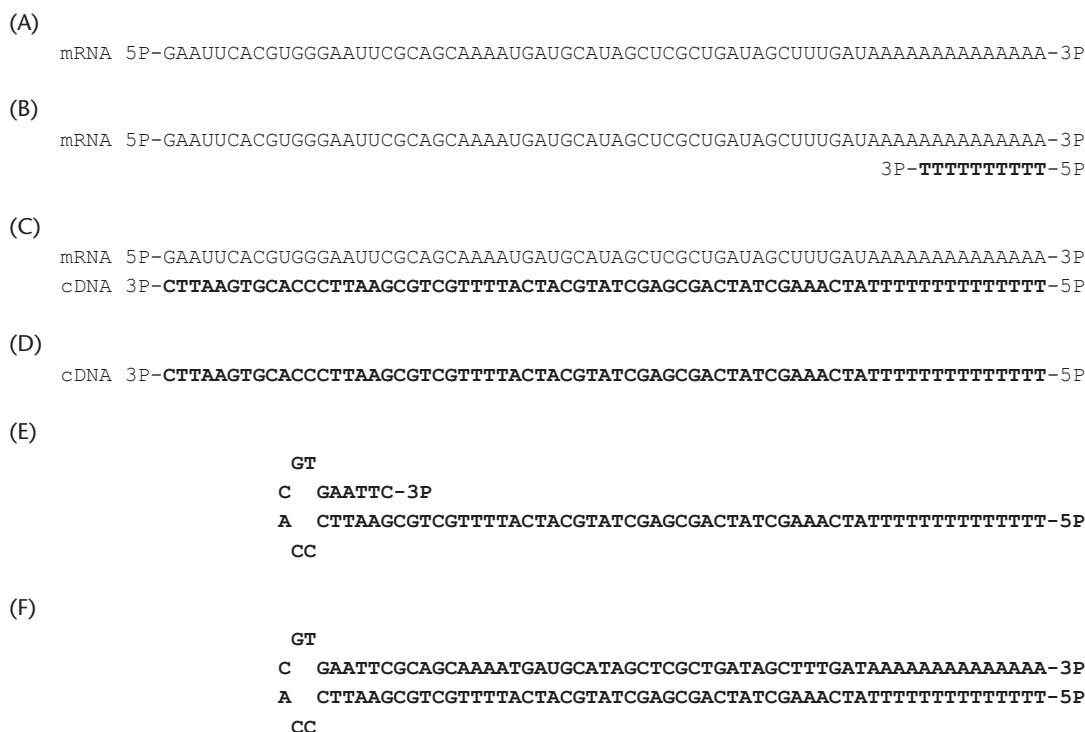
synthesis is begun by mixing the mRNA with the components necessary for the synthesis of cDNA. These components include the individual nucleotides (represented here as A, T, G, and C) and an enzyme called reverse transcriptase. As the name of this enzyme might suggest, it makes cDNA out of RNA. The enzyme requires a starting point: it needs a short stretch of DNA, called a primer, already sitting on the mRNA as a place to begin cDNA synthesis. Reverse transcriptase, like other DNA-synthesizing enzymes, can only begin at the 3P end of the primer. If you add a primer of poly(T) DNA to the reaction, it will base-pair with the poly(A) tail in the opposite orientation (Figure 1.2B).

Reverse transcription will then begin at the 3P end of the poly(T) and extend the DNA synthesis toward the 5P end of the mRNA (Figure 1.2C). The mRNA is then removed from the reaction, leaving only single-stranded cDNA behind (Figure 1.2D). The second strand of DNA now needs to be synthesized, but what will act as the 3P primer for this reaction? Multiple solutions have been devised but an early way to prime the reaction was to allow the cDNA to fold back on itself and self-prime (Figure 1.2E). The second strand synthesis was then completed (Figure 1.2F). Subsequent steps then clone the cDNA into vectors suitable for cloning and sequencing.

cDNA synthesis does have limitations. Synthesis of the first strand is not always efficient and the reverse transcriptase may fall off the mRNA before reaching the 5P end of the message. Since much of the early cDNA synthesis started with poly(T) priming, cDNAs in databases are often biased toward the 3P end. Reverse transcriptase is also error-prone so there may be mistakes in the sequence.

One type of cDNA mentioned above in Section 1.3 is called an Expressed Sequence Tag or EST. ESTs are often less than 500 nucleotides long even though they were derived from mRNAs thousands of nucleotides in length. What they lack in length is balanced out by quantity: ESTs are synthesized and sequenced in very high numbers (thousands). The result is often a very thorough sampling of the mRNAs expressed in that cell line, tissue, or organ. If nothing is done to normalize the mRNA population, the cDNA synthesized will be proportional to the abundance of the various mRNAs found in those cells. That is, abundant mRNAs will give rise to most of the ESTs, and rare mRNAs will give rise to rare ESTs or none at all. When you randomly pick ESTs to sequence, by chance you will sequence the

Figure 1.2 Synthesis of a cDNA from an mRNA. (A) mRNA. (B) A DNA primer is attached to the poly(A) tail. (C) Reverse transcriptase extends the cDNA to the 5P end of the mRNA. (D) cDNA after removal of the mRNA. (E) Formation of a “hairpin” at the 3P end of the cDNA acts as a primer for synthesis of the complementary strand (F).



abundant cDNAs repeatedly. So scientists will sequence ESTs by the thousands to find those rare EST sequences. ESTs will be discussed further in [Chapter 5](#).

1.10 SEQUENCE ANALYSIS AND DATA DISPLAY

The following example illustrates a very simple sequence analysis problem. As the analysis progresses, the display of data changes, demonstrating some of the variety of styles that you will see in this book.

Figure 1.3 shows the mRNA transcript from a human gene called Factor IX (pronounced “factor nine”). The Factor IX gene encodes a protein critical to the cascade of proteins that respond and work together to properly clot blood. Transcript sequences are conventionally shown in databases and in many publications as cDNA sequences, using “T” instead of “U.” In this form, the sequence is mostly uninformative, not providing any details except for general impressions about the nucleotide content and length (the sequence in [Figure 1.3](#) is 2802 nucleotides long).

But what if you knew two simple rules: protein-coding regions begin with “ATG” and end with “TAA,” “TGA,” or “TAG.” [Figure 1.4](#) shows bold and underlined the

Figure 1.3 The sequence of the mRNA for human Factor IX.

In GenBank and many other databases, sequence files are given unique identification numbers called “accession numbers.” This sequence is from GenBank accession number NM_000133.

```

ACCACTTTTCACAATCTGCTAGCAAAGGTTATGCAGCGCGTGAACATGATCATGGCAGAATCACCAGGCCT
CATCACCATCTGCCTTTTAGGATATCTACTCAGTGTCTGAATGTACAGTTTTTCTTGATCATGAAAACGCC
AACAAAATTTCTGAATCGGCCAAAGAGGTATAATTACGGTAAATTGGAAGAGTTTGTTCAGGGAACCTTG
AGAGAGAATGTATGGAAGAAAAGTGTAGTTTTGAAGAAGCACGAGAAGTTTTTGAAAACACTGAAAGAAC
AACTGAATTTTGAAGCAGTATGTTGATGGAGATCAGTGTGAGTCCAATCCATGTTTAAATGGCGGCAGT
TGCAAGGATGACATTAATTCCTATGAATGTTGGTGTCCCTTGGAATTTGAAGGAAAGAACTGTGAATTAG
ATGTAACATGTAACATTAAGAATGGCAGATGCGAGCAGTTTTGTAAAAATAGTGCTGATAACAAGGTGGT
TTGCTCCTGTACTGAGGGATATCGACTTGCAGAAAACCAGAAGTCTGTGAACCAGCAGTGCCATTTCCA
TGTGGAAGAGTTTCTGTTTTCACAACTTCTAAGCTACCCGTGCTGAGACTGTTTTCTCTGATGTGGACT
ATGTAATTTCTACTGAAGCTGAAACCATTTTGGATAACATCACTCAAAGCACCCAAATCATTTAATGACTT
CACTCGGGTTGTTGGTGGAGAAGATGCCAAACCAGGTCAATTCCTTGGCAGGTTGTTTTGAATGGTAAA
GTTGATGCATTCTGTGGAGGCTCTATCGTTAATGAAAAATGGATTGTAAGTGTCTGCCCACTGTGTTGAAA
CTGGTGTTAAAATTACAGTTGTGCGAGGTGAACATAATATTGAGGAGACAGAACATACAGAGCAAAAGCG
AAATGTGATTGCAATTATTCTCACCACAACATAAATGCAGCTATTAATAAGTACAAACCATGACATTGCC
CTTCTGGAAGTGGACGAACCTTAGTGCTAAACAGCTACGTTACACCTATTTCGATTGCTGACAAGGAAT
ACACGAACATCTTCTCAAATTTGGATCTGGCTATGTAAGTGGCTGGGGAAGAGTCTCCACAAAGGGAG
ATCAGCTTTAGTTCTTCAGTACCTTAGAGTTCACCTTGTTGACCGAGCCACATGTCTTCGATCTACAAAG
TTCACCATCTATAACAACATGTTCTGTGCTGGCTTCCATGAAGGAGGTAGAGATTATGTCAAGGAGATA
GTGGGGGACCCCATGTACTGAAGTGAAGGGACAGTTTCTTAAGTGAATATTAGCTGGGGTGAAGA
GTGTGCAATGAAAGGCAATATGGAATATATACCAAGGTATCCCGGTATGTCAACTGGATTAAAGAAAAA
ACAAAGCTCACTTAATGAAAGATGGATTTCCAAGGTTAATTCATTGGAATTGAAAATTAAACAGGCGCTCT
CACTAACTAATCACTTCCCATCTTTGTTAGATTGAAATATATACATTCATGATCATTGCTTTTCTCTC
TTTACAGGGGAGAATTTATATTTTACCTGAGCAAATTGATTAGAAAAATGAACCACTAGAGGAATATAA
TGTGTTAGGAAATTACAGTCATTTCTAAGGGCCAGCCCTTGACAAAAATTGTAAGTTAAATCTCCACT
CTGTCCATCAGATACTATGGTTCTCCACTATGGCAACTAACTCACTCAATTTCCCTCCTTAGCAGCATT
CCATCTTCCCGATCTTCTTTGCTTCTCCAACCAAAACATCAATGTTTATTAGTTCTGTATACAGTACAGG
ATCTTTGGTCTACTCTATCACAAGGCCAGTACCACACTCATGAAGAAAGAACACAGGAGTAGCTGAGAGG
CTAAACTCATCAAAAACACTACTCCTTTTCTCTACCTTATCTCAATCTTTTACCTTTTCCAAATCC
CAATCCCCAAATCAGTTTTTCTCTTTCTTACTCCCTCTCTCCCTTTTACCTCCATGGTCGTTAAAGGAG
AGATGGGGAGCATCTTCTGTTATCTGTACACAGTTATACATCATGATCAAAACCCAGACTTGCTTTC
CGTAGTGGAGACTTGCTTTTCAGAACATAGGGATGAAGTAAGGTGCCTGAAAAGTTTGGGGGAAAAGTTT
CTTTCAGAGAGTTAAGTTATTTTATATATATAATATATATATAAATATATAATATACAATATAAATATA
TAGTGTGTGTGTATGCGTGTGTGTAGACACACACGCATACACATATAATGGAAGCAATAAGCCATTCT
AAGAGCTTGTATGGTTATGGAGGTGTGACTAGGCATGATTTACGAGGCAAGATTGGCATATCATTGTA
ACTAAAAAGCTGACATTGACCCAGACATATTGTACTCTTTCTAAAAATAATAATAATGCTAACAGA
AAGAAGAGAACCGTTTCGTTTGCAATCTACAGCTAGTAGAGACTTTGAGGAAGAATTCAACAGTGTGTCTT
CAGCAGTGTTTCAGAGCCAAGCAAGAAGTTGAAGTTGCCTAGACCAGAGGACATAAGTATCATGTCTCCTT
TAAGTAGCATACCCGGAAGTGGAGAAGGGTGCAGCAGGCTCAAAGGCATAAGTCATTCCAATCAGCCAAC
TAAGTTGTCTTTTCTGGTTTCGTGTTCCACATGGAACATTTTGATTATAGTTAATCCTTCTATCTTGAA
TCTTCTAGAGAGTTGCTGACCAACTGACGTATGTTTCCCTTTGTGAATTAATAAAGTGGTGTCTGGTTC
AT

```

“ATG” and “TAA” triplets that bound the protein-coding region. You can now see that there are regions upstream and downstream of the coding region that do not code for protein. These are the 5P and 3P **untranslated regions (UTRs)**, respectively.

But these aren’t perfect rules. If you closely look at the sequence, there are many other instances of “ATG” and “TAA.” However, there are some additional constraints to consider. ATG can appear multiple times in a gene sequence, but often (but not always—this is biology!) the first ATG is used to start the coding region, as in this sequence. There are no other ATGs upstream of the one indicated in [Figure 1.4](#).

The protein-coding region is read three nucleotides at a time, starting at the ATG. These are called **codons**. The coding sequence can now be formatted to show all the codons ([Figure 1.5](#)). Although straightforward, this grouping step is completely dependent on the sequence being accurate. If the sequence was incorrect and a single nucleotide was missing or inserted, the grouping of three would be completely wrong from that point onward. A mistake involving two nucleotides would also be incorrect. However, if the insertion or deletion were a multiple of three, an event such as this would only have obvious consequences at the point of change, as all the other codons would be correct.

```

ACCACCTTTCACAATCTGCTAGCAAAGGTTATGCAGCGCGTGAACATGATCATGGCAGAATCACCAGGCCT
CATCACCATCTGCCTTTTAGGATATCTACTCAGTGCTGAATGTACAGTTTTCTTGATCATGAAAACGCC
AACAAAATTCGAATCGGCCAAAGAGGTATAATTCAGGTAAATTGGAAGAGTTTGTTCAGGGAACCTTG
AGAGAGAATGTATGGAAGAAAAGTGTAGTTTTGAAGAAGCACGAGAAGTTTTGAAAACACTGAAAGAAC
AACTGAATTTTGAAGCAGTATGTTGATGGAGATCAGTGTGAGTCCAATCCATGTTTAAATGGCGGCAGT
TGCAAGGATGACATTAATTCCTATGAATGTTGGTGCCCTTGGAATTGAAGGAAAGAACTGTGAATTAG
ATGTAACATGTAACATTAAGAATGGCAGATGCGAGCAGTTTTGTAAAAATAGTGTGATAACAAGGTGGT
TTGCTCCTGTACTGAGGGATATCGACTTGCAGAAAACCAGAAGTCTGTGAACCAGCAGTGCCATTTCCA
TGTGGAAGAGTTTCTGTTTCAAACTTCTAAGCTCAGCCGTGCTGAGACTGTTTTCTGTATGTGGACT
ATGTAATTTCTACTGAAGCTGAAACATTTTGGATAACATCACTCAAAGCACCCAATCATTTAATGACTT
CACTCGGGTTGTTGGTGGAGAAGATGCCAAACCAGGTCAATTCCTTGGCAGGTTGTTTTGAATGGTAAA
GTTGATGCATTCTGTGGAGGCTCTATCGTTAATGAAAATGGATTGTAACGTCTGCCACTGTGTTGAAA
CTGGTGTTAAAATTACAGTTGTGCGAGGTGAACATAATATTGAGGAGACAGAACATACAGAGCAAAAGCG
AAATGTGATTGCAATTATTCTCCACCAACTACAATGCAGCTATTAATAAGTACACCATGACATTGGC
CTTCTGGAACGTGGACGAACCTTAGTGCTAAACAGCTACGTTACACCTATTTGCATTGCTGACAAGGAAT
ACACGAACATCTTCCTCAAATTTGGATCTGGCTATGTAAGTGGCTGGGGAAGAGTCTCCACAAAGGGAG
ATCAGCTTTAGTTCTTCAGTACCTTAGAGTTCCACTGTGTGACCGAGCCACATGTCTCGATCTACAAAG
TTCACCATCTATAACAACATGTTCTGTGCTGGCTTCATGAAGGAGGTAGAGATTATGTCAAGGAGATA
GTGGGGGACCCCATGTTACTGAAGTGAAGGGACAGTTTCTTAAGTGAATATTAGCTGGGGTGAAGA
GTGTGCAATGAAAGGCAATATGGAATATATACCAAGGTATCCCGGTATGTCAACTGGATTAAGGAAAAA
ACAAAGCTCACTTAATGAAGATGGATTTCGAAGGTTAATTCATTGGAATTGAAAATTACAGGGCCTCT
CACTAATCAATCACTTTCCCATCTTTGTTAGATTTGAATATATACATTTATGATCATTTGCTTTTTCTC
TTTACAGGGGAGAATTTCATATTTTACCTGAGCAAATTGATTAGAAAATGGAACCACTAGAGGAATATAA
TGTGTTAGGAAATTACAGTCATTTCTAAGGGCCAGCCCTTGACAAAATTGTGAAGTTAAATCTCCACT
CTGTCCATCAGATACTATGTTTCTCCACTATGGCAACTAACTCACTCAATTTTCCCTCCTTAGCAGCATT
CCATCTTCCCGATCTTCTTTGCTTCTCCAACCAAAACATCAATGTTTATTAGTTCTGTATACAGTACAGG
ATCTTTGGTCTACTCTATCAAGGCCAGTACCACACTCATGAAGAAAGAACACAGGAGTAGCTGAGAGG
CTAAACTCATCAAAACACTACTCCTTTTCTCTACCCTATTCTCAATCTTTACCTTTTCCAAATCC
CAATCCCCAAATCAGTTTTTCTCTTTCTACTCCCTCTCTCCCTTTTACCCTCCATGGTCGTTAAAGGAG
AGATGGGGAGCATCTTTCTGTATATCTGTACACAGTTATACATGTCTATCAACCCAGACTTGTCTTC
CGTAGTGAGACTTGCTTTTCAACAATAGGGATGAAGTAAGGTGCCTGAAAAGTTTGGGGGAAAAGTTT
CTTTACAGAGAGTTAAGTTATTTTATATATATAATATATATAAAATATATAATATACAATATAAATATA
TAGTGTGTGTATGCGTGTGTGTAGACACACACGCATACACACATATAATGGAAGCAATAAGCCATTCT
AAGAGCTTGTATGGTTATGGAGGTCTGACTAGGCATGATTTACGAAAGGCAAGATTGGCATATCATTGTA
ACTAAAAAGCTGACATTGACCCAGACATATTGTACTCTTTCTAAAAATAATAATAATGCTAACAGA
AAGAAGAGAACCGTTTCGTTTGAATCTACAGCTAGTAGAGACTTTGAGGAAGAATTCAACAGTGTGTCTT
CAGCAGTGTTTCAGAGCAAGCAAGAAGTTGAAGTTGCTTAGACCAGAGGACATAAGTATCATGTCTCTCTT
TAAGTAGCATACCCCGAAGTGGAGAGAGGTGCAGCAGGCTCAAAGGCATAAGTCATTCCAATCAGCCAAC
TAAGTTGTCCTTTTCTGTTTCTGTGTTCACCATGGAACATTTTATTATAGTTAATCCTTCTATCTTGAA
TCTTCTAGAGAGTTGCTGACCAACTGACGTATGTTTCCCTTTGTGAATTAATAAACTGGTGTCTGGTTC
AT

```

Figure 1.4 Applying two rules for describing the human Factor IX mRNA sequence. Those two rules are (a) coding regions begin with “ATG” and (b) coding regions end with one of three terminator sequences, “TAA,” “TGA,” or “TAG.” Two of the many possible matches to these rules are in bold and underlined.

ACCACCTTTACAAATCTGCTAGCAAGGTT

ATG CAG CGC GTG AAC ATG ATC ATG GCA GAA TCA CCA GGC CTC ATC ACC ATC TGC CTT TTA GGA TAT CTA CTC AGT
 GCT GAA TGT ACA GTT TTT CTT GAT CAT GAA AAC GCC AAC AAA ATT CTG AAT CGG CCA AAG AGG TAT AAT TCA GGT
 AAA TTG GAA GAG TTT GTT CAA GGG AAC CTT GAG AGA GAA TGT ATG GAA GAA AAG TGT AGT TTT GAA GAA GCA CGA
 GAA GTT TTT GAA AAC ACT GAA AGA ACA ACT GAA TTT TGG AAG CAG TAT GTT GAT GGA GAT CAG TGT GAG TCC AAT
 CCA TGT TTA AAT GGC GGC AGT TGC AAG GAT GAC ATT AAT TCC TAT GAA TGT TGG TGT CCC TTT GGA TTT GAA GGA
 AAG AAC TGT GAA TTA GAT GTA ACA TGT AAC ATT AAG AAT GGC AGA TGC GAG CAG TTT TGT AAA AAT AGT GCT GAT
 AAC AAG GTG GTT TGC TCC TGT ACT GAG GGA TAT CGA CTT GCA GAA AAC CAG AAG TCC TGT GAA CCA GCA GTG CCA
 TTT CCA TGT GGA AGA GTT TCT GTT TCA CAA ACT TCT AAG CTC ACC CGT GCT GAG ACT GTT TTT CCT GAT GTG GAC
 TAT GTA AAT TCT ACT GAA GCT GAA ACC ATT TTG GAT AAC ATC ACT CAA AGC ACC CAA TCA TTT AAT GAC TTC ACT
 CGG GTT GTT GGT GGA GAA GAT GCC AAA CCA GGT CAA TTC CCT TGG GAG GTT GTT TTG AAT GGT AAA GTT GAT GCA
 TTC TGT GGA GGC TCT ATC GTT AAT GAA AAA TGG ATT GTA ACT GCT GCC CAC TGT GTT GAA ACT GGT GTT AAA ATT
 ACA GTT GTC GCA GGT GAA CAT AAT ATT GAG GAG ACA GAA CAT ACA GAG CAA AAG CGA AAT GTG ATT CGA ATT ATT
 CCT CAC CAC AAC TAC AAT GCA GCT ATT AAT AAG TAC AAC CAT GAC ATT GCC CTT CTG GAA CTG GAC GAA CCC TTA
 GTG CTA AAC AGC TAC GTT ACA CCT ATT TGC ATT GCT GAC AAG GAA TAC ACG AAC ATC TTC CTC AAA TTT GGA TCT
 GGC TAT GTA AGT GGC TGG GGA AGA GTC TTC CAC AAA GGG AGA TCA GCT TTA GTT CTT CAG TAC CTT AGA GTT CCA
 CTT GTT GAC CGA GCC ACA TGT CTT CGA TCT ACA AAG TTC ACC ATC TAT AAC AAC ATG TTC TGT GCT GGC TTC CAT
 GAA GGA GGT AGA GAT TCA TGT CAA GGA GAT AGT GGG GGA CCC CAT GTT ACT GAA GTG GAA GGG ACC AGT TTC TTA
 ACT GGA ATT ATT AGC TGG GGT GAA GAG TGT GCA ATG AAA GGC AAA TAT GGA ATA TAT ACC AAG GTA TCC CGG TAT
 GTC AAC TGG ATT AAG GAA AAA ACA AAG CTC ACT **TAA**
 TGAAGATGGATTTCGAAGGTTAATTCATTGGAATTGAAATTAACAGGGCCTCTCACTAATACTCACTTCCCCTCTTTTGTTAGATTTGAATATATACA
 TTCTATGATCATTTGCTTTTTCTCTTTACAGGGGAGAATTCATATTTTACCTGAGCAAATTGATTAGAAAATGGAACCACTAGAGGAATATAATGTGTAGG
 AAATTACAGTCATTTCTAAGGGCCAGCCCTTGACAAAATTGTGAAGTTAAATTTCTCCACTCTGTCCATCAGATACTATGGTTCTCCACTATGGCAACTAAC
 TCACTCAATTTTCCCTCCTTAGCAGCATTCATCTTCCGATCTTCTTTGCTTCTCCAACCAAAACATCAATGTTTATTAGTTCTGTATACAGTACAGGATC
 TTTGGTCTACTCTATCACAGGCCAGTACCACACTCATGAAGAAAGAACACAGGAGTAGCTGAGAGGCTAAAACTCATCAAAAACACTACTCCTTTTCCTCT
 ACCCTATTCTCAATCTTTTACCTTTTCCAAATCCCAATCCCAATCAGTTTTTCTCTTTCTTACTCCCTCTCTCCCTTTTACCTCCATGGTCGTTAAAG
 GAGAGATGGGGAGCATCATTCTGTATACCTTCTGTACACAGTTATACATGTCTATCAAAACCCAGACTTGCTTCCGTAGTGAGACTTGCTTTTCAGAACATA
 GGGATGAAGTAAGTGCCTGAAAAGTTTGGGGGAAAAGTTTCTTTCAGAGAGTTAAGTTATTTTATATATATAATATATATAAAATATATAATATACAAT
 AATAATATATAGTGTGTGTATGCGTGTGTGTAGACACACAGCATACACACATATAATGGAAGCAATAAGCCATTCTAAGAGCTTGTATGGTTATGGAGG
 TCTGACTAGGCATGATTTACGAAGGCAAGATTGGCATATCATTGTAACATAAAAAAGCTGACATTGACCCAGACATATTGTACTCTTTCTAAAAATAATAAT
 AATAATGCTAACAGAAAGAAGAGAACCGTTCGTTTGCAATCTACAGCTAGTAGAGACTTTGAGGAAGAATTCAACAGTGTGTCTTCAGCAGTGTTCAGAGCC
 AAGCAAGAAGTTGAAGTTGCCTAGACCAGAGGACATAAGTATCATGTCTCTTTAACTAGCATACCCCGAAGTGGAGAAGGGTGCAGCAGGCTCAAGGCAT
 AAGTCATTCCAATCAGCCAACTAAGTTGTCTTTTCTGGTTTCGTGTTCCACCATGGAACATTTTGATTATAGTTAATCCTTCTATCTTGAATCTTCTAGAGA
 GTTGCTGACCAACTGACGTATGTTTCCCTTTGTGAATTAATAAACTGGTGTCTGGTTCAT

Figure 1.5 Coding regions are read as triplets. The human Factor IX mRNA with the start and termination codons bold and underlined. The coding region has been divided into the three-base codons. The 5P and 3P untranslated regions (5P and 3P UTRs, respectively) appear before and after the coding region.

Scanning by eye, you can see that there are no other terminator codons—TAA, TAG, or TGA—within this coding region. But there are other ATG triplets, including two just downstream of the first one. All of these codons are translated into a polypeptide chain according to the genetic code.

If you wanted to, you could count the 462 codons to deduce the length of the protein as 461 amino acids (the terminator codon does not encode an amino acid). There are many software programs to do this for you, but you should also get into the habit of examining a sequence by eye as well. After all, software only finds things it is designed to look for but you may notice something that is not yet described.

There are programs that take DNA sequence as input and translate this into an amino acid sequence using the genetic code (to be covered later, in [Chapter 4](#)). [Figure 1.6](#) shows this translation below each codon using the one-letter code for the amino acids. This figure is a little complex since it includes both the nucleotides, some grouped as three-letter codons, and one-letter amino acids. It has its value, though; for example, you can see that there can be multiple codons for each amino acid. Methionine (M) is always ATG, and tryptophan (W) is always TGG, but valine (V) can be GTG, GTT, GTA, or GTC.

[Figure 1.7](#), which shows just the amino acid sequence, is a much simpler figure to study. If you knew the biochemical properties of the amino acids, you might recognize regions that are hydrophilic or hydrophobic. Based on the sequence you see, regions of amino acids that tend to fold into helical structures or sheets might be noticed. Or you might recognize certain groups of amino acids that often have attached sugar groups. These structural features may tell a story about

ACCACCTTTCACAATCTGCTAGCAAAGGTT

ATG CAG CGC GTG AAC ATG ATC ATG GCA GAA TCA CCA GGC CTC ATC ACC ATC TGC CTT TTA GGA TAT CTA CTC AGT
M Q R V N M I M A E S P G L I T I C L L G Y L L S

GCT GAA TGT ACA GTT TTT CTT GAT CAT GAA AAC GCC AAC AAA ATT CTG AAT CGG CCA AAG AGG TAT AAT TCA GGT
A E C T V F L D H E N A N K I L N R P K R Y N S G

AAA TTG GAA GAG TTT GTT CAA GGG AAC CTT GAG AGA GAA TGT ATG GAA GAA AAG TGT AGT TTT GAA GAA GCA CGA
K L E E F V Q G N L E R E C M E E K C S F E E A R

GAA GTT TTT GAA AAC ACT GAA AGA ACA ACT GAA TTT TGG AAG CAG TAT GTT GAT GGA GAT CAG TGT GAG TCC AAT
E V F E N T E R T T E F W K Q Y V D G D Q C E S N

CCA TGT TTA AAT GGC GGC AGT TGC AAG GAT GAC ATT AAT TCC TAT GAA TGT TGG TGT CCC TTT GGA TTT GAA GGA
P C L N G G S C K D D I N S Y E C W C P F G F E G

AAG AAC TGT GAA TTA GAT GTA ACA TGT AAC ATT AAG AAT GGC AGA TGC GAG CAG TTT TGT AAA AAT AGT GCT GAT
K N C E L D V T C N I K N G R C E Q F C K N S A D

AAC AAG GTG GTT TGC TCC TGT ACT GAG GGA TAT CGA CTT GCA GAA AAC CAG AAG TCC TGT GAA CCA GCA GTG CCA
N K V V C S C T E G Y R L A E N Q K S C E P A V P

TTT CCA TGT GGA AGA GTT TCT GTT TCA CAA ACT TCT AAG CTC ACC CGT GCT GAG ACT GTT TTT CCT GAT GTG GAC
F P C G R V S V S Q T S K L T R A E T V F P D V D

TAT GTA AAT TCT ACT GAA GCT GAA ACC ATT TTG GAT AAC ATC ACT CAA AGC ACC CAA TCA TTT AAT GAC TTC ACT
Y V N S T E A E T I L D N I T Q S T Q S F N D F T

CGG GTT GTT GGT GGA GAA GAT GCC AAA CCA GGT CAA TTC CCT TGG CAG GTT GTT TTG AAT GGT AAA GTT GAT GCA
R V V G G E D A K P G Q F P W Q V V L N G K V D A

TTC TGT GGA GGC TCT ATC GTT AAT GAA AAA TGG ATT GTA ACT GCT GCC CAC TGT GTT GAA ACT GGT GTT AAA ATT
F C G G S I V N E K W I V T A A H C V E T G V K I

ACA GTT GTC GCA GGT GAA CAT AAT ATT GAG GAG ACA GAA CAT ACA GAG CAA AAG CGA AAT GTG ATT CGA ATT ATT
T V V A G E H N I E E T E H T E Q K R N V I R I I

CCT CAC CAC AAC TAC AAT GCA GCT ATT AAT AAG TAC AAC CAT GAC ATT GCC CTT CTG GAA CTG GAC GAA CCC TTA
P H H N Y N A A I N K Y N H D I A L L E L D E P L

GTG CTA AAC AGC TAC GTT ACA CCT ATT TGC ATT GCT GAC AAG GAA TAC ACG AAC ATC TTC CTC AAA TTT GGA TCT
V L N S Y V T P I C I A D K E Y T N I F L K F G S

GGC TAT GTA AGT GGC TGG GGA AGA GTC TTC CAC AAA GGG AGA TCA GCT TTA GTT CTT CAG TAC CTT AGA GTT CCA
G Y V S G W G R V F H K G R S A L V L Q Y L R V P

CTT GTT GAC CGA GCC ACA TGT CTT CGA TCT ACA AAG TTC ACC ATC TAT AAC AAC ATG TTC TGT GCT GGC TTC CAT
L V D R A T C L R S T K F T I Y N N M F C A G F H

GAA GGA GGT AGA GAT TCA TGT CAA GGA GAT AGT GGG GGA CCC CAT GTT ACT GAA GTG GAA GGG ACC AGT TTC TTA
E G G R D S C Q G D S G G P H V T E V E G T S F L

ACT GGA ATT ATT AGC TGG GGT GAA GAG TGT GCA ATG AAA GGC AAA TAT GGA ATA TAT ACC AAG GTA TCC CGG TAT
T G I I S W G E E C A M K G K Y G I Y T K V S R Y

GTC AAC TGG ATT AAG GAA AAA ACA AAG CTC ACT **TAA**
V N W I K E K T K L T Stop

TGAAAGATGGATTTCGAAGGTTAATTCATTGGAATTGAAAATTAACAGGGCCTCTCACTAACTAATCACTTTCCCATCTTTTGTGTAGATTGAATATATACA
TTCTATGATCATTGCTTTTTCTCTTTACAGGGGAGAATTCATATTTTTACCTGAGCAAATGATTAGAAAATGGAACCACTAGAGGAATATAATGTGTAGG
AAATTACAGTCATTTCTAAGGGCCAGCCCTTGACAAAATGTGAAGTTAAATTTCTCCACTCTGTCCATCAGATACTATGGTTCTCCACTATGGCAACTAAC
TCACCAATTTTCCCTCCTTAGCAGCATTCATCTTCCCGATCTTCTTTGCTTCTCCAACCAAAACATCAATGTTTATTAGTTCTGTATACAGTACAGGATC
TTTGGTCTACTCTATCACAGGCCAGTACCACACTCATGAAGAAAGAACAGGAGTAGCTGAGAGGCTAAAACTCATCAAAAACACTACTCCTTTTCTCT
ACCCTATTCTCTCAATCTTTTACCTTTTCCAAATCCCAATCCCCAAATCAGTTTTTCTCTTTCTACTCCCTCTCTCCCTTTTACCCTCCATGGTCGTTAAAG
GAGAGATGGGGAGCATCTTCTGTTATACTTCTGTACACAGTTATACATGTCTATCAAACCCAGACTTGCTTCCGTAGTGAGAGACTTGCTTTTTCAGAACATA
GGGATGAAGTAAGGTGCCTGAAAAGTTTGGGGGAAAAGTTTCTTTCAGAGAGTTAAGTTATTTTATATATATAATATATATAAAAATATATAATATACAAT
ATAAATATATAGTGTGTGTATGCGTGTGTGTAGACACACACGCATACACATATAATGGAAGCAATAAGCCATTCTAAGAGCTTGATGGTTATGGAGG
TCTGACTAGGCATGATTTTCAAGGCAAGATTGGCATATCATTTGTAACATAAAAAAGCTGACATGACCCAGACATATTGTACTCTTTCTTCTAAAAATAAAT
AATAATGCTAACAGAAAGAGAGAACCGTTCGTTTGCAATCTACAGCTAGTAGAGACTTTGAGGAAGAATTCAACAGTGTGTCTTCAGCAGTGTTCAGAGCC
AAGCAAGAAGTTGAAGTTGCTAGACCAGAGGACATAAGTATCATGTCTCTTTAACTAGCATACCCCGAAGTGAGAGGGTGCAGCAGGCTCAAAGGCAT
AAGTCATTCGAATCAGCCAACTAAGTTGTCTTTTCTGTTTTCGTTTCACCATGGAACATTTTGATTATAGTTAATCTTCTATCTTGAATCTTCTAGAG
AGTTGCTGACCAACTGACGTATGTTTCCCTTTGTGAATTAATAAACTGGTGTCTGTTTCAT

Figure 1.6 Coding-region triplets are translated into amino acids. Each three-base codon can be translated into an amino acid using the genetic code. The one-letter representations for the amino acids appear in this figure (for example, “M” stands for methionine, “Q” stands for glutamine, and so on).

Figure 1.7 The protein sequence of human Factor IX. Here is the translation of the coding region appearing in Figure 1.6. This protein is 461 amino acids long.

MQRVNMIMAESPLITICLLGYLLSAECTVFLDHENANKILNRPKRYNSGKLEEFVQGNL
ERECEMEKCSFEEAREVFENTERTEFWKQYVDGDQCESNPCLNGGSKDDINSYECWCP
FGFEGKNCELDVTCNIKNGRCEQFCKNSADNKVVCSTEGYRLAENQKSCEPAVPFPCGR
VSVSQTSKLTRAETVFPDQYVNSTEATILDNITQSTQSFNDFTRVVGGEDAKPGQFPW
QVVLNGKVDFAFCGGSIVNEKWIVTAAHCVETGVKITVVAGEHNIEETEHEQKRNVIIR
PHHNYNAAINKYNHDIALLELDEPLVLSYVTPICIAADKEYTNIFLKFSSGYVSGWGRVF
HKGRSALVLQYLRVPLVDRATCLRSTKFTIYNNMFCAGFHGGGRDSCQGDSSGPHVTEVE
GTSFLTGIISWGEECAMKGKYGITYTKVSRVYNWIKETKLT

the function of this protein. Luckily, we have bioinformatics programs that will tell us about these biochemical properties, structures, and modifications. Protein analysis, along with structures and their visualization, will appear in [Chapter 8](#).

This simple example illustrates how analysis of a raw sequence ([Figure 1.3](#)) can be broken down into steps and additional information can be extracted through the application of distinct rules. Knowing the intermediate steps makes you more aware of the dependencies. The next two examples show how sequences can be compared.

The Factor IX transcript, analyzed above, is a product of the Factor IX gene that is over 38,000 nucleotides long. A single nucleotide mutation, changing a G to a T at coordinate 25,531, results in hemophilia B, a severe bleeding disorder ([Figure 1.8](#)).

Single nucleotide changes elsewhere in the gene are quite common but tolerated, either because they do not change the protein sequence or do not disrupt splicing or any form of regulation, or do not make changes to the protein that biochemically compromise the function of Factor IX. A comparison between the human and chimpanzee Factor IX proteins ([Figure 1.9](#)) illustrates that conservative variation in at least one amino acid position can be tolerated in primates.

The Factor IX protein sequence can be divided into domains with different biochemical properties and functions. These domains interact with each other and with other proteins to function properly. They attain a three-dimensional structure via folding although they are depicted in a linear form in [Figure 1.10](#). In this figure, the domains of interest are gray boxes while other amino acids, not described here, are in white boxes. Domain one, at the N-terminus of the protein, functions to direct the protein to the endoplasmic reticulum of liver cells, from where it is secreted into the blood. When the protein is secreted, this first domain is cleaved off by a protein called signal peptidase. Twelve glutamic acid residues in domain two (also called the “Gla” domain) are modified by the enzyme gamma-carboxylase to become gamma-carboxyglutamic acid residues. Domain three is an “epidermal growth factor (EGF)-like” domain that binds calcium. Skipping ahead, domain five is a peptidase domain that cleaves another protein in the clotting cascade, the X protein. This protease function only becomes active once the Factor IX protein is cleaved into two peptides, and this cut occurs in domain four. That is why domain four is called the “activation peptide.” This cleavage results in two polypeptides, the Factor IX light chain consisting of domains two and three, and the Factor IX heavy chain consisting of domain five. The heavy and light chains remain covalently linked to each other by a disulfide bond between two cysteines. To transform from precursor to functional protein, Factor IX interacts with at least four other molecules, as described above.

Finally, here are two more views of the human Factor IX gene showing the context within genomic DNA. In [Figure 1.11A](#), the entire 38,000-nucleotide gene is

Figure 1.8 Hemophilia B mutation. A simple pairwise alignment between a small portion of the normal sequence and a sequence from a hemophilia B patient is shown. A vertical bar links positions where the two sequences are identical. A single nucleotide change in the 38,059 base pair human Factor IX gene can cause hemophilia B. The sequence shown in this figure spans the location of the G-to-T mutation found at gene coordinate 25,531 in GenBank record K02402.

Normal	GATGCCAAACCAGGTCAATTCCCTTGGCAGGTACTTTATACTGATGGTGTGTCAAACTG
Mutation	GATGCCAAACCAGGTCAATTCCCTTGGCAGTTACTTTATACTGATGGTGTGTCAAACTG

Query	1	MQRVNMIMAESPLITICLLGYLLSAECTVFLDHENANKILNRPKRYNSGKLEEFVQGNL	60
Sbjct	1	MQRVNMIMAESPLITICLLGYLLSAECTVFLDHENANKILNRPKRYNSGKLEEFVQGNL	60
Query	61	ERECMEEEKCSFEEAREVFENTERTEFWKQYVDGDQCESNPCLNGGCKDDINSYECWCP	120
Sbjct	61	ERECMEEEKCSFEEAREVFENTERTEFWKQYVDGDQCESNPCLNGGCKDDINSYECWCP	120
Query	121	FGFEGKNCELDVTCNIKNGRCEQFCCKNSADNKVVCSTEGYRLAENQKSCEPAVPFPCGR	180
Sbjct	121	FGFEGKNCELDVTCNIKNGRCEQFCCKNSADNKVVCSTEGYRLAENQKSCEPAVPFPCGR	180
Query	181	VSVSQTSKLTRAETVFPDQVDYVNSTEAEITLDNITQSTQSFNDFTRVVGGEDAKPGQFPW	240
Sbjct	181	VSVSQTSKLTRAETVFPDQVDYVNSTEAEITLDNITQSTQSFNDFTRVVGGEDAKPGQFPW	240
Query	241	QVVLNGKVDAFCCGSIVNEKWIVTAAHCVETGVKITVVAGEHNIETEHTEQKRNVIIRII	300
Sbjct	241	QVVLNGKVDAFCCGSIVNEKWIVTAAHCVETGVKITVVAGEHNIETEHTEQKRNVIIRII	300
Query	301	PHHNYNAAINKYNHDIALLEDEPLVLNSYVTPICIAADKEYTNIFLFGSGYVSGWGRVF	360
Sbjct	301	PHHNYNAAINKYNHDIALLEDEPLVLNSYVTPICIAADKEYTNIFLFGSGYVSGWGRVF	360
Query	361	HKGRSALVLQYLRLVPLVDRATCLRSTKFTIYNNMFCAGFHEGGRDSCQGDGSGGPHVTEVE	420
Sbjct	361	HKGRSALVLQYLRLVPLVDRATCLRSTKFTIYNNMFCAGFHEGGRDSCQGDGSGGPHVTEVE	420
Query	421	GTSFLTGIISWGEECAMKGKGYIYTKVSRVNVNIKEKTKLT	461
Sbjct	421	GTSFLTGIISWGEECAMKGKGYIYTKVSRVNVNIKEKTKLT	461

Figure 1.9 Alignment of human (Query) and chimpanzee (Sbjct) Factor IX proteins. Instead of a vertical bar to signify identity, as seen in Figure 1.8, this alignment places identical amino acids between the two sequences. There is only one amino acid difference over the 461 amino acid length (see if you can spot it). This change is biochemically conservative and so the space between the E (glutamic acid) and the D (aspartic acid) is indicated by "+" rather than a blank space indicating "no identity." The chimpanzee protein sequence is from the NCBI RefSeq sequence file NP_001129063.1 and the human sequence is from file NP_000124.1.

shown as the black arrow (labeled *F9*) near the center of the figure. Factor IX has several genetic neighbors including steroid-5- α -reductase, alpha polypeptide 1 pseudogene 1 (*SRD5A1P1*). A **pseudogene** is a sequence derived from a known functional gene but which is somehow defective structurally. There are over 15,000 pseudogenes in the human genome. These may represent opportunities for new functional genes to arise, or be defective genes that are destined to accumulate mutations until they are no longer recognizable as once being similar to any known gene. Another genomic neighbor is *MCF2* which encodes a protein capable of transforming normal tissue culture cells into a cancerous state. Note that the *MCF2* gene is over 126,000 nucleotides long and has the opposite orientation of the Factor IX gene, as indicated by the direction of the arrows. Genes can be found on both DNA strands and their sizes vary tremendously.

Figure 1.11B shows the location of the Factor IX gene on the X chromosome. Chromosomes are one long piece of genomic DNA with features that allow their identification and orientation. The X chromosome is 155 million nucleotides long, with a constriction near the center called the centromere. Throughout the chromosome are regions that, upon staining, show dark and light bands. Factor IX is one of 1500 genes and pseudogenes on this chromosome.

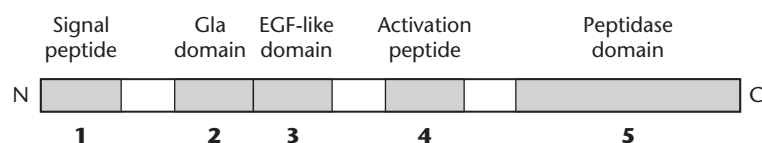


Figure 1.10 Factor IX protein domains. The Factor IX protein contains five major domains (gray boxes), each with specific functions in the molecule. They are depicted as boxes on a line representing the length of the protein, from N-terminus (N) to the C-terminus (C).

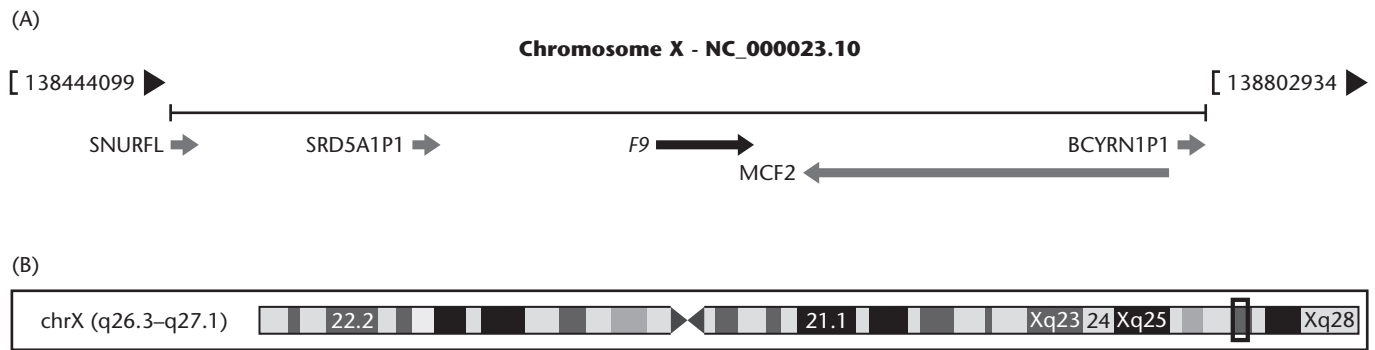


Figure 1.11 The position of the Factor IX gene on chromosome X.

(A) Detail of the part of chromosome X that contains the Factor IX gene (the locus). The Factor IX gene is shown as a black arrow pointing to the right, 5P to 3P. It is labeled using the official gene symbol, "F9." Note that the genes shown differ in size and orientation. This screenshot is taken from the NCBI Gene database. (B) The human X chromosome with the banding pattern similar to what is seen with a light microscope. On the far right, the boxed area indicates the approximate region shown in (A). This screenshot is taken from the University of California at Santa Cruz Genome Browser.

1.11 SUMMARY

This chapter serves as an introduction to sequence analysis, which is, perhaps, the first specialty within bioinformatics and has become the cornerstone of interpreting the deluge of data that we are experiencing today. As this chapter described, sequence data come from animals, plants, and microbes and from research and medicine. In the coming chapters you will acquire the practical skills of using sequence analysis tools, giving you access to this wealth of information.

FURTHER READING

- Anonymous (2011) Microbiology by numbers. *Nat. Rev. Microbiol.* 9, 628. A collection of interesting statistics about microbes.
- Benson DA, Karsch-Mizrachi I, Lipman DJ et al. (2011) GenBank. *Nucleic Acids Res.* 39 (Database issue), D32–37.
- Cordain L, Eaton SB, Sebastian A et al. (2005) Origins and evolution of the Western diet: health implications for the 21st century. *Am. J. Clin. Nutr.* 81, 341–354.
- Feuk L, Carson AR & Scherer SW (2006) Structural variation in the human genome. *Nat. Rev. Genet.* 7, 85–97.
- Gan X, Stegle O, Behr J et al. (2011) Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* 477, 419–423.
- Green RE, Krause J, Briggs AW et al. (2010) A draft sequence of the Neandertal genome. *Science* 328, 710–722.
- Itan Y, Powell A, Beaumont MA et al. (2009) The origins of lactase persistence in Europe. *PLoS Comput. Biol.* 5, e1000491.
- Johnson M, Gallagher K, Porter G et al. A baffling illness. Milwaukee Journal Sentinel, December 19, 2010. Sifting through the DNA haystack. Milwaukee Journal Sentinel, December 22, 2010. Embracing a risk. Milwaukee Journal Sentinel, December 26, 2010. An award-winning series of newspaper articles about the use of genomic DNA sequencing to help diagnose a patient.
- Li J, Yang T, Wang L et al. (2009) Whole genome distribution and ethnic differentiation of copy number variation in Caucasian and Asian populations. *PLoS One* 4, e7958 (DOI: 10.1371/journal.pone.0007958).
- Lupski JR, Reid JG, Gonzaga-Jauregui C et al. (2010) Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *New Engl. J. Med.* 362, 1181–1191.
- McDermott R, Tingley D, Cowden J et al. (2009) Monoamine oxidase A gene (MAOA) predicts behavioral aggression following provocation. *Proc. Natl Acad. Sci. USA* 106, 2118–2123.
- Pelak K, Shianna KV, Ge D et al. (2010) The characterization of twenty sequenced human genomes. *PLoS Genet.* 6, e1001111 (DOI:10.1371/journal.pgen.1001111). Genomic sequencing accurately identified those with hemophilia and found a number of genes that were "knocked out."

- Perry GH, Dominy NJ, Claw KG et al. (2007) Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* 39, 1256–1260.
- Pontius JU, Wagner L & Schuler GD (2003) UniGene: a unified view of the transcriptome. In *The NCBI Handbook*. Bethesda, MD: National Center for Biotechnology Information.
- Reich D, Green RE, Kircher M et al. (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* 468, 1053–1060.
- Roach JC, Glusman G, Smit AF et al. (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 328, 636–639. The genomic sequencing of two healthy parents and their two children affected by genetic disorders.
- Schuenemann VJ, Bos K, DeWitte S et al. (2011) Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of *Yersinia pestis* from victims of the Black Death. *Proc. Natl Acad. Sci. USA* 108, E746–752.
- Sturm RA (2009) Molecular genetics of human pigmentation diversity. *Hum. Mol. Genet.* 18, R9–17. Includes a beautiful picture of 19 forearms showing the range in skin pigmentation.
- Tishkoff SA, Reed FA, Ranciaro A et al. (2007) Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* 39, 31–40.
- Tripp S & Grueber M (2011). Economic impact of the human genome project. Battelle Memorial Institute.
- Yi X, Liang Y, Huerta-Sanchez E et al. (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 329, 75–78.
- Zhu J, He F, Hu S & Yu J (2008) On the nature of human housekeeping genes. *Trends Genet.* 24, 481–484. Housekeeping genes are expressed in all cell types, taking care of basic functions such as transcription, translation, and cell division.

Internet resources

The NCBI has a collection of electronic textbooks available through their Website which is a good place to search for explanations of the biology and technology mentioned in this chapter: www.ncbi.nlm.nih.gov/books.

GenBank release notes: <ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt>



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

ACAAGGGACTAGAGAAACCAAAA

AGAAACCAAAACGAAAGGTGCAGAA

AACGAAAGGTGCAGAAGGGGAAACAGATGCAGA

GAAGGGGAAACAGATGCAGAAAGCATC

AGAAAGCATC

ACAAGGGACTAGAGAAACCAAAACGAAAGGTGCAGAAGGGGAAACAGATGCAGAAAGCATC

ACAAGGGACTAGAGAAACCAAAA

AGAAACCAAAACGAAAGGTGCAGAA

AACGAAAGGTGCAGAAGGGG

GAAGGGG

CHAPTER 2

Introduction to Internet Resources

Key concepts

- Searching the medical and scientific literature: PubMed, iHOP, OMIM
- Searching the patent literature
- Gene classifications: Ontology
- Sequence collections such as Gene and UniGene

2.1 INTRODUCTION

How do you first learn about an interesting topic in biology? Is it in a lecture? A book? From a newspaper article or a conversation? Or do you observe it yourself in the laboratory or during a walk in the woods? However you first learn about a topic, you may wish to learn more and the early steps in this journey should include a search of the medical and scientific literature. Here you find a virtual mountain of past observations, experiments, discussions, comparisons, speculations, and explanations about our world. The benefits of consulting past observations are too numerous to list but include saving you time by directing your next steps to understanding. Do you make more observations, or have they already taken place? Do you care to repeat them, or do you build on past accomplishments? You have entered the field of science because you are naturally curious, and want explanations, so why not learn from thousands who have walked before you?

This chapter introduces the tools and resources for obtaining information from published works and Internet databases. Here, and throughout the book, we will be exploring some of the best Websites in the world for genetic information. Our launch point is the Website for the National Center for Biotechnology Information (NCBI). This Website will be used extensively throughout this book and this introduction should give you the tools to work beyond the topics covered. The foundations of many observations are the genes that act alone or in concert to give rise to the phenotype, disease, behavior, or organism. In this chapter you will not be analyzing any sequences, but you will learn ways to find and better understand them. You will see a mix of locations and approaches with the ultimate goal of showing you a path to your objective.

2.2 THE NCBI WEBSITE AND ENTREZ

The NCBI Website is one of the major hubs of bioinformatics resources and innovation in the world, and provides a wealth of information. The NCBI home page

(**Figure 2.1**) welcomes you with quick access to commonly used tools or databases (“Popular Resources”), and links to collections of resources (left sidebar). The “Site Map” and “All Resources,” also on the left sidebar, give you access to everything in one place and it is interesting to scroll through all that is offered here.

It is important to remember that this Website is not just a collection of hyperlinks. The scientific community is trying to cope with truly massive amounts of biological information and the scientists and computer experts at the NCBI have created and maintain many powerful solutions to the problems of data storage, retrieval, and display. Many other wonderful Websites will be introduced in this book, but the NCBI is certainly a major player in the world of bioinformatics.

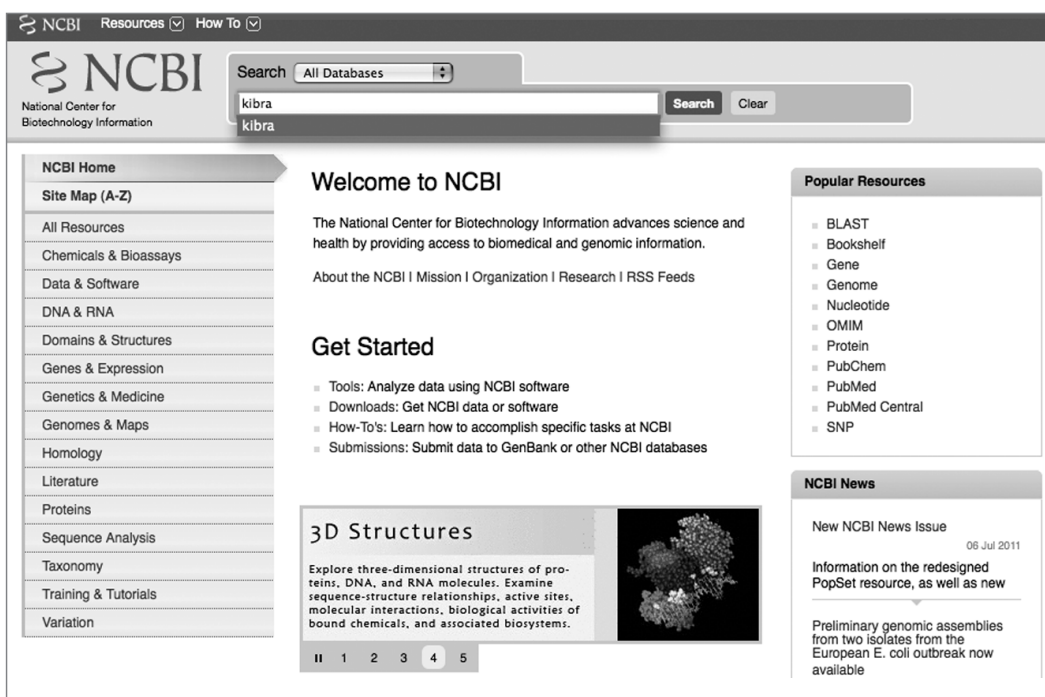
A primary interface between you and this wealth of information is the Entrez (pronounced “on-tray”; it is French for “enter”) retrieval system, the NCBI’s own search engine. On most pages on the Website, there is a very simple text field along with a drop-down menu which launches your searches of any or all of the databases available through Entrez. This text field appears at the top of **Figure 2.1**. In this example, *kibra*, the name of a gene, has been entered. Hitting the return key or clicking the “Search” button will launch a very fast search of 38 NCBI databases.


When the page refreshes, the majority of the page is a large section listing numerous databases, each with a brief description. The number of hits to each database is displayed to the left (**Figure 2.2**). Some databases have zero hits but this is understandable. Unless there was an organism named “*kibra*” you would not expect to find any hits in the Taxonomy database, for example. But on display are a number of choices for information, sequences, and other forms of data.

Clicking on the question mark next to each brief database description provides a longer explanation. To explore the hits in an individual database, just click on the number. Note that on the Entrez results page seen in **Figure 2.2**, the drop-down menu of individual databases is no longer available. But if you hit the back button or navigate to any of the individual results (for example, Nucleotide) you will find the familiar drop-down menu. Many of the individual databases will be described in detail later in this book. In this section of the chapter, we will focus on “PubMed: biomedical literature citations and abstracts.”

Figure 2.1 Home page for the NCBI (www.ncbi.nlm.nih.gov).

At the top of the page is a drop-down menu where you can choose a specific database to search. In this example, all the NCBI databases will be searched with the term “*kibra*.”






































NCBI  **Entrez, The Life Sciences Search Engine**

ARCH | SITE MAP | PubMed | All Databases | Human Genome | GenBank | Map Viewer

Search across databases [Help](#)

- Result counts displayed in gray indicate one or more terms not found

52  PubMed: biomedical literature citations and abstracts	2  Books: online books
116  PubMed Central: free, full text journal articles	2  OMIM: online Mendelian Inheritance in Man
none  Site Search: NCBI web and FTP sites	

272  Nucleotide: Core subset of nucleotide sequence records	none  dbGaP: genotype and phenotype
3  EST: Expressed Sequence Tag records	13  UniGene: gene-oriented clusters of transcript sequences
none  GSS: Genome Survey Sequence records	2  CDD: conserved protein domain database
246  Protein: sequence database	1574  Clone: integrated data for clone resources
19  Genome: whole genome sequences	none  UniSTS: markers and mapping data
1  Structure: three-dimensional macromolecular structures	none  PopSet: population study data sets
none  Taxonomy: organisms in GenBank	5395  GEO Profiles: expression and molecular abundance profiles
5374  SNP: short genetic variations	none  GEO DataSets: experimental sets of GEO data
193  dbVar: Genomic structural variation	none  Epigenomics: Epigenetic maps and data sets
124  Gene: gene-centered information	none  PubChem BioAssay: bioactivity screens of chemical substances
none  SRA: Sequence Read Archive	none  PubChem Compound: unique small molecule chemical structures
109  BioSystems: Pathways and systems of interacting molecules	3  PubChem Substance: deposited chemical substance records
1  HomoloGene: eukaryotic homology groups	none  Protein Clusters: a collection of related protein sequences
230  Probe: sequence-specific reagents	none  OMIA: online Mendelian Inheritance in Animals
none  BioProject: aggregated biological research project data	none  BioSample: biological material descriptions

2.3 PubMed

PubMed is a very large literature database, covering all major journals of biology and medicine. When you use Entrez to search PubMed, a number of searches are performed. First, your queries are used against an index of Medical Subject Headings (MeSH), which is a controlled vocabulary that describes the contents of a published paper. Queries are also used against indices from several fields including author. For example, searching with “white” will find papers on white blood cells or the *Drosophila* mutation called white, but also authors named White, too. You can specify which specific fields are searched, such as author and title/abstract, to eliminate many unwanted hits.

Clicking on the PubMed hits in Entrez brings you to a page providing more details: title, authors, and journal (**Figure 2.3**). The search term “kibra” found 52 hits in this example, sorted by date. In the upper-right corner of this figure, notice that related searches are suggested, including “kibra memory,” “kibra lats,” “hippo kibra,” and “kibra expression.” Clicking on these links will automatically return hits found with these terms, subsets of the original 52 since these hits must have both search terms present. Filters for the listed results, such as date, are found on the left side of the page.

The default display is 20 references to a page, which can save page-loading time by not providing long lists of hits should you have them. This display is also customizable. Clicking on the “Display Settings” hyperlink (upper left) opens a box (**Figure 2.4**). You can vary the information displayed, the citations per page, and the sorting order. Similar display settings are available on all of the Entrez results pages (for example, Nucleotide searches).

Back in the PubMed results (**Figure 2.3**), clicking on the hypertext title of the citation takes you to a page displaying the title, authors, and full abstract, if available (**Figure 2.5**). The information on this page can be exported using the “Send to:”

Figure 2.2 NCBI Entrez results page.

The search for the term “kibra” performed in **Figure 2.1** returned hits in many databases, including PubMed, Gene, OMIM, and UniGene. These specific databases will be explored in this chapter while others such as the Nucleotide, EST, Protein, and Structure databases will be covered in later chapters.



Search all of them or only what you need?

There are benefits for searching all the databases even if you are just looking for a nucleotide sequence. Scanning the results page, you may find it informative to see that there are only a few PubMed references, or that there are many. The same applies to all the other databases. So glance at the other hit numbers and let those bits of information provide you a little more background on your topic of interest.

The screenshot shows the PubMed search results for the term "kibra". The top navigation bar includes "NCBI", "Resources", "How To", "My NCBI", and "Sign In". The search bar contains "kibra" and the "Search" button is visible. Below the search bar, there are links for "RSS", "Save search", and "Advanced".

On the left side, there are filters for "Text availability" (Abstract available, Free full text available, Full text available), "Publication dates" (5 years, 10 years, Custom range...), "Species" (Humans, Other Animals), "Article types" (Review, more...), and "Languages" (English, more...). There is also a "Clear all" button and a "Choose additional filters" link.

The main content area shows "Display Settings: Summary, 20 per page, Sorted by Recently Added". It indicates "Did you mean: *kibra* (1 items)". Below this, it says "See 47 articles about kibra gene function" and "See also: kibra kibra ortholog in the Gene database". It lists "kibra in *Drosophila melanogaster* | *Homo sapiens* | *Mus musculus* | All 4 Gene records".

The results section shows "Results: 1 to 20 of 52". The first two results are visible:

- Positive Feedback and Mutual Antagonism Combine to Polarize Crumbs in the *Drosophila* Follicle Cell Epithelium.**
Fletcher GC, Lucas EP, Brain R, Tournier A, Thompson BJ.
Curr Biol. 2012 May 31. [Epub ahead of print]
PMID: 22658591 [PubMed - as supplied by publisher]
[Related citations](#)
- KIBRA exhibits MST-independent functional regulation of the Hippo signaling pathway in mammals.**
Moleirinho S, Chang N, Sims AH, Tilston-Lünel AM, Angus L, Steele A, Boswell V, Barnett SC, Ormandy C, Faratian D, Gunn-Moore FJ, Reynolds PA.
Oncogene. 2012 May 21. doi: 10.1038/onc.2012.196. [Epub ahead of print]
PMID: 22614006 [PubMed - as supplied by publisher]
[Related citations](#)

On the right side, there are sections for "Filters: Manage Filters", "Related searches" (kibra memory, kibra lats, hippo kibra, kibra expression, kibra merlin), "Titles with your search terms" (Kibra functions as a tumor suppressor protein that regulates Hippo signaling in [Dev Cell. 2010], Kibra is a regulator of the Salvador/Warts/Hippo signaling network. [Dev Cell. 2010], Common Kibra alleles are associated with human memory performance. [Science. 2006]), and "15 free full-text articles in PubMed Central" (Drosophila PI4KIIIalpha is required in follicle cells for oocyte polarization and F- [Development. 2011]).

Figure 2.3 PubMed result for the search term “kibra.” By clicking on the PubMed hits in the Entrez results page, seen in Figure 2.2, you are brought to a page listing those hits individually. Here, two hits are visible, showing title, authors, and journal citation. Clicking on the title will bring you to the full citation along with an abstract, if available.

The screenshot shows the "Display Settings" dialog box for PubMed. It has three main sections: "Format", "Items per page", and "Sort by".

- Format:** Radio buttons for Summary (selected), Summary (text), Abstract, Abstract (text), MEDLINE, XML, and PMID List.
- Items per page:** Radio buttons for 5, 10, 20 (selected), 50, 100, and 200.
- Sort by:** Radio buttons for Recently Added (selected), Pub Date, First Author, Last Author, Journal, and Title.

An "Apply" button is located at the bottom right of the dialog box.

Figure 2.4 Display settings for PubMed. Each Entrez results subsection has a “Display Settings” menu, tailored for the data to be displayed. Shown here are the PubMed Display Settings where you can view the list of hits in multiple ways (Format), vary the number of citations listed per page (Items per page), and sort the results.

tool on the right, which opens a dialog box (Choose Destination). In the figure, this box partially obscures the link to the full article on the far right. This link is often the logo of a journal, in this case *Science* magazine.

The authors' names are hypertext (Figure 2.5) so if you want to search for all articles published by that author, just click on their name and the page refreshes, listing their articles. In addition, the author's name appears in the search window (Figure 2.6A). Along with the author's name (Papassotiropoulos A), PubMed inserts a field label ([Author]) to direct the search to author names, ignoring other fields. You may do this manually, but PubMed also provides a form to construct these and other types of searches, so there is no need to memorize the field labels. This form is available through the “Advanced search” link visible in this and earlier figures near the top of the page.

In Advanced Search (Figure 2.6B) a drop-down menu (under “Builder”) offers many choices, only some of which are visible in this screenshot. You pick the field from the menu (Title/Abstract is this example), enter text in the window next to

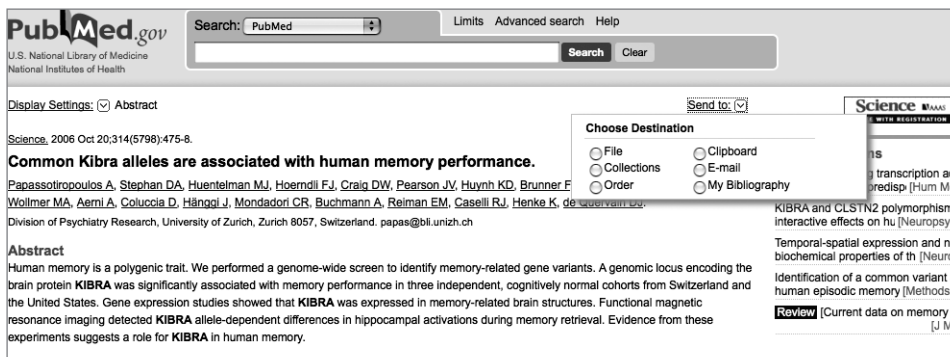


Figure 2.5 A single PubMed result. Clicking on any of the titles of the PubMed hits will take you to the full citation and an abstract, if available. The author names are hypertext; clicking on them will find other articles by that author. A “Send to:” menu, shown already open in this figure, allows you to export the citation in a variety of ways. There will also be a link to the full journal article (in this case, a link to *Science* magazine) if it is available online. Not all articles are free to view.

(A)

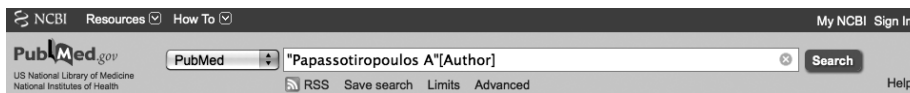
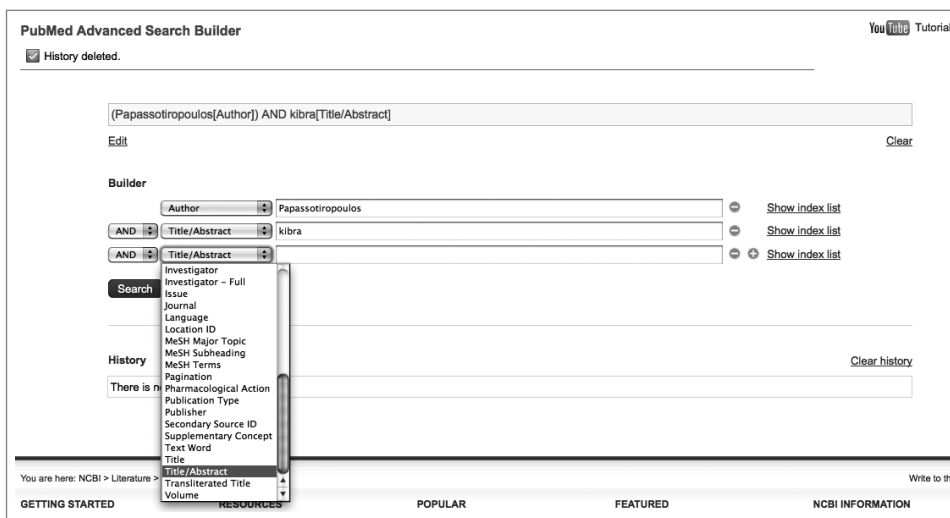


Figure 2.6 More specific searches in PubMed.

(A) Clicking on the hypertext author name of a citation will create a more specific PubMed search. There is now a field description called “Author” in square brackets and this will direct PubMed to only display articles by this author. (B) Clicking on the “Advanced search” hypertext in the PubMed title bar takes you to a query builder where many specific fields of the database can be selected from a drop-down menu, and the Boolean terms “AND,” “OR,” and “NOT” can be selected from a separate menu (right). The “Limits” hyperlink, shown in this and earlier figures on PubMed, allows the placement of other constraints upon the search such as dates or language.

(B)



your choice, and the form automatically constructs your **query** in the text box at the top of the page. This form also offers a choice of “AND,” “OR,” and “NOT” to construct the logic for your approach (the default is “AND”). These choices are available next to the Builder drop-down menus. For example, let’s say that there were two authors with the same name, “Papassotiropoulos A,” and one has published articles about genes while the other has published papers about heart surgery. You may wish to construct “(Papassotiropoulos A[Author]) NOT surgery[Title/Abstract]” to generate a list of articles that are more specific.

2.4 GENE NAME EVOLUTION

Gene names can often be considered a moving target. If you search PubMed for “kibra” you find 41 hits. But using the official gene symbol, *WWC1*, given to this gene some time after it was first described, you only get 28 hits. If you do an “OR” search, kibra OR *WWC1*, you get 41 hits, suggesting that kibra is the term to use.

There are thousands of laboratories around the world and many laboratories make the same discoveries and publish their names for the same gene. Even the same laboratory could have multiple names for the same gene as their research