# BIOINFORMATICS AND FUNCTIONAL GENOMICS

*Second Edition*

**Jonathan Pevsner**

Department of Neurology, Kennedy Krieger Institute

*and*

Department of Neuroscience and Division of Health Sciences Informatics, The Johns Hopkins School of Medicine, Baltimore, Maryland


A JOHN WILEY & SONS, INC., PUBLICATION

# Bioinformatics and Functional Genomics

# BIOINFORMATICS AND FUNCTIONAL GENOMICS

*Second Edition*

**Jonathan Pevsner**

Department of Neurology, Kennedy Krieger Institute

*and*

Department of Neuroscience and Division of Health Sciences Informatics, The Johns Hopkins School of Medicine, Baltimore, Maryland


A JOHN WILEY & SONS, INC., PUBLICATION

Wiley-Blackwell is an imprint of John Wiley & Sons, formed by the merger of Wiley's global Scientific, Technical, and Medical business with Blackwell Publishing.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada



Cover illustration includes detail from Leonardo da Vinci (1452–1519), dated c.1506–1507, courtesy of the Schlossmuseum (Weimar).

ISBN: 978-0-470-08585-1

***Library of Congress Cataloging-in-Publication Data is available.***

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

For Barbara, Ava and Lillian with all my love.

# Contents in Brief

# Contents

# Preface to the Second Edition

The Neurobehavioral Unit of the Kennedy Krieger Institute has 16 hospital beds. Most of the patients are children who have been diagnosed with autism, and most engage in self-injurious behavior. They engage in self-biting, self-hitting, head-banging, and other destructive behaviors. In most cases, we do not understand the genetic contributions to such behaviors, limiting the available strategies for treatment. In my research, I am motivated to understand molecular changes that underlie childhood brain diseases. The field of bioinformatics provides tools we can use to understand disease processes through the analysis of molecular sequence data. More broadly, bioinformatics facilitates our understanding of the basic aspects of biology including development, metabolism, adaptation to the environment, genetics (e.g., the basis of individual differences), and evolution.

Since the publication of the first edition of this textbook in 2003, the fields of bioinformatics and genomics have grown explosively. In the preface to the first edition (2003) I noted that tens of billions of base pairs (gigabases) of DNA had been deposited in GenBank. Now in 2009 we are reaching tens of trillions (terabases) of DNA, presenting us with unprecedented challenges in how to store, analyze, and interpret sequence data. In this second edition I have made numerous changes to the content and organization of the book. All of the chapters are rewritten, and about 90% of the figures and tables are updated. There are two new chapters, one on functional genomics and one on the eukaryotic chromosome. I now focus on the globins as examples throughout the book. Globins have a special place in the history of biology, as they were among the first proteins to be identified (in the 1830s) and sequenced (in the 1950s and 1960s). The first protein to have its structure solved by X-ray crystallography was myoglobin (Chapter 11); molecular phylogeny was applied to the globins in the 1960s (Chapter 7); and the globin gene loci were among the first to be sequenced (in the 1980s; see Chapter 16).

The fields of bioinformatics and genomics are far too broad to be understood by one person. Thus many textbooks are written by multiple authors, each of whom brings a deeper knowledge of the subject matter. I hope that this book at least offers the benefit of a single author's vision of how to present the material. This is essentially two textbooks: one on bioinformatics (parts I and II) and one on genomics (part III). I feel that presenting bioinformatics on its own would be incomplete without further applying those approaches to sequence analysis of genomes across the tree of life. Similarly I feel that it is not possible to approach genomics without first treating the bioinformatics tools that are essential engines of that field.

As with the previous edition a companion website is available which provides up-to-date web links referred to in the book and PowerPoint slides arranged by

chapter (www.bioinfbook.org). A resource site for instructors is also available giving detailed solutions to problems (www.wiley.com/go/pevsnerbioinformatics).

In preparing each edition of this book I read many papers and reviewed several thousand websites. I sincerely apologize to those authors, researchers and others whose work I did not cite. It is a great pleasure to acknowledge my colleagues who have helped in the preparation of this book. Some read chapters including Jef Boeke (Chapter 12), Rafael Irizarry (Chapter 9), Stuart Ray (Chapter 7), Ingo Ruczinski (Chapter 11), and Sarah Wheelan (Chapters 3 and 5–7). I thank many students and faculty at Johns Hopkins and elsewhere who have provided critical feedback, including those who have lectured in bioinformatics and genomics courses (Judith Bender, Jef Boeke, Egbert Hoiczyk, Ingo Ruczinski, Alan Scott, David Sullivan, David Valle, and Sarah Wheelan). Many others engaged in helpful discussions including Charles D. Cohen, Bob Cole, Donald Coppock, Laurence Frelin, Hugh Gelch, Gary W. Goldstein, Marjan Gucek, Ada Hamosh, Nathaniel Miller, Akhilesh Pandey, Elisha Roberson, Kirby D. Smith, Jason Ting, and N. Varg. I thank my wife Barbara for her support and love as I prepared this book.

# Preface to the First Edition

## Origins of this Book

This book emerged from lecture notes I prepared several years ago for an introductory bioinformatics and genomics course at the Johns Hopkins School of Medicine. The first class consisted of about 70 graduate students and several hundred auditors, including postdoctoral fellows, technicians, undergraduates, and faculty. Those who attended the course came from a broad variety of fields—students of genetics, neuroscience, immunology or cell biology, clinicians interested in particular diseases, statisticians and computer scientists, virologists and microbiologists. They had a common interest in wanting to understand how they could apply the tools of computer science to solve biological problems. This is the domain of bioinformatics, which I define most simply as the interface of computer science and molecular biology. This emerging field relies on the use of computer algorithms and computer databases to study proteins, genes, and genomes. Functional genomics is the study of gene function using genome-wide experimental and computational approaches.

## Comparison

At its essence, the field of bioinformatics is about comparisons. In the first third of the book we learn how to extract DNA or protein sequences from the databases, and then to compare them to each other in a pairwise fashion or by searching an entire database. For the student who has a gene of particular interest, a natural question is to ask "what other genes (or proteins) are related to mine?"

In the middle third of the book, we move from DNA to RNA (gene expression) and to proteins. We again are engaged in a series of comparisons. We compare gene expression in two cell lines with or without drug treatment, or a wildtype mouse heart versus a knockout mouse heart, or a frog at different stages of development. These comparisons extend to the world of proteins, where we apply the tools of proteomics to complex biological samples under assorted physiological conditions. The alignment of multiple, related DNA or protein sequences is another form of comparison. These relationships can be visualized in a phylogenetic tree.

The last third of the book spans the tree of life, and this provides another level of comparison. Which forms of human immunodeficiency virus threaten us, and how can we compare the various HIV subtypes to learn how we might develop a vaccine? How are a mosquito and a fruitfly related? What genes do vertebrates such as fish and humans share in common, and which genes are unique to various phylogenetic lineages?

I believe that these various kinds of comparisons are what distinguish the newly emerging fields of bioinformatics and genomics from traditional biology. Biology has always concerned comparisons; in this book I quote 19th century biologists such as Richard Owen, Ernst Haeckel, and Charles Darwin who engaged in comparative studies at the organismal level. The problems we are trying to solve have not changed substantially. We still seek a more complete understanding of the unifying concepts of biology, such as the organization of life from its constituent parts (e.g., genes and proteins), the behavior of complex biological systems, and the continuity of life through evolution. What *has* changed is how we pursue this more complete understanding. This book describes databases filled with raw information on genes and gene products and the tools that are useful to analyze these data.

## The Challenge of Human Disease

My training is as a molecular biologist and neuroscientist. My laboratory studies the molecular basis of childhood brain disorders such as Down syndrome, autism, and lead poisoning. We are located at the Kennedy Krieger Institute, a hospital for children for developmental disorders. (You can learn more about this Institute at http://www.kennedykrieger.org.) Each year over 10,000 patients visit the Institute. The hospital includes clinics for children with a variety of conditions including language disorders, eating disorders, autism, mental retardation, spina bifida, and traumatic brain injury. Some have very common disorders, such as Down syndrome (affecting about 1:700 live births) and mental retardation. Others have rare disorders, such as Rett syndrome or adrenoleukodystrophy.

We are at a time when the number of base pairs of DNA deposited in the world's public repositories has reached tens of billions, as described in Chapter 2. We have obtained the first sequence of the human genome, and since 1995 hundreds of genomes have been sequenced. Throughout the book, you can follow the progress of science as we learn how to sequence DNA, and study its RNA and protein products. At times the pace of progress seems dazzling.

Yet at the same time we understand so little about human disease. For thousands of diseases, a defect in a single gene causes a pathological effect. Even as we discover the genes that are defective in diseases such as cystic fibrosis, muscular dystrophy, adrenoleukodystrophy, and Rett syndrome, the path to finding an effective treatment or cure is obscure. But single gene disorders are not nearly as common as complex diseases such as autism, depression, and mental retardation that are likely due to mutations in multiple genes. And all genetic disease is not nearly as common as infectious disease. We know little about why one strain of virus infects only humans, while another closely related species infects only chimpanzees. We do not understand why one bacterial strain may be pathogenic, while another is harmless. We have not learned how to develop an effective vaccine against any eukaryotic pathogen, from protozoa (such as *Plasmodium falciparum* that causes malaria) to parasitic nematodes.

The prospects for making progress in these areas are very encouraging specifically because of the recent development of new bioinformatics tools. We are only now beginning to position ourselves to understand the genetic basis of both disease-causing agents and the hosts that are susceptible. Our hope is that the information so rapidly accumulating in new bioinformatics databases can be translated through research into insights into human disease and biology in general.

## Note to Readers

This book describes over 1,000 websites related to bioinformatics and functional genomics. All of these sites evolve over time (and some become extinct). In an effort to keep the web links up-to-date, a companion website (http://www.bioinfbook.org) maintains essentially all of the website links, organized by chapter of the book. We try our best to maintain this site over time. We use a program to automatically scan all the links each month, and then we update them as necessary.

An additional site is available to instructors, including detailed solutions to problems (see http://www.wiley.com).

## Acknowledgments

Writing this book has been a wonderful learning experience. It is a pleasure to thank the many people who have contributed. In particular, the intellectual environment at the Kennedy Krieger Institute and the Johns Hopkins School of Medicine has been extraordinarily rich. These chapters were developed from lectures in an introductory bioinformatics course. The Johns Hopkins faculty who lectured during its first three years were Jef Boeke (yeast functional genomics), Aravinda Chakravarti (human disease), Neil Clarke (protein structure), Kyle Cunningham (yeast), Garry Cutting (human disease), Rachel Green (RNA), Stuart Ray (molecular phylogeny), and Roger Reeves (the human genome). I have benefited greatly from their insights into these areas.

I gratefully acknowledge the many reviewers of this book, including a group of anonymous reviewers who offered extremely constructive and detailed suggestions. Those who read the book include Russ Altman, Christopher Aston, David P. Leader, and Harold Lehmann (various chapters), Conover Talbot (Chapters 2 and 18), Edie Sears (Chapter 3), Tom Downey (Chapter 7), Jef Boeke (Chapter 8 and various other chapters), Michelle Nihei and Daniel Yuan (Chapter 8), Mario Amzel and Ingo Ruczinski (Chapter 9), Stuart Ray (Chapter 11), Marie Hardwick (Chapter 13), Yukari Manabe (Chapter 14), Kyle Cunningham and Forrest Spencer (Chapter 15), and Roger Reeves (Chapter 16). Kirby D. Smith read Chapter 18 and provided insights into most of the other chapters as well. Each of these colleagues offered a great deal of time and effort to help improve the content, and each served as a mentor. Of the many students who read the chapters I mention Rong Mao, Ok-Hee Jeon, and Vinoy Prasad. I particularly thank Mayra Garcia and Larry Frelin who provided invaluable assistance throughout the writing process. I am grateful to my editor at John Wiley & Sons, Luna Han, for her encouragement.

I also acknowledge Gary W. Goldstein, President of the Kennedy Krieger Institute, and Solomon H. Snyder, my chairman in the Department of Neuroscience at Johns Hopkins. Both provided encouragement, and allowed me the opportunity to write this book while maintaining an academic laboratory.

On a personal note, I thank my family for all their love and support, as well as N. Varg, Kimberly Reed, and Charles Cohen. Most of all, I thank my fiancée Barbara Reed for her patience, faith, and love.

# Foreword

Ask 10 investigators in human genetics what resources they need most and it is highly likely that computational skills and tools will be at the top of the list. Genomics, with its reliance on microarrays, genotyping, high throughput sequencing and the like, is intensely data-rich and for this reason is impossible to disentangle from bioinformatics. This text, with its clear descriptions, practical examples and focus on the overlaps and interdependence of these two fields, is thus an essential resource for students and practitioners alike.

Interestingly, bioinformatics and genomics are both relatively recent disciplines. Each emerged in the course of the Human Genome Project (HGP) that was conceived in the mid-1980s and began officially on October 1, 1990. As the HGP matured from its initial focus on gene maps in model organisms to the massive efforts to produce a reference human whole genome sequence, there was an increasing need for computational biology tools to store, analyze and disseminate large amounts of sequence data. For this reason, genomics increasingly relied on bioinformatics and, in turn, the field of bioinformatics flourished. Today, no serious student of genomics can imagine life without bioinformatics. This interdependence continues to grow by leaps and bounds as the questions and activities of investigators in genomics become bolder and more expansive; consider, for example, whole genome association studies (GWAS), the ENCODE project, the challenge of copy number variants, the 1000 Genomes project, epigenomics, and the looming growth of personal genome sequences and their analysis.

This textbook provides a clear and timely introduction to both bioinformatics and genomics. It is organized so that each chapter can correspond to a lecture for a course on bioinformatics or genomics and, indeed, we have used it this way for our students. Also, for readers not taking courses, the book provides essential background material. For computer scientists and biologists alike the book offers explanations of available methods and the kinds of problems for which they can be used. The sections on bioinformatics in the first part of the book describe many of the basic tools that are used to analyze and compare DNA and protein sequences. The tone is inviting as the reader is guided to learn to use different software by example. Multiple approaches for solving particular problems, such as sequence alignment and molecular phylogeny, are presented. The middle part of the book introduces functional genomics. Here again the focus is on helping the reader to learn how to do analyses (such as microarray data analysis or protein structure prediction) in a practical way. A companion website provides many data sets, so the student can get experience in performing analyses. Chapter 12 provides a roadmap to the very complicated topic of functional genomics, spanning a range of techniques and model organisms used to study gene function. The last third of

the book provides a survey of the tree of life from a genomics perspective. There is an attempt to be comprehensive, and at the same time, to present the material in an interesting way, highlighting the fascinating features that make each genome unique.

Far from being a dry account of the facts of genomics and bioinformatics, the book offers many features that highlight the vitality of this field. There are discussions throughout about how to critically evaluate the performance of different software. For example, there are 'competitions' in which different research groups perform computational analyses on data sets that have been validated with some 'gold standard', allowing false positive and false negative error rates to be determined. These competitions are described in areas such as microarray data analysis (Chapter 9), mass spectrometry (Chapter 10), protein structure prediction (Chapter 11), or gene prediction (Chapter 16). The book also includes descriptions of important movements in the fields of bioinformatics and genomics, ranging from the RefSeq project for organizing sequences to the ENCODE and HapMap projects. Similarly, there is a rich description of the historical context for different aspects of bioinformatics and genomics, such as Garrod's views on disease (Chapter 20); Ohno's classic 1970 book on genome duplication (Chapter 17); and, the earliest attempts to create alignments and phylogenetic trees of the globins.

Where will the fields of bioinformatics and genomics go in the next five to 10 years? The opportunities are vast and any prediction will certainly be incomplete, but it is certain that the rapid technological advances in sequencing will provide an unprecedented view of human genetic variation and how this relates to phenotype. In the area of human disease studies, genome-wide association studies can be expected to lead to the identification of hundreds of genes underlying complex disorders. Finally, our understanding of evolution and its relevance to medicine will expand dramatically. Dr Pevsner's valuable book will help the student or researcher access the tools and learn the principles that will enable this exciting research.

David Valle, M.D.
*Henry J. Knott Professor and Director McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine*

# Part I

# Analyzing DNA, RNA, and Protein Sequences in Databases

account of this very identity of composition. Hence the opinion is not unworthy of a closer investigation, that gelatine, when taken in the dissolved state, is again converted, in the body, into cellular tissue, membrane and cartilage; that it may serve for the reproduction of such parts of these tissues as have been wasted, and for their growth.

And when the powers of nutrition in the whole body are affected by a change of the health, then, even should the power of forming blood remain the same, the organic force by which the constituents of the blood are transformed into cellular tissue and membranes must necessarily be enfeebled by sickness. In the sick man, the intensity of the vital force, its power to produce metamorphoses, must be diminished as well in the stomach as in all other parts of the body. In this condition, the uniform experience of practical physicians shows that gelatinous matters in a dissolved state exercise a most decided influence on the state of the health. Given in a form adapted for assimilation, they serve to husband the vital force, just as may be done, in the case of the stomach, by due preparation of the food in general. Brittleness in the bones of graminivorous animals is clearly owing to a weakness in those parts of the organism whose function it is to convert the constituents of the blood into cellular tissue and membrane; and if we can trust to the reports of physicians who have resided in the East, the Turkish women, in their diet of rice, and in the frequent use of enemata of strong soup, have united the conditions necessary for the formation both of cellular tissue and of fat.

---

# PART II.

## THE METAMORPHOSIS OF TISSUES.

1. The absolute identity of composition in the chief constituents of blood and the nitrogenized compounds in vegetable food would, some years ago, have furnished a plausible reason for denying the accuracy of the chemical analysis leading to such a result. At that period, experiment had not as yet demonstrated the existence of numerous compounds, both containing nitrogen and devoid of that element, which with the greatest diversity in external characters, yet possess the very same composition in 100 parts; nay, many of which even contain the same absolute amount of equivalents of each element. Such examples are now very frequent, and are known by the names of *isomeric* and *polymeric* compounds.

2. Cyanuric acid, for example, is a nitrogenized compound which crystallizes in beautiful transparent octahedrons, easily soluble in water and in acids, and very permanent. Cyamelide is a second body, absolutely insoluble in water and acids, white and opaque like porcelain or magnesia. Hydrated cyanic acid is a third compound, which is a liquid more volatile than pure acetic acid, which blisters the skin, and cannot be brought in contact with water without being instantaneously resolved into new products. These three substances not only yield, on analysis, absolutely the same relative weights of the same elements, but they may be converted and reconverted into one another, even in hermetically closed vessels —that is, without the aid of any foreign matter. (See Appendix, 21.) Again, among those substances which contain no nitrogen, we have aldehyde, a combustible liquid miscible with water, which boils at the temperature of the hand, attracts oxygen from the atmosphere with avidity, and is thereby changed into acetic acid. This compound cannot be preserved, even in close vessels; for after some hours or days, its consistence, its volatility, and its power of absorbing oxygen, all are changed. It deposits long, hard, needle-shaped crystals, which at 212° are not volatilized, and the supernatant liquid is no longer aldehyde. It now boils at 140°, cannot be mixed with water, and when cooled to a moderate degree crystallizes in a form like ice. Nevertheless, analysis has proved, that these three bodies, so different in their characters, are identical in composition. (21.)

3. A similar group of three occurs in the case of albumen, fibrine, and caseine. They ◀ differ in external character, but contain exactly the same proportions of organic elements.

When animal albumen, fibrine, and caseine are dissolved in a moderately strong solution of caustic potash, and the solution is exposed for some time to a high temperature, these substances are decomposed. The addition of acetic acid to the solution causes, in all three, the separation of a gelatinous translucent precipitate, which has exactly the same characters and composition, from whichever of the three substances above mentioned it has been obtained.

Mulder, to whom we owe the discovery of this compound, found, by exact and careful analysis, that it contains the same organic elements, and exactly in the same proportion, as the animal matters from which it is prepared; insomuch, that if we deduct from the analysis of albumen, fibrine, and caseine, the ashes they yield when incinerated, as well as the sulphur and phosphorus they contain, and then calculate the remainder for 100 parts, we obtain the same result as

*The study of bioinformatics includes the analysis of proteins. In the first half of the nineteenth century the Dutch researcher Gerardus Johannes Mulder (1802–1880), advised by the Swedish chemist Jöns Jacob Berzelius (1779–1848), studied the "albuminous" substances or proteins fibrin, albumin from blood, albumin from egg (ovalbumin), and the coloring matter of blood (hemoglobin). Mulder and others extracted and purified these proteins and believed that they all shared the same elemental composition ($C_{400}H_{260}N_{100}O_{120}$), with varying amounts of phosphorus and sulfur. Justus Liebig (1803–1873) believed that the composition of protein was $C_{48}H_{36}N_6O_{14}$. This page, from Liebig's Animal Chemistry, or Organic Chemistry in its Applications to Physiology and Pathology (1847, p. 36), discusses albumin, fibrin, and casein (see arrowhead).*

# 1

# Introduction

Bioinformatics represents a new field at the interface of the twentieth-century revolutions in molecular biology and computers. A focus of this new discipline is the use of computer databases and computer algorithms to analyze proteins, genes, and the complete collections of deoxyribonucleic acid (DNA) that comprises an organism (the genome). A major challenge in biology is to make sense of the enormous quantities of sequence data and structural data that are generated by genome-sequencing projects, proteomics, and other large-scale molecular biology efforts. The tools of bioinformatics include computer programs that help to reveal fundamental mechanisms underlying biological problems related to the structure and function of macromolecules, biochemical pathways, disease processes, and evolution.

According to a National Institutes of Health (NIH) definition, bioinformatics is "research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, analyze, or visualize such data." The related discipline of computational biology is "the development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems."

The NIH Bioinformatics Definition Committee findings are reported at ► http://www.bisti.nih.gov/CompuBioDef.pdf. For additional definitions of bioinformatics and functional genomics, see Boguski (1994), Luscombe et al. (2001), Ideker et al. (2001), and Goodman (2002).

While the discipline of bioinformatics focuses on the analysis of molecular sequences, genomics and functional genomics are two closely related disciplines. The goal of genomics is to determine and analyze the complete DNA sequence of an organism, that is, its genome. The DNA encodes genes, which can be expressed as ribonucleic acid (RNA) transcripts and then in many cases further translated into

*Bioinformatics and Functional Genomics, Second Edition*. By Jonathan Pevsner

protein. Functional genomics describes the use of genomewide assays in the study of gene and protein function.

The aim of this book is to explain both the theory and practice of bioinformatics and genomics. The book is especially designed to help the biology student use computer programs and databases to solve biological problems related to proteins, genes, and genomes. Bioinformatics is an integrative discipline, and our focus on individual proteins and genes is part of a larger effort to understand broad issues in biology, such as the relationship of structure to function, development, and disease. For the computer scientist, this book explains the motivations for creating and using algorithms and databases.

## Organization of the Book

There are three main sections of the book. The first part (Chapters 2 to 7) explains how to access biological sequence data, particularly DNA and protein sequences (Chapter 2). Once sequences are obtained, we show how to compare two sequences (pairwise alignment; Chapter 3) and how to compare multiple sequences (primarily by the Basic Local Alignment Search Tool [BLAST]; Chapters 4 and 5). We introduce multiple sequence alignment (Chapter 6) and show how multiply aligned sequences can be visualized in phylogenetic trees (Chapter 7). Chapter 7 thus introduces the subject of molecular evolution.

The second part of the book describes functional genomics approaches to RNA and protein and the determination of gene function (Chapters 8 to 12). The central dogma of biology states that DNA is transcribed into RNA then translated into protein. We will examine bioinformatic approaches to RNA, including both noncoding and coding RNAs. We then describe the technology of DNA microarrays and examine microarray data analysis (Chapter 9). From RNA we turn to consider proteins from the perspective of protein families, and the analysis of individual proteins (Chapter 10) and protein structure (Chapter 11). We conclude the middle part of the book with an overview of the rapidly developing field of functional genomics (Chapter 12).

Since 1995, the genomes have been sequenced for several thousand viruses, prokaryotes (bacteria and archaea), and eukaryotes, such as fungi, animals, and plants. The third section of the book covers genome analysis (Chapters 13 to 20). Chapter 13 provides an overview of the study of completed genomes and then descriptions of how the tools of bioinformatics can elucidate the tree of life. We describe bioinformatics resources for the study of viruses (Chapter 14) and bacteria and archaea (Chapter 15; these are two of the three main branches of life). Next we examine the eukaryotic chromosome (Chapter 16) and explore the genomes of a variety of eukaryotes, including fungi (Chapter 17), organisms from parasites to primates (Chapter 18), and then the human genome (Chapter 19). Finally, we explore bioinformatic approaches to human disease (Chapter 20).

## Bioinformatics: The Big Picture

We can summarize the fields of bioinformatics and genomics with three perspectives. The first perspective on bioinformatics is the cell (Fig. 1.1). The central dogma of molecular biology is that DNA is transcribed into RNA and translated into protein. The focus of molecular biology has been on individual genes, messenger RNA

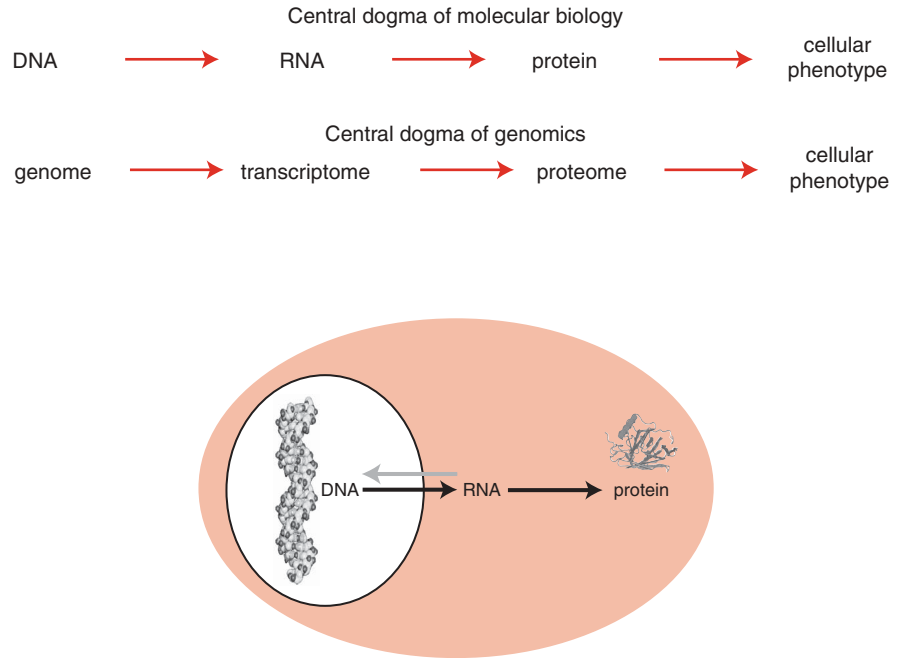



*FIGURE 1.1. The first perspective of the field of bioinformatics is the cell. Bioinformatics has emerged as a discipline as biology has become transformed by the emergence of molecular sequence data. Databases such as the European Molecular Biology Laboratory (EMBL), GenBank, and the DNA Database of Japan (DDBJ) serve as repositories for hundreds of billions of nucleotides of DNA sequence data (see Chapter 2). Corresponding databases of expressed genes (RNA) and protein have been established. A main focus of the field of bioinformatics is to study molecular sequence data to gain insight into a broad range of biological problems.*

(mRNA) transcripts as well as noncoding RNAs, and proteins. A focus of the field of bioinformatics is the complete collection of DNA (the genome), RNA (the transcriptome), and protein sequences (the proteome) that have been amassed (Henikoff, 2002). These millions of molecular sequences present both great opportunities and great challenges. A bioinformatics approach to molecular sequence data involves the application of computer algorithms and computer databases to molecular and

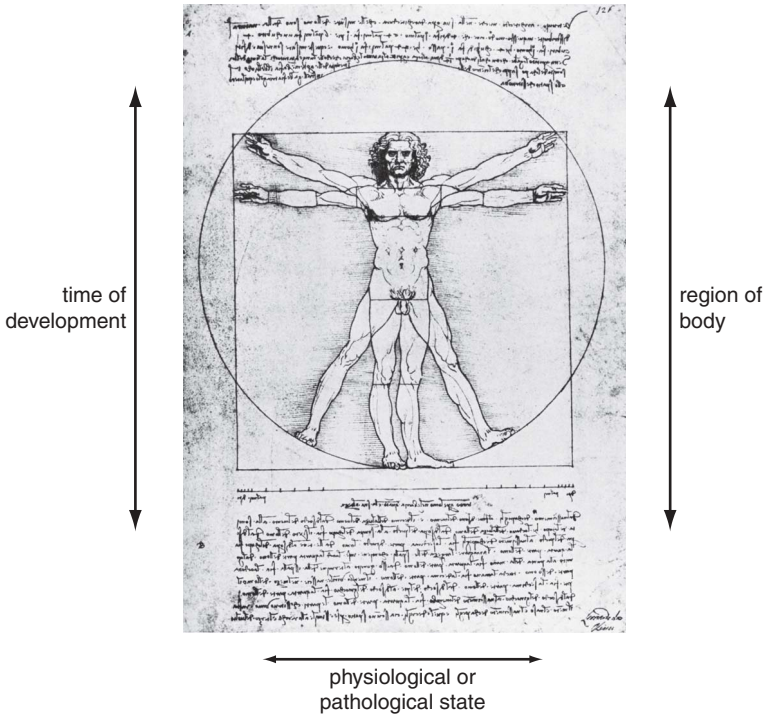



*FIGURE 1.2. The second perspective of bioinformatics is the organism. Broadening our view from the level of the cell to the organism, we can consider the individual's genome (collection of genes), including the genes that are expressed as RNA transcripts and the protein products. Thus, for an individual organism bioinformatics tools can be applied to describe changes through developmental time, changes across body regions, and changes in a variety of physiological or pathological states.*

cellular biology. Such an approach is sometimes referred to as functional genomics. This typifies the essential nature of bioinformatics: biological questions can be approached from levels ranging from single genes and proteins to cellular pathways and networks or even whole genomic responses (Ideker et al., 2001). Our goals are to understand how to study both individual genes and proteins and collections of thousands of genes or proteins.

From the cell we can focus on individual organisms, which represents a second perspective of the field of bioinformatics (Fig. 1.2). Each organism changes across different stages of development and (for multicellular organisms) across different regions of the body. For example, while we may sometimes think of genes as static entities that specify features such as eye color or height, they are in fact dynamically regulated across time and region and in response to physiological state. Gene expression varies in disease states or in response to a variety of signals, both intrinsic and environmental. Many bioinformatics tools are available to study the broad biological questions relevant to the individual: there are many databases of expressed

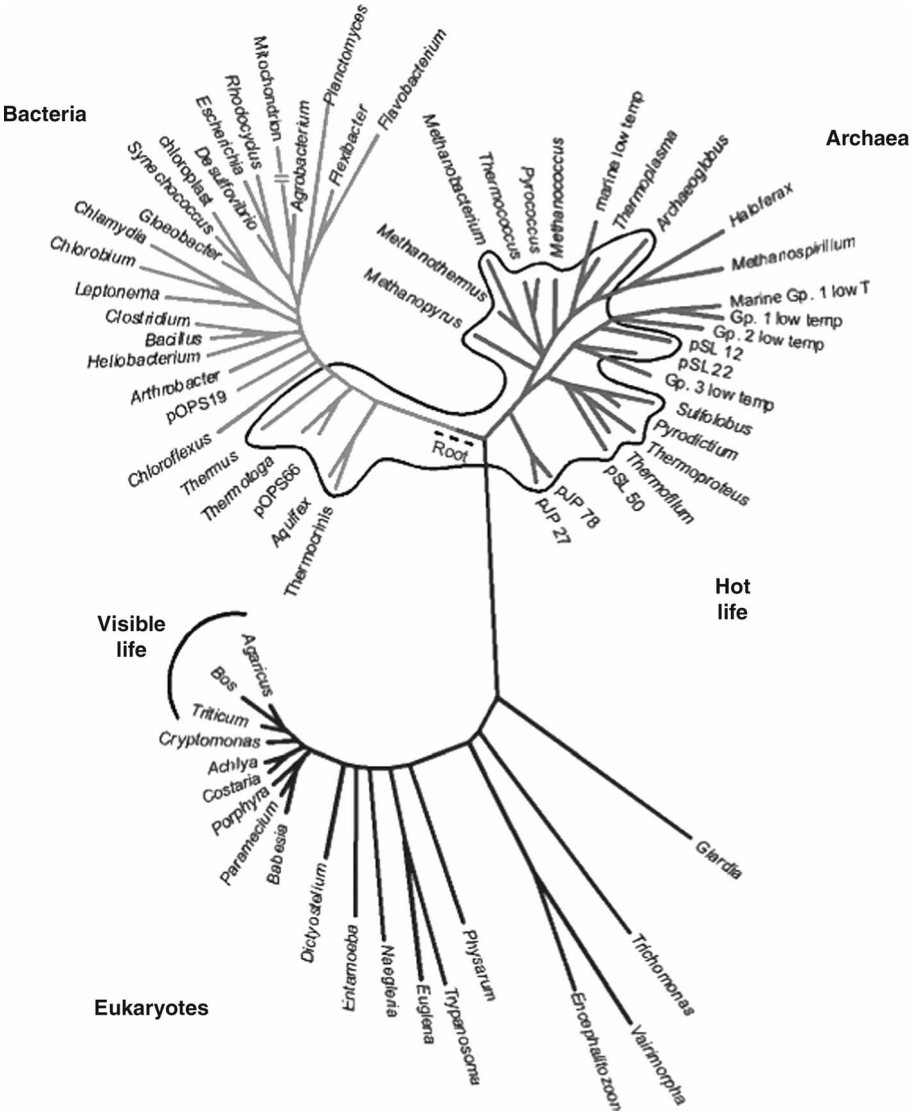
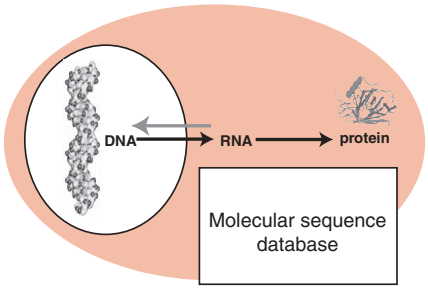



*FIGURE 1.3. The third perspective of the field of bioinformatics is represented by the tree of life. The scope of bioinformatics includes all of life on Earth, including the three major branches of bacteria, archaea, and eukaryotes. Viruses, which exist on the borderline of the definition of life, are not depicted here. For all species, the collection and analysis of molecular sequence data allow us to describe the complete collection of DNA that comprises each organism (the genome). We can further learn the variations that occur between species and among members of a species, and we can deduce the evolutionary history of life on Earth. (After Barns et al., 1996 and Pace, 1997.) Used with permission.*

genes and proteins derived from different tissues and conditions. One of the most powerful applications of functional genomics is the use of DNA microarrays to measure the expression of thousands of genes in biological samples.

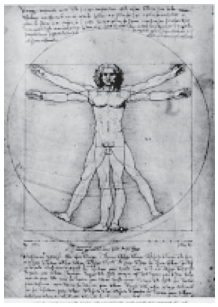
At the largest scale is the tree of life (Fig. 1.3) (Chapter 13). There are many millions of species alive today, and they can be grouped into the three major branches of bacteria, archaea (single-celled microbes that tend to live in extreme environments), and eukaryotes. Molecular sequence databases currently hold DNA sequences from over 150,000 different organisms. The complete genome sequences of thousands of organisms are now available, including organellar and viral genomes. One of the main lessons we are learning is the fundamental unity of life at the molecular level. We are also coming to appreciate the power of comparative genomics, in which genomes are compared. Through DNA sequence analysis we are learning how chromosomes evolve and are sculpted through processes such as chromosomal duplications, deletions, and rearrangements, as well as through whole genome duplications (Chapters 16 to 18).

Figure 1.4 presents the contents of this book in the context of these three perspectives of bioinformatics.



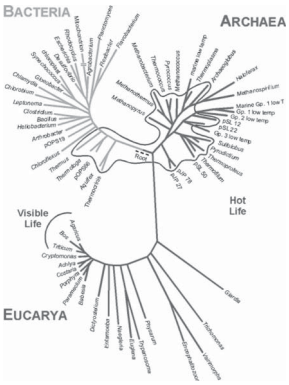

Part 1: Analyzing DNA, RNA, and protein sequences

Chapter 1: Introduction
Chapter 2: How to obtain sequences
Chapter 3: How to compare two sequences
Chapters 4 and 5: How to compare a sequence to all other sequences in databases
Chapter 6: How to multiply align sequences
Chapter 7: How to view multiply aligned sequences as phylogenetic trees



Part 2: Genome-wide analysis of RNA and protein

Chapter 8: Bioinformatics approaches to RNA
Chapter 9: Microarray data analysis
Chapter 10: Protein analysis and protein families
Chapter 11: Protein structure
Chapter 12: Functional genomics




Part 3: Genome analysis

Chapter 13: The tree of life
Chapter 14: Viruses
Chapter 15: Prokaryotes
Chapter 16: The eukaryotic chromosome
Chapter 17: The fungi
Chapter 18: Eukaryotes from parasites to plants to primates
Chapter 19: The human genome
Chapter 20: Human disease

*FIGURE 1.4. Overview of the chapters in this book.*

## A Consistent Example: Hemoglobin

Throughout this book, we will focus on the globin gene family to provide a consistent example of bioinformatics and genomics concepts. The globin family is one of the best characterized in biology.

- Historically, hemoglobin was one of the first proteins to be studied, having been described in the 1830s and 1840s by Mulder, Liebig, and others.
- Myoglobin, a globin that binds oxygen in the muscle tissue, was the first protein to have its structure solved by x-ray crystallography (Chapter 11).
- Hemoglobin, a tetramer of four globin subunits (principally $\alpha_2\beta_2$ in adults), is the main oxygen carrier in blood of vertebrates. Its structure was also one of the earliest to be described. The comparison of myoglobin, alpha globin, and beta globin protein sequences represents one of the earliest applications of multiple sequence alignment (Chapter 6), and led to the development of amino acid substitution matrices used to score protein relatedness (Chapter 3).
- In the 1980s as DNA sequencing technology emerged, the globin loci on human chromosomes 16 (for $\alpha$ globin) and 11 (for $\beta$ globin) were among the first to be sequenced and analyzed. The globin genes are exquisitely regulated across time (switching from embryonic to fetal to adult forms) and with tissue-specific gene expression. We will discuss these loci in the description of the control of gene expression (Chapter 16).
- While hemoglobin and myoglobin remain the best-characterized globins, the family of homologous proteins extends to two separate classes of plant globins, invertebrate hemoglobins (some of which contain multiple globin domains within one protein molecule), bacterial homodimeric hemoglobins (consisting of two globin subunits), and flavohemoglobins that occur in bacteria, archaea, and fungi. Thus the globin family is useful as we survey the tree of life (Chapters 13 to 18).

Another protein we will use as an example is retinol-binding protein (RBP4), a small, abundant secreted protein that binds retinol (vitamin A) in blood (Newcomer and Ong, 2000). Retinol, obtained from carrots in the form of vitamin A, is very hydrophobic. RBP4 helps transport this ligand to the eye where it is used for vision. We will study RBP4 in detail because it has a number of interesting features:

- There are many proteins that are homologous to RBP4 in a variety of species, including human, mouse, and fish ("orthologs"). We will use these as examples of how to align proteins, perform database searches, and study phylogeny.
- There are other human proteins that are closely related to RBP4 ("paralogs"). Altogether the family that includes RBP4 is called the lipocalins, a diverse group of small ligand-binding proteins that tend to be secreted into extracellular spaces (Akerstrom et al., 2000; Flower et al., 2000). Other lipocalins have fascinating functions such as apoliprotein D (which binds cholesterol), a pregnancy-associated lipocalin, aphrodisin (an "aphrodisiac" in hamsters), and an odorant-binding protein in mucus.

- There are bacterial lipocalins, which could have a role in antibiotic resistance (Bishop, 2000). We will explore how bacterial lipocalins could be ancient genes that entered eukaryotic genomes by a process called lateral gene transfer.
- Because the lipocalins are small, abundant, and soluble proteins, their biochemical properties have been characterized in detail. The three-dimensional protein structure has been solved for several of them by x-ray crystallography (Chapter 11).
- Some lipocalins have been implicated in human disease.

## Organization of the Chapters

The chapters of this book are intended to provide both the theory of bioinformatics subjects as well as a practical guide to using computer databases and algorithms. Web resources are provided throughout each chapter. Chapters end with brief sections called Perspective and Pitfalls. The perspective feature describes the rate of growth of the subject matter in each chapter. For example, a perspective on Chapter 2 (access to sequence information) is that the amount of DNA sequence data deposited in GenBank is undergoing an explosive rate of growth. In contrast, an area such as pairwise sequence alignment, which is fundamental to the entire field of bioinformatics (Chapter 3), was firmly established in the 1970s and 1980s. But even for fundamental operations such as multiple sequence alignment (Chapter 6) and molecular phylogeny (Chapter 7) dozens of novel, ever-improving approaches are introduced at a rapid rate. For example, hidden Markov models and Bayesian approaches are being applied to a wide range of bioinformatics problems.

The pitfalls section of each chapter describes some common difficulties encountered by biologists using bioinformatics tools. Some errors might seem trivial, such as searching a DNA database with a protein sequence. Other pitfalls are more subtle, such as artifacts caused by multiple sequence alignment programs depending upon the type of paramters that are selected. Indeed, while the field of bioinformatics depends substantially on analyzing sequence data, it is important to recognize that there are many categories of errors associated with data generation, collection, storage, and analysis. We address the problems of false positive and false negative results in a variety of searches and analyses.

Each chapter offers multiple-choice quizzes, which test your understanding of the chapter materials. There are also problems that require you to apply the concepts presented in each chapter. These problems may form the basis of a computer laboratory for a bioinformatics or genomics course.

The references at the end of each chapter are accompanied by an annotated list of recommended articles. This suggested reading section includes classic papers that show how the principles described in each chapter were discovered. Particularly helpful review articles and research papers are highlighted.

## A Textbook for Courses on Bioinformatics and Genomics

This is a textbook for two separate courses: one is an introduction to bioinformatics (and uses Chapters 1 to 12 [Parts 1 and 2]), and one is an introduction to genomics (and uses Chapters 13 to 20 [Part 3]). In a sense, the discipline of bioinformatics

serves biology, facilitating ways of posing and then answering questions about proteins, genes, and genomes. The third part of this book surveys the tree of life from the perspective of genes and genomes. Progress in this field could not occur at its current pace without the bioinformatics tools described in the first parts of the book.

Often, students have a particular research area of interest, such as a gene, a physiological process, a disease, or a genome. It is hoped that in the process of studying globins and other specific proteins and genes throughout this book, students can also simultaneously apply the principles of bioinformatics to their own research questions.

Web material for this book is available at ► http//www.wiley.com/go/pevsnerbioinformatics.

In teaching courses on bioinformatics and genomics at Johns Hopkins, it has been helpful to complement lectures with computer labs. These labs and many other resources are posted on the website for this book (► http://www.bioinfbook.org). That site contains many relevant URLs, organized by chapter. Each chapter makes references to web documents posted on the site. For example, if you see a figure of a phylogenetic tree or a sequence alignment, you can easily retrieve the raw data and make the figure yourself.

Another feature of the Johns Hopkins bioinformatics course is that each student is required to discover a novel gene by the last day of the course. The student must begin with any protein sequence of interest and perform database searches to identify genomic DNA that encodes a protein no one has described before. This problem is described in detail in Chapter 5 (and summarized in web document 5.15 at ► http://www.bioinfbook.org/chapter5). The student thus chooses the name of the gene and its corresponding protein and describes information about the organism and evidence that the gene has not been described before. Then, the student creates a multiple sequence alignment of the new protein (or gene) and creates a phylogenetic tree showing its relation to other known sequences.

Each year, some beginning students are slightly apprehensive about accomplishing this exercise, but in the end all of them succeed. A benefit of this exercise is that it requires a student to actively use the principles of bioinformatics. Most students choose a gene (or protein) relevant to their own research area, while others find new lipocalins or globins.

For a genomics course, students select a genome of interest and describe five aspects in depth (described at the start of Chapter 13): (1) What are the basic feature of the genome, such as its size, number of chromosomes, and other features? (2) A comparative genomic analysis is performed to study the relation of the species to its neighbors. (3) The student describes biological principles that are learned through genome analysis. (4) The human disease relevance is described. (5) Bioinformatics aspects are described, such as key databases or algorithms used for genome analysis.

Teaching bioinformatics and genomics is notable for the diversity of students learning this new discipline. Each chapter provides background on the subject matter. For more advanced students, key research papers are cited at the end of each chapter. These papers are technical, and reading them along with the chapters will provide a deeper understanding of the material. The suggested reading section also includes review articles.

## Key Bioinformatics Websites

The field of bioinformatics relies heavily on the Internet as a place to access sequence data, to access software that is useful to analyze molecular data, and as a place to integrate different kinds of resources and information relevant to biology. We will

describe a variety of websites. Initially, we will focus on the three main publicly accessible databases that serve as repositories for DNA and protein data. In Chapter 2 we begin with the National Center for Biotechnology Information (NCBI), which hosts GenBank. The NCBI website offers a variety of other bioinformatics-related tools. We will gradually introduce the European Bioinformatics Institute (EBI) web server, which hosts a complementary DNA database (EMBL, the European Molecular Biology Laboratory database). We will also introduce the DNA Database of Japan (DDBJ). The research teams at GenBank, EMBL, and DDBJ share sequence data on a daily basis. Throughout this book we will highlight the key genome browser hosted by the University of California, Santa Cruz (UCSC). A general theme of the discipline of bioinformatics is that many databases are closely interconnected. Throughout the chapters of this book we will introduce over 1,000 additional websites that are relevant to bioinformatics.

## Suggested Reading

Overviews of the field of bioinformatics have been written by Mark Gerstein and colleagues (Luscombe et al., 2001), Claverie et al. 2001, and Yu et al. 2004. Kaminski 2000 also introduces bioinformatics, with practical suggestions of websites to visit. Russ Altman 1998 discusses the relevance of bioinformatics to medicine, while David Searls 2000 introduces bioinformatics tools for the study of genomes. An approach to learning about the current state of bioinformatics education is to read about the perspectives of the programs at Yale (Gerstein et al., 2007), Stanford (Altman and Klein, 2007), and in Australia (Cattley, 2004).

## References


Akerstrom, B., Flower, D. R., and Salier, J. P. Lipocalins: Unity in diversity. *Biochim. Biophys. Acta* **1482**, 1–8 (2000).

Altman, R. B. Bioinformatics in support of molecular medicine. *Proc. AMIA Symp.*, 53–61 (1998).

Altman, R. B., and Klein, T. E. Biomedical informatics training at Stanford in the 21st century. *J. Biomed. Inform.* **40**, 55–58 (2007).

Barns, S. M., Delwiche, C. F., Palmer, J. D., and Pace, N. R. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proc. Natl. Acad. Sci. USA* **93**, 9188–9193 (1996).

Bishop, R. E. The bacterial lipocalins. *Biochim. Biophys. Acta* **1482**, 73–83 (2000).

Boguski, M. S. Bioinformatics. *Curr. Opin. Genet. Dev.* **4**, 383–388 (1994).

Cattley, S. A review of bioinformatics degrees in Australia. *Brief. Bioinform.* **5**, 350–354 (2004).

Claverie, J. M., Abergel, C., Audic, S., and Ogata, H. Recent advances in computational genomics. *Pharmacogenomics* **2**, 361–372 (2001).

Flower, D. R., North, A. C., and Sansom, C. E. The lipocalin protein family: Structural and sequence overview. *Biochim. Biophys. Acta* **1482**, 9–24 (2000).

Gerstein, M., Greenbaum, D., Cheung, K., and Miller, P. L. An interdepartmental Ph.D. program in computational biology and bioinformatics: The Yale perspective. *J. Biomed Inform.* **40**, 73–79 (2007).

Goodman, N. Biological data becomes computer literate: New advances in bioinformatics. *Curr. Opin. Biotechnol.* **13**, 68–71 (2002).

Henikoff, S. Beyond the central dogma. *Bioinformatics* **18**, 223–225 (2002).

Ideker, T., Galitski, T., and Hood, L. A new approach to decoding life: Systems biology. *Annu. Rev. Genomics Hum. Genet.* **2**, 343–372 (2001).

Kaminski, N. Bioinformatics. A user's perspective. *Am. J. Respir. Cell Mol. Biol.* **23**, 705–711 (2000).

Liebig, J. *Animal Chemistry, or Organic Chemistry in its Applications to Physiology and Pathology.* James M. Campbell, Philadelphia, 1847.

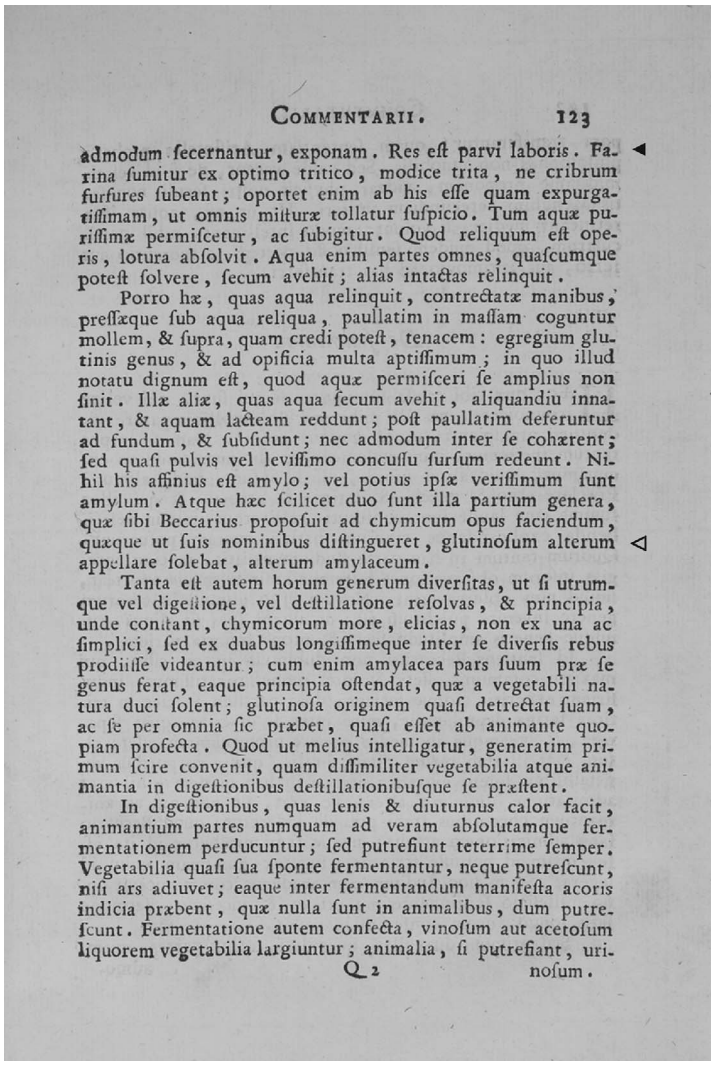
Luscombe, N. M., Greenbaum, D., and Gerstein, M. What is bioinformatics? A proposed definition and overview of the field. *Methods Inf. Med.* **40**, 346–358 (2001).

Newcomer, M. E., and Ong, D. E. Plasma retinol binding protein: Structure and function of the prototypic lipocalin. *Biochim. Biophys. Acta* **1482**, 57–64 (2000).

Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–740 (1997).

Searls, D. B. Bioinformatics tools for whole genomes. *Annu. Rev. Genomics Hum. Genet.* **1**, 251–279 (2000).

Yu, U., Lee, S. H., Kim, Y. J., and Kim, S. Bioinformatics in the post-genome era. *J. Biochem. Mol. Biol.* **37**, 75–82 (2004).



COMMENTARII. 123

admodum ſecernantur, exponam. Res eſt parvi laboris. Farina ſumitur ex optimo tritico, modice trita, ne cribrum furfures ſubeant; oportet enim ab his eſſe quam expurgatiſſimam, ut omnis miſturæ tollatur ſuſpicio. Tum aquæ puriſſimæ permiſcetur, ac ſubigitur. Quod reliquum eſt operis, lotura abſolvit. Aqua enim partes omnes, quaſcumque poteſt ſolvere, ſecum avehit; alias intactas relinquit.

Porro hæ, quas aqua relinquit, contrectatæ manibus, preſſæque ſub aqua reliqua, paullatim in maſſam coguntur mollem, & ſupra, quam credi poteſt, tenacem: egregium glutinis genus, & ad opificia multa aptiſſimum; in quo illud notatu dignum eſt, quod aquæ permiſceri ſe amplius non ſinit. Illæ aliæ, quas aqua ſecum avehit, aliquandiu innatant, & aquam lacteam reddunt; poſt paullatim deferuntur ad fundum, & ſubſidunt; nec admodum inter ſe cohærent; ſed quaſi pulvis vel leviſſimo concuſſu ſurſum redeunt. Nihil his affinius eſt amylo; vel potius ipſæ veriſſimum ſunt amylum. Atque hæc ſcilicet duo ſunt illa partium genera, quæ ſibi Beccarius propoſuit ad chymicum opus faciendum, quæque ut ſuis nominibus diſtingueret, glutinoſum alterum appellare ſolebat, alterum amylaceum.

Tanta eſt autem horum generum diverſitas, ut ſi utrumque vel digeſtione, vel deſtillatione reſolvas, & principia, unde conſtant, chymicorum more, elicias, non ex una ac ſimplici, ſed ex duabus longiſſimeque inter ſe diverſis rebus prodiiſſe videantur; cum enim amylacea pars ſuum præ ſe genus ferat, eaque principia oſtendat, quæ a vegetabili natura duci ſolent; glutinoſa originem quaſi detrectat ſuam, ac ſe per omnia ſic præbet, quaſi eſſet ab animante quopiam profecta. Quod ut melius intelligatur, generatim primum ſcire convenit, quam diſſimiliter vegetabilia atque animantia in digeſtionibus deſtillationibuſque ſe præſtent.

In digeſtionibus, quas lenis & diuturnus calor facit, animantium partes numquam ad veram abſolutamque fermentationem perducuntur; ſed putrefiunt teterrime ſemper. Vegetabilia quaſi ſua ſponte fermentantur, neque putreſcunt, niſi ars adiuvet; eaque inter fermentandum manifeſta acoris indicia præbent, quæ nulla ſunt in animalibus, dum putreſcunt. Fermentatione autem confecta, vinoſum aut acetoſum liquorem vegetabilia largiuntur; animalia, ſi putrefiant, uri-

Q 2 noſum.

*Chapter 2 introduces ways to access molecular data, including information about DNA and proteins. One of the first scientists to study proteins was Iacopo Bartolomeo Beccari (1682–1776), an Italian philosopher and physician who discovered protein as a component of vegetables. This image is from page 123 of the Bologna Commentaries, published in 1745 and written by a secretary on the basis of a 1728 lecture by Beccari. Beccari separated gluten (plant proteins) from wheaten flour. The passage beginning* Res est parvi laboris *("it is a thing of little labor"; see solid arrowhead) is translated as follows (Beach, 1961, p. 362):*

*"It is a thing of little labor. Flour is taken of the best wheat, moderately ground, the bran not passing though the sieve, for it is necessary that this be fully purged away, so that all traces of a mixture have been removed. Then it is mixed with pure water and kneaded. What is left by this procedure, washing clarifies. Water carries off with itself all it is able to dissolve, the rest remains untouched. After this, what the water leaves is worked with the hands, and pressed upon in the water that has stayed. Slowly it is drawn together in a doughy mass, and beyond what is possible to be believed, tenacious, a remarkable sort of glue, and suited to many uses; and what is especially worthy of note, it cannot any longer be mixed with water. The other particles, which water carries away with itself, for some time float and render the water milky; but after a while they are carried to the bottom and sink; nor in any way do they adhere to each other; but like powder they return upward on the lightest contact. Nothing is more like this than starch, or rather this truly is starch. And these are manifestly the two sorts of bodies which Beccari displayed through having done the work of a chemist and he distinguished them by their names, one being appropriately called glutinous (see open arrowhead) and the other amylaceous."*

*In addition to purifying gluten, Beccari identified it as an "animal substance" in contrast to starch, a "vegetable substance," based on differences on how they decomposed with heat or distillation. A century later Jons Jakob Berzelius proposed the word protein, and he also posited that plants form "animal materials" that are eaten by herbivorous animals.*

# 2

# Access to Sequence Data and Literature Information

## Introduction to Biological Databases

All living organisms are characterized by the capacity to reproduce and evolve. The genome of an organism is defined as the collection of DNA within that organism, including the set of genes that encode proteins. In 1995 the complete genome of a free-living organism was sequenced for the first time, the bacterium *Haemophilus influenzae* (Fleischmann et al., 1995; Chapters 13 and 15). In the few years since then the genomes of thousands of organisms have been completely sequenced, ushering in a new era of biological data acquisition and information accessibility. Publicly available databanks now contain billions of nucleotides of DNA sequence data collected from over 260,000 different organisms (Kulikova et al., 2007). The goal of this chapter is to introduce the databases that store these data and strategies to extract information from them.

Three publicly accessible databases store large amounts of nucleotide and protein sequence data: GenBank at the National Center for Biotechnology Information (NCBI) of the National Institutes of Health (NIH) in Bethesda (Benson et al., 2009), the DNA Database of Japan (DDBJ) at the National Institute of

*Bioinformatics and Functional Genomics, Second Edition*. By Jonathan Pevsner

GenBank is at ► http://www.ncbi.nlm.nih.gov/Genbank; DDBJ is at ► http://www.ddbj.nig.ac.jp/; and EMBL/EBI is at ► http://www.ebi.ac.uk/. You can visit the INSDC at ►http://www.insdc.org/. By November 2008 the total number of sequenced bases had passed 97 billion.

Genetics in Mishima (Miyazaki et al., 2004), and the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database at the European Bioinformatics Institute (EBI) in Hinxton, England (Kulikova et al., 2007). These three databases share their sequence data daily. They are coordinated by the International Nucleotide Sequence Database Collaboration (INSDC), which announced in August 2005 that the total amount of sequenced DNA had reached 100 billion base pairs.

In addition to GenBank, DDBJ, and EBI, there are other categories of bioinformatics databases that contain DNA and/or protein sequence data:

- Whole-genome shotgun (WGS) sequences and the Short Read Archive (Chapter 13 and discussed below) are not formally part of GenBank, but contain even more DNA sequences.
- Databases such as Ensembl, NCBI, and the genome browser at the University of California, Santa Cruz (UCSC) provide annotation of the human genome and other genomes (see below).
- Some contain nucleotide and/or protein sequence data that are relevant to a particular gene or protein (such as kinases). Other databases are specific to particular chromosomes or organelles (Chapters 16 to 18).
- A variety of databases include information on sequences sharing common properties that have been grouped together. For example, the Protein Family (Pfam) database consists of several thousand families of homologous proteins.
- Hundreds of databases contain sequence information related to genes that are mutated in human disease. These databases are described in Chapter 20.
- Many specialized databases focus on particular organisms (such as yeast); examples are listed in the section on genomes (Chapters 13 to 20).
- There are databases devoted to particular types of nucleic acids or proteins or properties of these macromolecules. Examples are databases of gene expression (see Chapters 8 and 9), databases of transfer RNA (tRNA) molecules, databases of tissue-specific protein expression (see Chapter 10), or databases of gene regulatory regions such as 3′-untranslated regions (see Chapter 16).

Pfam (► http://www.sanger.ac.uk/Software/Pfam/) and other related databases are described in Chapters 6 (multiple sequence alignment) and 10 (protein families).

Some bioinformatics databases do not contain nucleotide or protein sequence data as their main function. Instead, they contain information that may link to individual genes or proteins.

- Literature databases contain bibliographic references relevant to biological research and in some cases contain links to full-length articles. We will describe two of these databases, PubMed and the Sequence Retrieval System (SRS), in this chapter.
- Structure databases contain information on the structure of proteins and other macromolecules. These databases are described in Chapter 10 (on proteins) and Chapter 11 (on protein structure).

## GenBank: Database of Most Known Nucleotide and Protein Sequences

While the sequence information underlying DDBJ, EBI, and GenBank is equivalent, we begin our discussion with GenBank. GenBank is a database consisting of most

known public DNA and protein sequences (Benson et al., 2009). In addition to storing these sequences, GenBank contains bibliographic and biological annotation. Data from GenBank are available free of charge from the National Center for Biotechnology Information (NCBI) in the National Library of Medicine at the NIH (Wheeler et al., 2007).

## Amount of Sequence Data

GenBank currently contains about 100 billion nucleotides from 100 million sequences (release 168). The growth of GenBank in terms of both nucleotides of DNA and number of sequences from 1982 to 2008 is summarized in Fig. 2.1*a*. Over the period 1982 to the present, the number of bases in GenBank has doubled approximately every 18 months.

The WGS division consists of sequences generated by high throughput sequencing efforts. Since 2002, WGS sequences have been available at NCBI, but they are not considered part of the GenBank releases. As indicated in Fig. 2.1, the number of base pairs of DNA included among WGS sequences (136 billion base pairs in release 168, October 2008) is larger than the size of GenBank.

While the amount of sequence data in GenBank has risen rapidly, the arrival of next-generation sequencing technology, described in Chapter 13, is instantly leading

Between December 2007 and December 2008, over 15 billion base pairs (bp) of DNA were added to GenBank, an average of 42 million bp per day. In comparison, the first eukaryotic genome to be completed (*Saccharomyces cerevisiae;* Chapter 17) is about 13 million bp in size.

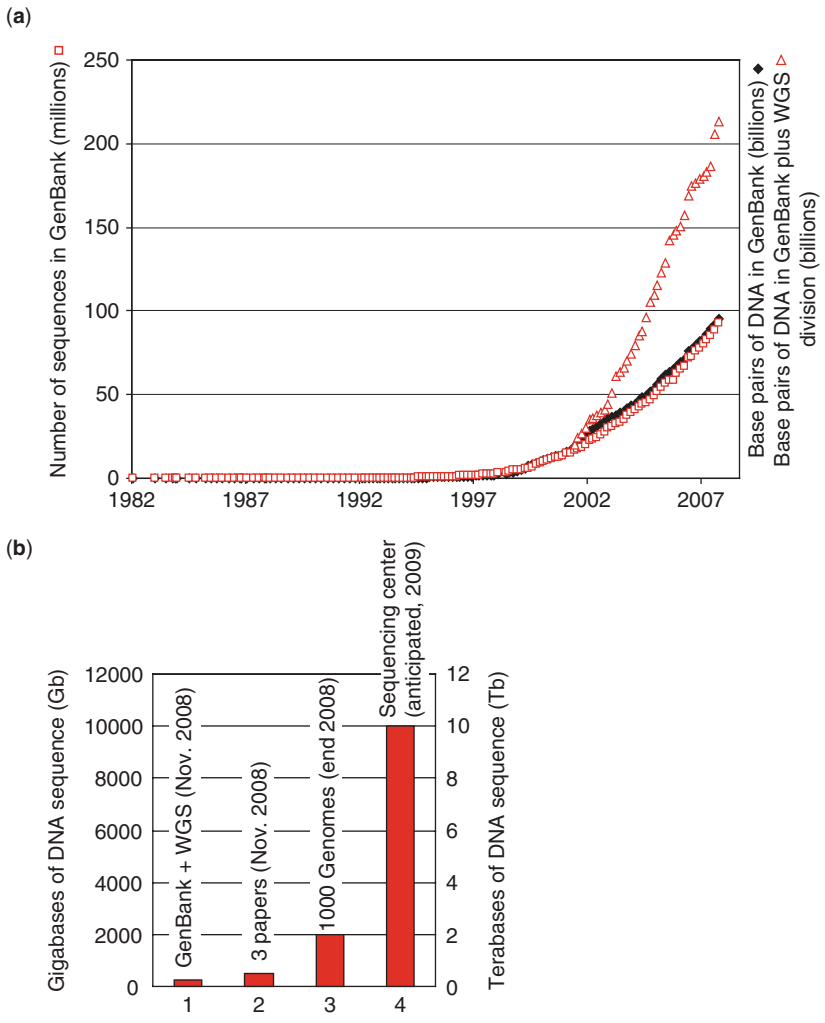



*FIGURE 2.1. (a) Growth of GenBank from release 3 (1982) to release 168 (October 2008). Data were plotted from the GenBank release notes at ► http://www.ncbi.nlm.nih.gov/Genbank/. Additional DNA sequences from the whole genome shotgun sequencing projects, begun in 2002, are shown. (b) The amount of sequenced DNA is vastly increasing. Bar 1 indicates the amount of DNA in GenBank plus WGS as shown in panel (a). Bar 2 indicates the amount of DNA sequence (492 gigabases) reported in three research articles published in a single issue of Nature (Bentley et al., 2008; Wang et al., 2008; Ley et al., 2008). Bar 3 indicates the amount of DNA sequence (2 terabases) expected to be generated by the end of 2008 as part of the 1000 Genomes Project (Chapter 13); 1 terabase was reported by the Wellcome Trust Sanger Institute in a six-month period in 2008. Bar 4 indicates the amount of sequence data (10 terabases) it is anticipated will be generated in 2009 alone by a typical major genome sequencing center.*

to a vast new influx of DNA sequence data (Fig. 2.1*b*). Next-generation sequencing involves the generation of massive amounts of sequence data, such as 1 billion bases (1 Gb) in a single experiment that is completed in a matter of days. In a single issue of the journal *Nature* in November 2008 Bentley et al. described the sequencing of an individual of Nigerian ancestry, Wang et al. reported the DNA sequence of an Asian individual, and Ley et al. analyzed the genome sequence of a tumor sample. Together, these three papers involved the generation and analysis of 492 gigabases (Gb) of DNA sequence. By the end of 2008 the 1000 Genomes Project generated several terabases of data. For major sequencing centers (such as those at the Wellcome Trust Sanger Institute, Beijing Genomics Institute Shenzhen, the Broad Institute of MIT and Harvard, Washington University School of Medicine's Genome Sequencing Center, and Baylor College of Medicine's Human Genome Sequencing Center) it is estimated that each will generate approximately 10 terabases in the year 2009. According to a Wellcome Trust Sanger Institute press release in 2008, that center now produces as much sequence data every 2 minutes as was generated in the first five years at GenBank. Thus the amount of DNA sequence generated by next-generation sequencing technologies has already dwarfed the amount of sequence in GenBank. Such data are available through the Trace Archive at NCBI and the Ensembl Trace Server at EBI, including the Short Read Archive that was initiated in 2007.

You can download all of the sequence data in GenBank at the website ► ftp://ftp.ncbi.nih.gov/genbank. For release 158.0 in February 2007, the total size of these files is about 250 gigabytes ($250 \times 10^9$ bytes). By comparison, all the words in the United States Library of Congress add up to 20 terabytes ($20 \times 10^{12}$ bytes; 20 trillion bytes). And the particle accelerator used by physicists at CERN near Geneva (► http://public.web.cern.ch/Public/) collects petabytes of data each year ($10^{15}$ bytes; 1 quadrillion bytes).

## Organisms in GenBank

Over 260,000 different species are represented in GenBank, with over 1000 new species added per month (Benson et al., 2009). The number of organisms represented in GenBank is shown in Table 2.1. We will define the bacteria, archaea, and eukaryotes in detail in Chapters 13 to 18. Briefly, eukaryotes have a nucleus and are often multicellular, whereas bacteria do not have a nucleus. Archaea are single-celled organisms, distinct from eukaryotes and bacteria, which constitute a third major branch of life. Viruses, which contain nucleic acids (DNA or RNA) but can only replicate in a host cell, exist at the borderline of the definition of living organisms.

We have seen so far that GenBank is very large and growing rapidly. From Table 2.1 we see that the organisms in GenBank consist mostly of eukaryotes. Of the microbes, there are currently over 25 times more bacteria than archaea represented in GenBank.

**TABLE 2-1** Taxa Represented in GenBank

| Ranks: | Higher Taxa | Genus | Species | Lower Taxa | Total |
|---|---|---|---|---|---|
| Archaea | 89 | 106 | 502 | 105 | 802 |
| Bacteria | 996 | 1,857 | 13,973 | 4,973 | 21,799 |
| Eukaryota | 15,205 | 45,066 | 167,764 | 13,200 | 241,235 |
| Fungi | 1,096 | 3,307 | 18,699 | 1,058 | 24,160 |
| Metazoa | 11,113 | 27,222 | 73,062 | 6,643 | 118,040 |
| Viridiplantae | 1,849 | 12,557 | 69,729 | 4,869 | 89,004 |
| Viruses | 445 | 294 | 5,054 | 33,909 | 39,702 |
| All taxa | 16,756 | 47,331 | 191,956 | 52,217 | 308,260 |

*Source*: From ► http://www.ncbi.nlm.nih.gov/Taxonomy/txstat.cgi (November 2008).

**TABLE 2-2** Twenty Most Sequenced Organisms in GenBank

| Entries | Bases | Species | Common Name |
|---|---|---|---|
| 11,550,460 | 13,148,670,755 | *Homo sapiens* | Human |
| 7,255,650 | 8,361,230,436 | *Mus musculus* | Mouse |
| 1,757,685 | 6,060,823,765 | *Rattus norvegicus* | Rat |
| 2,086,880 | 5,235,078,866 | *Bos taurus* | Cow |
| 3,181,318 | 4,600,009,751 | *Zea mays* | Corn |
| 2,489,204 | 3,551,438,061 | *Sus scrofa* | Pig |
| 1,591,342 | 2,978,804,803 | *Danio rerio* | Zebrafish |
| 1,205,529 | 1,533,859,717 | *Oryza sativa* | Rice |
| 228,091 | 1,352,737,662 | *Strongylocentrotus purpuratus* | Purple sea urchin |
| 1,673,038 | 1,142,531,302 | *Nicotiana tabacum* | Tobacco |
| 1,413,112 | 1,088,892,839 | *Xenopus (Silurana)* | Western clawed frog |
| 212,967 | 996,533,885 | *Pan troglodytes* | Chimpanzee |
| 780,860 | 913,586,921 | *Drosophila melanogaster* | Fruit fly |
| 2,211,104 | 912,500,625 | *Arabidopsis thaliana* | Thale cress |
| 650,374 | 905,797,007 | *Vitis vinifera* | Wine grape |
| 804,246 | 871,336,795 | *Gallus gallus* | Chicken |
| 77,069 | 803,847,320 | *Macaca mulatta* | Rhesus macaque |
| 1,215,319 | 748,031,972 | *Ciona intestinalis* | Sea squirt |
| 1,224,224 | 744,373,069 | *Canis lupus* | Dog |
| 1,725,913 | 680,988,452 | *Glycine max* | Soybean |

*Source*: From ► ftp://ftp.nebinith.gov/genbank/gbrel.txt (GenBank release 168.0, October 2008).

The number of entries and bases of DNA/RNA for the 20 most sequenced organisms in GenBank is provided in Table 2.2 (excluding chloroplast and mitochondrial sequences). This list includes some of the most common model organisms that are studied in biology. Notably, the scientific community is studying a series of mammals (e.g., human, mouse, cow), other vertebrates (chicken, frog), and plants (corn, rice, bread wheat, wine grape). Different species are useful for a variety of different studies. Bacteria, archaea, and viruses are absent from the list in Table 2.2 because they have relatively small genomes.

We will discuss how genomes of various organisms are selected for complete sequencing in Chapter 13.

To help organize the available information, each sequence name in a GenBank record is followed by its data file division and primary accession number. (Accession numbers are defined below.) The following codes are used to designate the data file divisions:

1. PRI: primate sequences
2. ROD: rodent sequences
3. MAM: other mammalian sequences
4. VRT: other vertebrate sequences
5. INV: invertebrate sequences
6. PLN: plant, fungal, and algal sequences
7. BCT: bacterial sequences
8. VRL: viral sequences
9. PHG: bacteriophage sequences

The International Human Genome Sequencing Consortium adopted the Bermuda Principles in 1996, calling for the rapid release of raw genomic sequence data. You can read about recent versions of these principles at ► http://www.genome.gov/10506376.

The terms STS, GSS, EST, and HTGS are defined bellow.

10. SYN: synthetic sequences
11. UNA: unannotated sequences
12. EST: EST sequences (expressed sequence tags)
13. PAT: patent sequences
14. STS: STS sequences (sequence-tagged sites)
15. GSS: GSS sequences (genome survey sequences)
16. HTG: HTGS sequences (high throughput genomic sequences)
17. HTC: HTC sequences (high throughput cDNA sequences)
18. ENV: environmental sampling sequences

Beta globin is sometimes called hemoglobin-beta. In general, a gene does not always have the same name as the corresponding protein. Indeed there is no such thing as a "hemoglobin gene" because globin genes encode globin proteins, and the combination of these globins with heme forms the various types of hemoglobin. Often, multiple investigators study the same gene or protein and assign different names. The human genome organization (HUGO) Gene Nomenclature Committee (HGNC) has the critical task of assigning official names to genes and proteins. See ► http://www.gene.ucl.ac.uk/nomenclature/.

## Types of Data in GenBank

There is an enormous number of molecular sequences in GenBank. We will next look at some of the basic kinds of data present in GenBank. Afterward, we will address strategies to extract the data you want from GenBank.

We start with an example. We want to find out the sequence of human beta globin. A fundamental distinction is that DNA, RNA-based, and protein sequences are stored in discrete databases. Furthermore, within each database, sequence data are represented in a variety of forms. For example, beta globin may be described at the DNA level (e.g. as a gene), at the RNA level (as a messenger RNA [mRNA] transcript), and at the protein level (see Fig. 2.2). Because RNA is relatively unstable, it is typically converted to complementary DNA (cDNA), and a variety

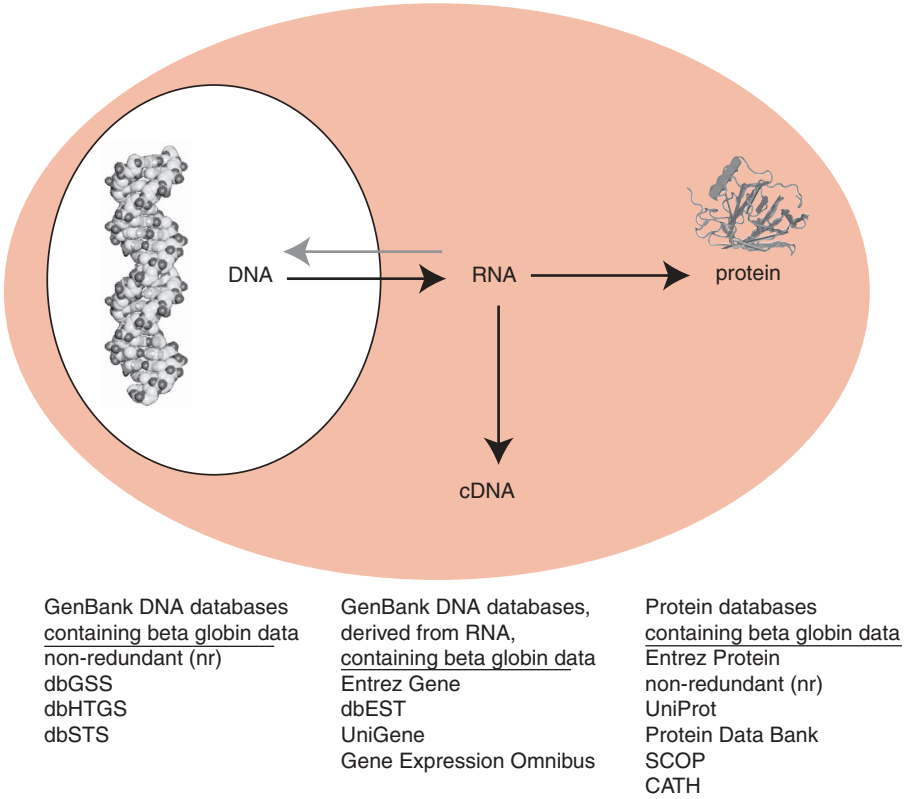



*FIGURE 2.2. Types of sequence data in GenBank and other databases using human beta globin as an example. Note that "globin" may refer to a gene or other DNA feature, an RNA transcript (or its corresponding complementary DNA), or a protein. There are specialized databases corresponding to each of these three levels. See text for abbreviations. There are many other databases (not listed) that are not part of GenBank and NCBI; note that SwissProt, PDB, and PIR are protein databases that are independent of GenBank. The raw nucleotide sequence data in GenBank, DDBJ, and EBI are equivalent.*

of databases contain cDNA sequences corresponding to RNA transcripts. Thus for our example of beta globin, the various forms of sequence data include the following.

## Genomic DNA Databases

- Beta globin is part of a chromosome. In the case of human RBP we will see that its gene is situated on chromosome 11 (Chapter 16, on the eukaryotic chromosome).
- Beta globin may be a part of a large fragment of DNA such as a cosmid, bacterial artificial chromosome (BAC), or yeast artificial chromosome (YAC) that may contain several genes. A BAC is a large segment of DNA (typically about 200,000 base pairs [bp], or 200 kilobases [kb]) that is cloned into bacteria. Similarly, YACs are used to clone large amounts of DNA into yeast. BACs and YACs are useful vectors with which to sequence large portions of genomes.
- Beta globin is present in databases as a gene. The gene is the functional unit of heredity (further defined in Chapter 16), and it is a DNA sequence that typically consists of regulatory regions, protein-coding exons, and introns. Often, human genes are 10 to 100 kb in size.
- Beta globin is present as a sequence-tagged site (STS)—that is, as a small fragment of DNA (typically 500 bp long) that is used to link genetic and physical maps and which is part of a database of sequence-tagged sites (dbSTS).

Human chromosome 11, which is a mid-sized chromosome, contains about 1800 genes and is about 134,000 kilobases (kb) in length.

## cDNA Databases Corresponding to Expressed Genes

Beta globin is represented in databases as an expressed sequence tag (EST), that is, a cDNA sequence derived from a particular cDNA library. If one obtains a tissue such as liver, purifies RNA, then converts the RNA to the more stable form of cDNA, some of the cDNA clones contained in that cDNA are likely to encode beta globin.

In GenBank, the convention is to use the four DNA nucleotides when referring to DNA derived from RNA.

## Expressed Sequence Tags (ESTs)

The database of expressed sequence tags (dbEST) is a division of GenBank that contains sequence data and other information on "single-pass" cDNA sequences from a number of organisms (Boguski et al., 1993). An EST is a partial DNA sequence of a cDNA clone. All cDNA clones, and thus all ESTs, are derived from some specific RNA source such as human brain or rat liver. The RNA is converted into a more stable form, cDNA, which may then be packaged into a cDNA library (refer to Fig. 2.2). ESTs are typically randomly selected cDNA clones that are sequenced on one strand (and thus may have a relatively high sequencing error rate). ESTs are often 300 to 800 bp in length. The earliest efforts to sequence ESTs resulted in the identification of many hundreds of genes that were novel at the time (Adams et al., 1991).

In November, 2008 GenBank had over 58,000,000 ESTs. We discuss ESTs further in Chapter 8.

Currently, GenBank divides ESTs into three major categories: human, mouse, and other. Table 2.3 shows the 10 organisms from which the greatest number of ESTs has been sequenced. Assuming that there are 22,000 human genes (see Chapter 19) and given that there are about 8.1 million human ESTs, there is currently an average of over 300 ESTs corresponding to each human gene.

**TABLE 2-3** Top Ten Organisms for Which ESTs Have Been Sequenced

| Organisms | Common Name | Number of ESTs |
|---|---|---|
| *Homo sapiens* | Human | 8,138,094 |
| *Mus musculus + domesticus* | Mouse | 4,850,602 |
| *Zea mays* | Maize | 2,002,585 |
| *Arabidopsis thaliana* | Thale cress | 1,526,133 |
| *Bos taurus* | Cattle | 1,517,139 |
| *Sus scrofa* | Pig | 1,476,546 |
| *Danio rerio* | Zebrafish | 1,379,829 |
| *Glycine max* | Soybean | 1,351,356 |
| *Xenopus (Silurana) tropicalis* | Western clawed frog | 1,271,375 |
| *Oryza sativa* | Rice | 1,220,908 |

Many thousand of cDNA libraries have been generated from a variety of organism, and the total number of public entries is currently over 58 million.
*Source*: ► http://www.nebi.nlm.nin.gov/dbEST/dbEST_summary.html (dbEST release 022307, November 2008).

## ESTs and UniGene

To find the entry for beta globin, go to ► http://www.ncbi.nlm.nih.gov, select All Databases then click UniGene, select human, then enter beta globin or HBB. The UniGene accession number is Hs.523443; note that Hs refers to *Homo sapiens*. To see the DNA sequence of a typical EST, click on an EST accession number from the UniGene page (e.g., AA970968.1), then follow the link to the GenBank entry in Entrez Nucleotide.

We are using beta globin as a specific example. If you want to type "globin" as a query, you will simply get more results from any database—in UniGene, you will find over 100 entries corresponding to a variety of globin genes in various species.

The UniGene project has become extremely important in the effort to identify protein-coding genes in newly sequenced genomes. We discuss this in Chapters 13 and 16.

The goal of the UniGene (unique gene) project is to create gene-oriented clusters by automatically partitioning ESTs into nonredundant sets. Ultimately there should be one UniGene cluster assigned to each gene of an organism. There may be as few as one EST in a cluster, reflecting a gene that is rarely expressed, to tens of thousands of ESTs, associated with a highly expressed gene. We discuss UniGene clusters further in Chapter 8 (on gene expression). There are over 100 organisms currently represented in UniGene, 71 of which are listed in Table 2.4.

For human beta globin, there is only a single UniGene entry. This entry currently has 2400 human ESTs that match the beta globin gene. This large number of ESTs reflects how abundantly the beta globin gene has been expressed in cDNA libraries that have been sequenced. A UniGene cluster is a database entry for a gene containing a group of corresponding ESTs (Fig. 2.3).

There are now thought to be approximately 22,000 human genes (see Chapter 19). One might expect an equal number of UniGene clusters. However, in practice, there are more UniGene clusters than there are genes—currently, there are about 120,000 human UniGene clusters. This discrepancy could occur for three reasons. (1) Clusters of ESTs could correspond to distinct regions of one gene. In that case there would be two (or more) UniGene entries corresponding to a single gene (see Fig. 2.3). Two UniGene clusters may properly cluster into one, and the number of UniGene clusters may collapse over time. (2) In the past several years it has become appreciated that much of the genome is transcribed at low levels (see Chapter 8). Currently, 40,000 human UniGene clusters consist of a single EST, and over 76,000 UniGene clusters consist of just one to four ESTs. These could reflect authentic genes that have not yet been appreciated by other means of gene identification. Alternatively they may represent rare transcription events of unknown biological relevance. (3) Some DNA may be transcribed during the creation of a cDNA library without corresponding to an authentic transcript. Thus it is a cloning artifact. We discuss the criteria for defining a eukaryotic gene in Chapter 16. Alternative splicing (Chapter 8) may introduce apparently new clusters of genes because the spliced exon is not homologous to the rest of the sequence.

**TABLE 2-4** Seventy-One Organisms Represented in UniGene

| Group | No. | Species |
|---|---|---|
| Chordata: Mammalia | 12 | *Bos taurus* (cattle), *Canis familiaris* (dog), *Equus caballus* (horse), *Homo sapiens* (human), *Macaca fascicularis* (crab-eating macaque), *Macaca mulatta* (rhesus monkey), *Mus musculus* (mouse), *Oryctolagus cuniculus* (rabbit), *Ovis aries* (sheep), *Rattus norvegicus* (Norway rat), *Sus scrofa* (pig), *Trichosurus vulpecula* (silver-gray brushtail possum) |
| Chordata: Actinopterygii | 8 | *Danio rerio* (zebrafish), *Fundulus heteroclitus* (killifish), *Gasterosteus aculeatus* (three spined stickleback), *Oncorhynchus mykiss* (rainbow trout), *Oryzias latipes* (Japanese medaka), *Pimephales promelas* (fathead minnow), *Salmo salar* (Atlantic salmon), *Takifugu rubripes* (pufferfish) |
| Chordata: Amphibia | 2 | *Xenopus laevis* (African clawed frog), *Xenopus tropicalis* (western clawed frog) |
| Chordata: Ascidiacea | 3 | *Ciona intestinalis*, *Ciona savignyi*, *Molgula tectiformis* |
| Chordata: Aves | 2 | *Gallus gallus* (chicken), *Taeniopygia guttata* (zebra finch) |
| Chordata: Cephalochordata | 1 | *Branchiostoma floridae* (Florida lancelet) |
| Chordata: Hyperoartia | 1 | *Petromyzon marinus* (sea lamprey) |
| Echinodermata: Echinoidea | 1 | *Strongylocentrotus purpuratus* (purple sea urchin) |
| Arthopoda: Insecta | 6 | *Aedes aegypti* (yellow fever mosquito), *Anopheles gambiae* (African malaria mosquito), *Apis mellifera* (honey bee), *Bombyx mori* (domestic silkworm), *Drosophila melanogaster* (fruit fly), *Tribolium castaneum* (red flour beetle) |
| Nematoda: Chromadorea | 1 | *Caenorhabditis elegans* (nematode) |
| Platyhelminthes: Trematoda | 2 | *Schistosoma japonicum*, *Schistosoma mansoni* |
| Cnidaria: Hydrozoa | 1 | *Hydra magnipapillata* |
| Streptophyta: Bryopsida | 1 | *Physcomitrella patens* |
| Streptophyta: Coniferopsida | 3 | *Picea glauca* (white spruce), *Picea sitchensis* (Sitka spruce), *Pinus taeda* (loblolly pine) |
| Streptophyta: Eudicotyledons | 18 | *Aquilegia formosa* × *Aquilegia pubescens*, *Arabidopsis thaliana* (thale cress), *Brassica napus* (rape), *Citrus sinensis* (Valencia orange), *Glycine max* (soybean), *Gossypium hirsutum* (upland cotton), *Gossypium raimondii*, *Helianthus annuus* (sunflower), *Lactuca sativa* (garden lettuce), *Lotus japonicus*, *Malus* × *domestica* (apple), *Medicago truncatula* (barrel medic), *Nicotiana tabacum* (tobacco), *Populus tremula* × *Populus tremuloides*, *Populus trichocarpa* (western balsam poplar), *Solanum lycopersicum* (tomato), *Solanum tuberosum* (potato), *Vitis vinifera* (wine grape) |
| Streptophyta: Liliopsida | 6 | *Hordeum vulgare* (barley), *Oryza sativa* (rice), *Saccharum officinarum* (sugarcane), *Sorghum bicolor* (sorghum), *Triticum aestivum* (wheat), *Zea mays* (maize) |
| Chlorophyta: Chlorophyceae | 1 | *Chlamydomonas reinhardtii* |
| Dictyosteliida: Dictyostelium | 1 | *Dictyostelium discoideum* (slime mold) |
| Apicomplexa: Coccidia | 1 | *Toxoplasma gondii* |

*Source*: UniGene ► http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene (November 2008).

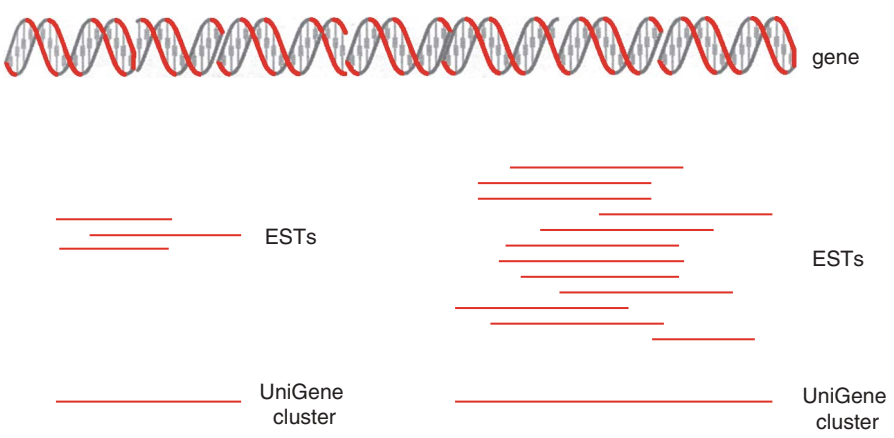



*FIGURE 2.3. Schematic description of UniGene clusters. Expressed sequence tags (ESTs) are mapped to a particular gene and to each other. The number of ESTs that constitute a UniGene cluster ranges from 1 to tens of thousands; on average there are 300 human ESTs per cluster. Sometimes, as shown in the diagram, separate UniGene clusters correspond to distinct regions of a gene. Eventually, as genome sequencing increases our ability to define and annotate full-length genes, these two UniGene clusters would be collapsed into one single cluster. Ultimately, the number of UniGene clusters should equal the number of genes in the genome.*

## Sequence-Tagged Sites (STSs)

As of November 2008 there are 1.3 million STSs, derived from 300 organisms.

The dbSTS is an NCBI site containing STSs, which are short genomic landmark sequences for which both DNA sequence data and mapping data are available (Olson et al., 1989). STSs have been obtained from several hundred organisms, including primates and rodents (Table 2.5). A typical STS is approximately the size of an EST. Because they are sometimes polymorphic, containing short sequence repeats (Chapter 16), STSs can be useful for mapping studies.

## Genome Survey Sequences (GSSs)

There are currently 24 million GSS entries from over 800 organisms (November 2008). The top four organisms (Table 2.6) account for about a third of all entries. This database is accessed via ► http://www.ncbi.nlm.nih.gov/projects/dbGSS/.

The GSS division of GenBank is similar to the EST division, except that its sequences are genomic in origin, rather than cDNA (mRNA). The GSS division contains the following types of data (see Chapters 13 and 16):

- Random "single-pass read" genome survey sequences
- Cosmid/BAC/YAC end sequences
- Exon-trapped genomic sequences
- The *Alu* polymerase chain reaction (PCR) sequences

**TABLE 2-5 Organisms from Which STSs Have Been Obtained**

| Organism | Approximate Number of STSs |
|---|---|
| *Homo sapiens* | 324,000 |
| *Pan troglodytes* | 161,000 |
| *Macaca mulatta* | 72,000 |
| *Mus musculus* | 56,000 |
| *Rattus norvegicus* | 50,000 |

These are the organisms with the most UniSTS entries.
*Source*: ► http://www.ncbi.nlm.nih.gov/genome/sts/unists_stats.html (November 2008).

**TABLE 2-6** Selected Organisms from Which GSSs Have Been Obtained. For a discussion of Metagenomes see Chapter 13

| Organism | Approximate Number of Sequences |
|---|---|
| Marine metagenome | 2,643,000 |
| *Zea mays* + subsp. *mays* (maize) | 2,091,000 |
| *Mus musculus* + *domesticus* (mouse) | 1,864,000 |
| *Nicotiana tabacum* (tobacco) | 1,421,000 |
| *Homo sapiens* (human) | 1,214,000 |
| *Canis lupus familiaris* (dog) | 854,000 |

*Source*: ► http://www.ncbi.nlm.nih.gov/dbGSS/dbGSS_summary.html (November 2008).

All searches of the Entrez Nucleotide database provide results that are divided into three sections: GSS, ESTs, and "CoreNucleotide" (that is, the remaining nucleotide sequences). Recent holdings of the GSS database are listed in Table 2.6.

## High Throughput Genomic Sequence (HTGS)

The HTGS division was created to make "unfinished" genomic sequence data rapidly available to the scientific community. It was done in a coordinated effort between the three international nucleotide sequence databases: DDBJ, EMBL, and GenBank. The HTGS division contains unfinished DNA sequences generated by the high throughput sequencing centers.

The HTGS home page is ► http://www.ncbi.nlm.nih.gov/HTGS/ and its sequences can be searched via BLAST (see Chapters 4 and 5).

## Protein Databases

The name beta globin may refer to the DNA, the RNA, or the protein. As a protein, beta globin is present in databases such as the nonredundant (nr) database of GenBank (Benson et al., 2009), the SwissProt database (Boeckmann et al., 2003), UniProt (UniProt Consortium 2007), and the Protein Data Bank (Kouranov et al., 2006).

We have described some of the basic kinds of sequence data in GenBank. We will next turn our attention to Entrez and the other programs in NCBI and elsewhere, which allow you to access GenBank, EMBL, and DDBJ data and related literature information. In particular, we will introduce the NCBI website, one of the main web-based resources in the field of bioinformatics.

# NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION

## Introduction to NCBI: Home Page

The NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information (Wheeler et al., 2007). The NCBI home page is shown in Fig. 2.4. Across the top bar of the website, there are seven categories: PubMed, Entrez, BLAST, OMIM, Books, Taxonomy, and Structure.

Extremely useful tutorials are available for Entrez, PubMed, and other NCBI resources at ► http://www.ncbi.nlm.nih.gov/Education/. You can also access this from the education link on the NCBI home page (► http://www.ncbi.nlm.nih.gov).

## PubMed

PubMed is the search service from the National Library of Medicine (NLM) that provides access to over 18 million citations in MEDLINE (Medical Literature,

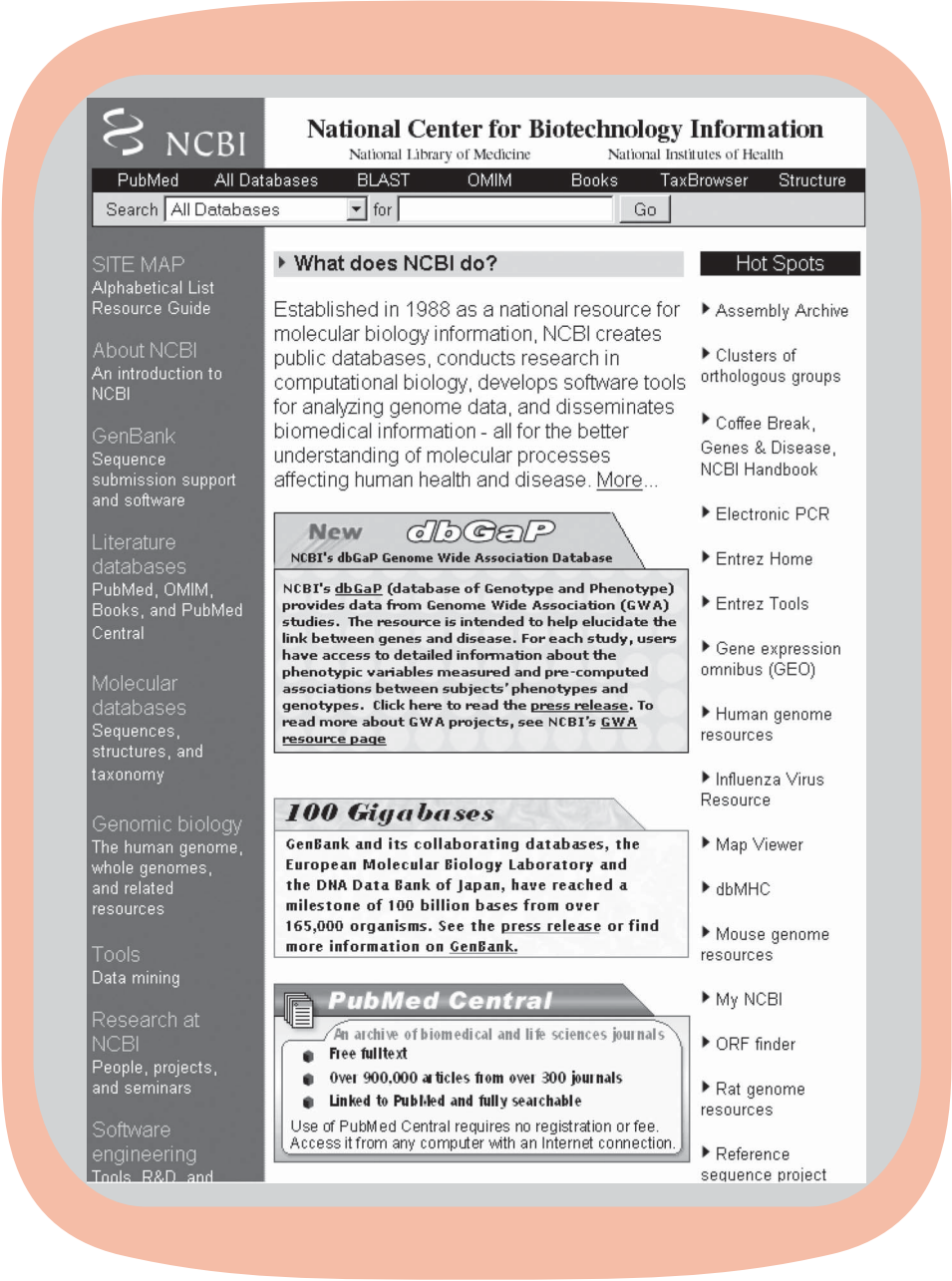



*FIGURE 2.4. The main page of the National Center for Biotechnology Information (NCBI) website (► http://www.ncbi.nlm.nih.gov). Across the top bar, sections include PubMed, Entrez and Books (described in this chapter), BLAST (Chapters 3–5), Taxonomy (Chapters 13–19), Structure (Chapter 11), and Online Mendelian Inheritance in Man (OMIM, Chapter 20). Note that the left sidebar includes tutorials within the Education section.*

Analysis, and Retrieval System Online) and other related databases, with links to participating online journals.

## Entrez

Entrez integrates the scientific literature, DNA and protein sequence databases, three-dimensional protein structure data, population study data sets, and assemblies of complete genomes into a tightly coupled system. PubMed is the literature component of Entrez.

### BLAST

BLAST (Basic Local Alignment Search Tool) is NCBI's sequence similarity search tool designed to support analysis of nucleotide and protein databases (Altschul et al., 1990, 1997). BLAST is a set of similarity search programs designed to explore all of the available sequence databases regardless of whether the query is protein or DNA. We explore BLAST in Chapters 3 to 5.

### OMIM

Online Mendelian Inheritance in Man (OMIM) is a catalog of human genes and genetic disorders. It was created by Victor McKusick and his colleagues and developed for the World Wide Web by NCBI (Hamosh et al., 2005). The database contains detailed reference information. It also contains links to PubMed articles and sequence information. We describe OMIM in Chapter 20 (on human disease).

### Books

NCBI offers several dozen books online. These books are searchable, and are linked to PubMed.

### Taxonomy

The NCBI taxonomy website includes a taxonomy browser for the major divisions of living organisms (archaea, bacteria, eukaryota, and viruses). The site features taxonomy information such as genetic codes and taxonomy resources and additional information such as molecular data on extinct organisms and recent changes to classification schemes. We will visit this site in Chapters 7 (on evolution) and 13 to 18 (on genomes and the tree of life).

### Structure

The NCBI structure site maintains the Molecular Modelling Database (MMDB), a database of macromolecular three-dimensional structures, as well as tools for their visualization and comparative analysis. MMDB contains experimentally determined biopolymer structures obtained from the Protein Data Bank (PDB). Structure resources at NCBI include PDBeast (a taxonomy site within MMDB), Cn3D (a three-dimensional structure viewer), and a vector alignment search tool (VAST) which allows comparison of structures. (See Chapter 11, on protein structure.)

The Protein Data Bank (► http://www.rcsb.org/pdb/) is the single worldwide repository for the processing and distribution of biological macromolecular structure data. We explore the PDB in Chapter 11.

## THE EUROPEAN BIOINFORMATICS INSTITUTE (EBI)

The EBI website is comparable to NCBI in its scope and mission, and it represents a complementary, independent resource. EBI features six core molecular databases (Brooksbank et al., 2003), as follows. (1) EMBL-Bank is the repository of DNA and RNA sequences that is complementary to GenBank and DDBJ (Kulikova et al., 2007). (2) SWISS-PROT and (3) TrEMBL are two protein databases that are described further below. (4) MSD is a protein structure database (see Chapter 11). (5) Ensembl is one of the three main genome browsers (described below). (6) ArrayExpress is one of the two main worldwide repositories for gene expression

You can access EBI at ►EBI at http://www.ebi.ac.uk/.

data, along with the Gene Expression Omnibus at NCBI; both are described in Chapter 8.

Throughout this book we will focus on both the NCBI and EBI websites. In many cases those sites begin with similar raw data and then provide distinct ways of organizing, analyzing, and displaying data across a broad range of bioinformatics applications. When you work on a problem, such as studying the structure or function of a particular gene, it is often helpful to explore the wealth of resources on both these sites. For example, each offers expert functional annotation of particular sequences and expert curation of the database. The NCBI and EBI websites increasingly offer an integration of their database resources so that one can link to information between the two sites with reasonable effort.

## Access to Information: Accession Numbers to Label and Identify Sequences

When you have a problem you are studying that involves any gene or protein, it is likely that you will need to find information about some database entries. You may begin your research problem with information obtained from the literature or you may have the name of a specific sequence of interest. Perhaps you have raw amino acid and/or nucleotide sequence data; we will explore how to analyze these (e.g. Chapters 3 to 5). The problem we will address now is how to extract information about your gene or protein of interest from databases.

An essential feature of DNA and protein sequence records is that they are tagged with accession numbers. An accession number is a string of about 4 to 12 numbers and/or alphabetic characters that are associated with a molecular sequence record. An accession number may also label other entries, such as protein structures or the results of a gene expression experiment (Chapters 8 and 9). Accession numbers from molecules in different databases have characteristic formats (Box 2.1). These formats vary because each database employs its own system. As you explore databases from which you extract DNA and protein data, try to become familiar with the different formats for accession numbers. Some of the various databases (Fig. 2.2) employ accession numbers that tell you whether the entry contains nucleotide or protein data.

DNA is usually sequenced on both strands. However, ESTs are often sequenced on one strand only, and thus they have a high error rate. We will discuss sequencing error rates in Chapter 13.

For a typical molecule such as beta globin there are thousands of accession numbers (Fig. 2.5). Many of these correspond to ESTs and other fragments of DNA that match beta globin. How can you assess the quality of sequence or protein data? Some sequences are full-length, while others are partial. Some reflect naturally occurring variants such as single nucleotide polymorphisms (SNPs; Chapter 16) or alternatively spliced transcripts (Chapter 8). Many of the sequence entries contain errors, particularly in the ends of EST reads. When we compare beta globin sequences derived from mRNA and from genomic DNA, we may expect them to match perfectly (or nearly so), but as we will see, discrepancies routinely occur.

In addition to accession numbers, NCBI also assigns unique sequence identification numbers that apply to the individual sequences within a record. GI numbers are assigned consecutively to each sequence that is processed. For example, the human beta globin DNA sequence associated with the accession number NM_000518.4 has a gene identifier GI:28302128. The suffix .4 on the accession number refers to a version number; NM_000518.3 has a different gene identifier, GI: 13788565.

## BOX 2-1
## Types of Accession Numbers

| Type of Record | Sample Accession Format |
|---|---|
| GenBank/EMBL/DDBJ nucleotide sequence records | One letter followed by five digits, e.g., X02775<br>Two letters followed by six digits, e.g., AF025334 |
| GenPept sequence records (which contain the amino acid translations from GenBank/EMBL/DDBJ records that have a coding region feature annotated on them) | Three letters and five digits, e.g., AAA12345 |
| Protein sequence records from SwissProt and PIR | Usually one letter and five digits, e.g., P12345. SwissProt numbers may also be a mixture of numbers and letters. |
| Protein sequence records from the Protein Research Foundation | A series of digits (often six or seven) followed by a letter, e.g., 1901178A |
| RefSeq nucleotide sequence records | Two letters, an underscore bar, and six or more digits, e.g., mRNA records (NM_*): NM_006744; genomic DNA contigs (NT_*): NT_008769 |
| RefSeq protein sequence records | Two letters (NP), an underscore bar, and six or more digits, e.g., NP_006735 |
| Protein structure records | PDB accessions generally contain one digit followed by three letters, e.g., 1TUP. They may contain other mixtures of numbers and letters (or numbers only). MMDB ID numbers generally contain four digits, e.g., 3973. |

## The Reference Sequence (RefSeq) Project

One of the most important recent developments in the management of molecular sequences is RefSeq. The goal of RefSeq is to provide the best representative sequence for each normal (i.e., nonmutated) transcript produced by a gene and for each normal protein product (Pruitt et al., 2009; Maglott et al., 2000). There may be hundreds of GenBank accession numbers corresponding to a gene, since GenBank is an archival database that is often highly redundant. However, there will be only one RefSeq entry corresponding to a given gene or gene product, or several RefSeq entries if there are splice variants or distinct loci.

To see and compare the three myoglobin RefSeq entries at the DNA and the protein levels, visit ► http://www.bioinfbook.org/chapter2 and select webdocument 2.1.

Consider human myoglobin as an example. There are three RefSeq entries (NM_005368, NM_203377, and NM_203378), each corresponding to a distinct splice variant. Each splice variant involves the transcription of different exons from a single gene locus. In this example, all three transcripts happen to encode an identical protein having the same amino acid sequence. Because the source of the transcript varies distinctly, each identical protein sequence is assigned its own protein accession number (NP_005359, NP_976311, and NP_976312, respectively).

Allelic variants, such as single base mutations in a gene, are not assigned different RefSeq accession numbers. However, OMIM and dbSNP (Chapters 16 and 20) do catalog allelic variants.

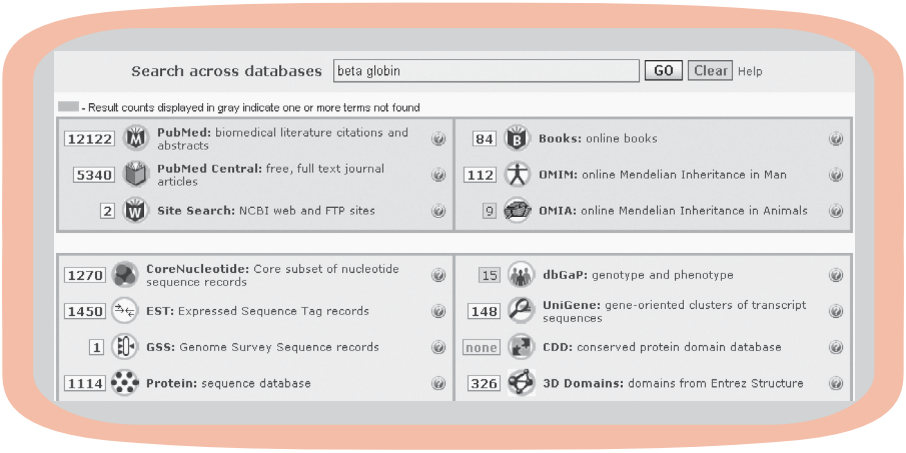



*FIGURE 2.5. There are thousands of accession numbers corresponding to many genes and proteins. A search with the query "beta globin" from the main page of NCBI shows the results across the databases of the Entrez search engine. There are over 1000 each of core nucleotide sequences, expressed sequence tags (ESTs), and proteins. The RefSeq project is particularly important in trying to provide the best representative sequence of each normal (nonmutated) transcript produced by a gene and of each distinct, normal protein sequence.*

A GenBank or RefSeq accession number refers to the most recent version of a given sequence. For example NM_000558.3 is currently a RefSeq identifier for human alpha globin. The suffix".3" is the version number. By default, if you do not specify a version number then the most recent version is provided. Try doing an Entrez nucleotide search for NM_000558.1 and you can learn about the revision history of that accession number. In Chapter 3 we will learn how to compare two sequences; you can blast NM_000558.1 against NM_000558.3 to see the differences, or view the results in web document 2.2 at ► http://www.bioinfbook.org/chapter2.

RefSeq entries are curated by the staff at NCBI, and are nearly nonredundant. However, there can be two proteins encoded by distinct genes sharing 100% amino acid identity. Each is assigned its own unique RefSeq identifier. For example, the alpha-1 globin and alpha-2 globin genes in human are physically separate genes that encode proteins with identical sequences. The encoded alpha-1 globin and alpha-2 globin proteins are assigned the RefSeq identifiers NP_000549 and NP_000508.

Refseq entries have different status levels (predicted, provisional, and reviewed), but in each case the RefSeq entry is intended to unify the sequence records. You can recognize a RefSeq accession by its format, such as NP_000509 (P stands for beta globin protein) or NM_006744 (for beta globin mRNA). A variety of RefSeq identifiers are shown in Table 2.7, and examples of beta globin identifiers are given in Table 2.8.

**TABLE 2-7 Formats of Accession Numbers for RefSeq Entries**

| Molecule | Accession Format | Genome |
|---|---|---|
| Complete genome | NC_123456 | Complete genomic molecules, including genomes, chromosomes, organelles, and plasmids |
| Genomic DNA | NW_123456<br>NW_123456789 | Intermediate genomic assemblies |
| Genomic DNA | NZ_ABCD12345678 | Collection of whole genome shotgun sequence data |
| Genomic DNA | NT_123456 | Intermediate genomic assemblies (BAC and/or WGS sequence data) |
| mRNA | NM_123456 or<br>NM_123456789 | Transcript products; mature mRNA protein-coding transcripts |
| Protein | NP_123456 or<br>NM_123456789 | Protein products (primarily full-length) |
| RNA | NR_123456 | Noncoding transcripts (e.g. structural RNAs, transcribed pseudogenes) |

There are currently 21 different RefSeq accession formats. The methods include expert manual curation, automated curation, or a combination. Abbreviations: BAC, bacterial artificial chromosome; WGS, whole genome shotgun (see Chapter 13).

*Source*: Adapted from ► http://www.ncbi.nlm.nih.gov/RefSeq/key.html#accessions (March 2007).

**TABLE 2-8** RefSeq Accession Numbers Corresponding to Human Beta Globin

| Category | Accession | Size | Description |
|---|---|---|---|
| DNA | NC_000011 | 134,452,384 bp | Genomic contig |
| DNA | NM_000518.4 | 626 bp | DNA corresponding to mRNA |
| DNA | NG_000007.3 | 81,706 bp | Genomic reference |
| DNA | NW_925006.1 | 1,606 bp | Alternate assembly |
| Protein | NP_000509.1 | 147 amino acids | Protein |

## The Consensus Coding Sequence (CCDS) Project

The Consensus Coding Sequence (CCDS) project was established to identify a core set of protein coding sequences that provide a basis for a standard set of gene annotations. The CCDS project is a collaboration between four groups (EBI, NCBI, the Wellcome Trust Sanger Institute, and the University of California, Santa Cruz [UCSC]). Currently, the CCDS project has been applied to the human and mouse genomes, and thus its scope is considerably more limited than RefSeq.

You can learn about the CCDS project at ► http://www.ncbi.nlm.nih.gov/projects/CCDS/.

## Access to Information via Entrez Gene at NCBI

How can one navigate through the bewildering number of protein and DNA sequences in the various databases? An emerging feature is that the various databases are increasingly interconnected, providing a variety of convenient links to each other and to algorithms that are useful for DNA, RNA, and protein analysis. Entrez Gene (formerly LocusLink) is particularly useful as a major portal. It is a curated database containing descriptive information about genetic loci (Maglott et al., 2007). You can obtain information on official nomenclature, aliases, sequence accessions, phenotypes, EC numbers, OMIM numbers, UniGene clusters, HomoloGene (a database that reports eukaryotic orthologs), map locations, and related websites.

Entrez Gene is accessed from the main NCBI web page (by clicking All Databases). Currently (November 2008), Entrez Gene encompasses about 5,700 taxa and 4.6 million genes. We will explore many of the resources within Entrez Gene in later chapters.

To illustrate the use of Entrez Gene we will search for human myoglobin. The result of entering an Entrez Gene search is shown in Fig. 2.6. Note that in performing this search, it can be convenient to restrict the search to a particular organism of interest. (This can be done using the "limits" tab on the Entrez Gene page.) The "Links" button (Fig. 2.6, top right) provides access to various other database entries on myoglobin. Clicking on the main link to the human myoglobin entry results in the following information (Fig. 2.7):

- At the top right, there is a table of contents for the Entrez Gene myoglobin entry. Below it are further links to myoglobin entries in NCBI databases (e.g. protein and nucleotide databases and PubMed), as well as external databases (e.g. Ensembl and UCSC; see below and Chapter 16).
- Entrez Gene provides the official symbol and name for human myoglobin, MB.
- A schematic overview of the gene structure is provided, hyperlinked to the Map Viewer (see below).
- There is a brief description of the function of MB, defining it as a carrier protein of the globin family.

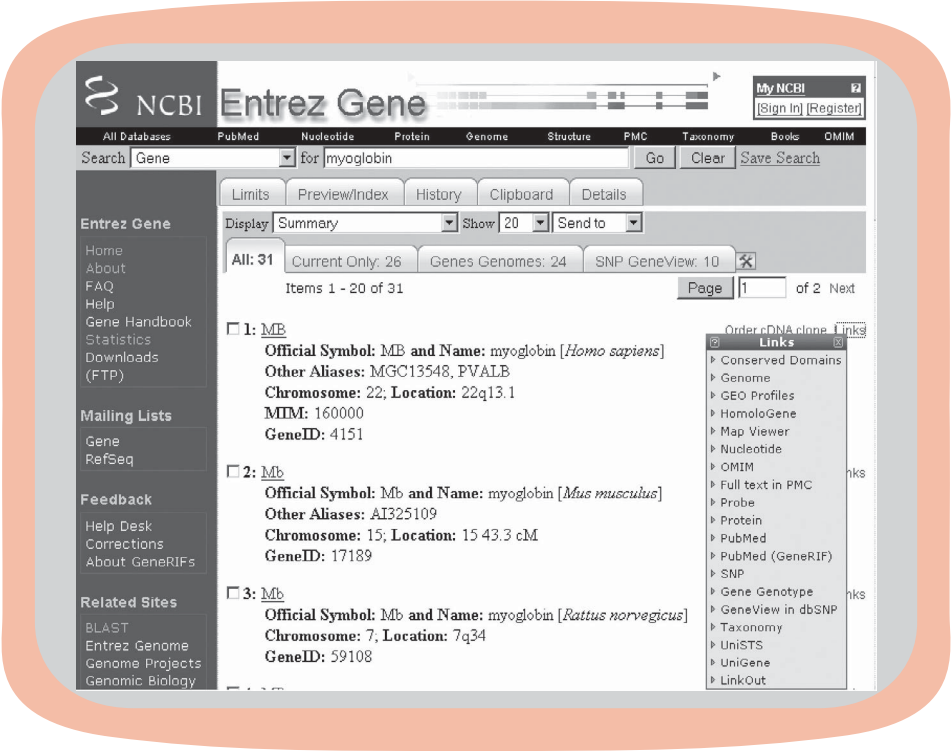



*FIGURE 2.6. Result of a search for "myoglobin" in Entrez Gene. Information is provided for a variety of organisms, including* Homo sapiens, Mus musculus, *and* Rattus norvegicus. *The links button (top right) provides access to information on myoglobin from a variety of other databases.*

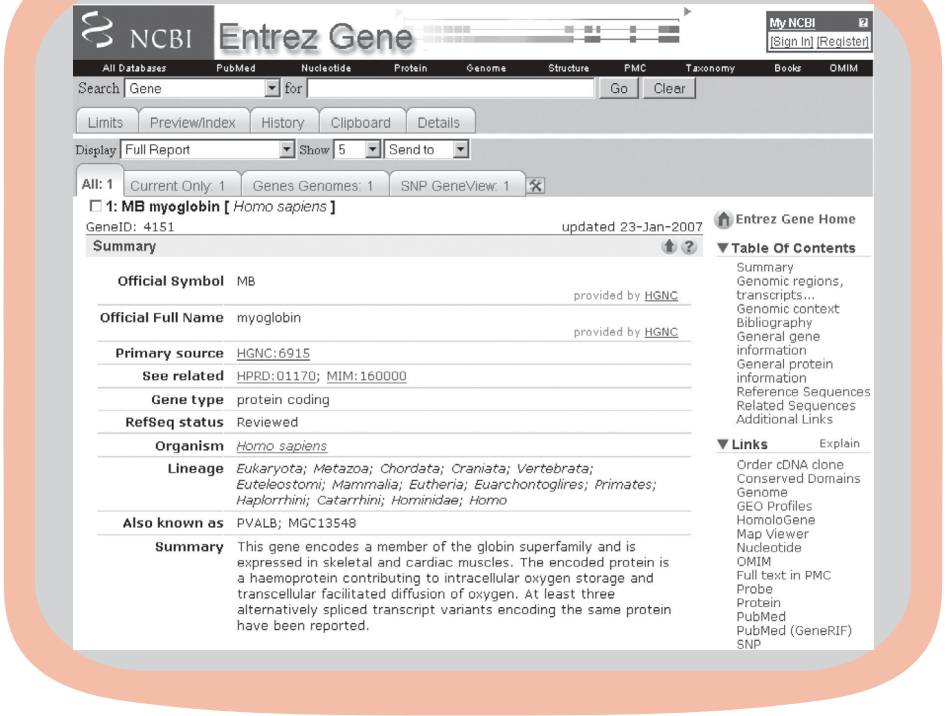



*FIGURE 2.7. Portion of the Entrez Gene entry for human myoglobin. Information is provided on the gene structure, chromosomal location, as well as a summary of the protein's function. RefSeq accession numbers are also provided (not shown); you can access them by clicking "Reference sequences" in the table of contents (top right). The menu (right sidebar) provides extensive links to additional databases, including PubMed, OMIM, UniGene, a variation database (dbSNP), HomoloGene (with information on homologs), a gene ontology database, and Ensembl viewers at EBI. We will describe these resources in later chapters.*

- The Reference Sequence (RefSeq) accession numbers are provided: NM_005368 for the DNA sequence encoding the longest myoglobin transcript and NP_005359 for the protein entry. GenBank accession numbers corresponding to myoglobin (both nucleotide and protein) are also provided.

Figure 2.8 shows the standard, default form of a typical Entrez Protein record (for myoglobin). It is simple to obtain a variety of formats by changing the Entrez display options. By using the Display pulldown menu (Fig. 2.8a) one can obtain


*FIGURE 2.8. Display of an Entrez Protein record for human myoglobin. This is a typical entry for any protein. (a) Top portion of the record. Key information includes the length of the protein (154 amino acids), the division (PRI, or primate), the accession number (NP_005359), the organism (H. sapiens), literature references, comments on the function of globins, and many links to other databases (right side). At the top of the page, the display option allows you to obtain this record in a variety of formats, such as FASTA (Figure 2.9). (b) Bottom portion of the record. This includes features such as the coding sequence (CDS). The amino acid sequence is provided at the bottom in the single letter amino acid code.*

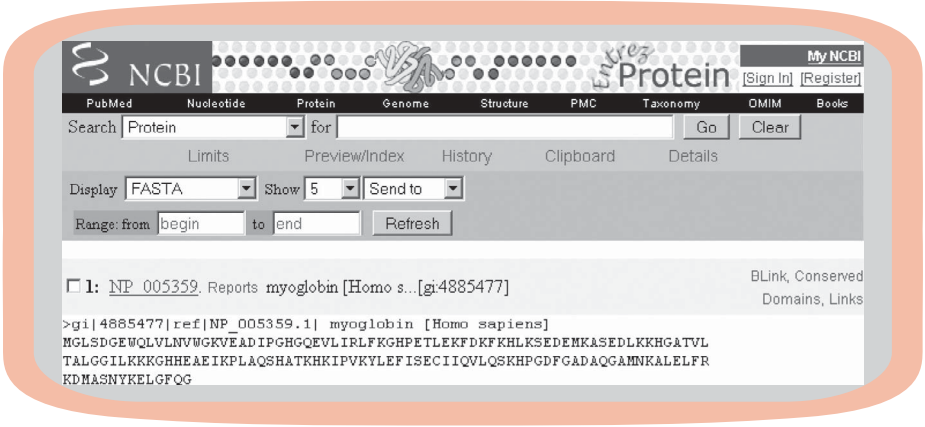



*FIGURE 2.9. The protein entry for human myogobin can be displayed in the FASTA format. This is easily accomplished by adjusting the "Display" pull-down menu from an Entrez protein record. The FASTA format is used in a variety of software programs that we will use in later chapters.*

FASTA is both an alignment program (described in Chapter 3) and a commonly used sequence format (further described in Chapter 4).

the commonly used FASTA format for protein (or DNA) sequences, as shown in Fig. 2.9. Note also that by clicking the CDS (coding sequence) link of an Entrez Protein or Entrez Nucleotide record (shown in Fig. 2.8b), you can obtain the nucleotides that encode a particular protein, typically beginning with a start methionine (ATG) and ending with a stop codon (TAG, TAA, or TGA). This can be useful for a variety of applications including multiple sequence alignment (Chapter 6) and molecular phylogeny (Chapter 7).

## Relationship of Entrez Gene, Entrez Nucleotide, and Entrez Protein

If you are interested in obtaining information about a particular DNA or protein sequence, it is reasonable to visit Entrez Nucleotide or Entrez Protein and do a search. A variety of search strategies are available, such as limiting the output to a particular organism or taxonomic group of interest, or limiting the output to RefSeq entries.

There are also many advantages to beginning your search through Entrez Gene. There, you can identify the official gene name, and you can be assured of the chromosomal location of the gene (thus providing unambigous information about which particular gene you are studying). Furthermore, each Entrez Gene entry includes a section of reference sequences that provides all the DNA and protein variants that are assigned RefSeq accession numbers.

## Comparison of Entrez Gene and UniGene

As described above, the UniGene project assigns one cluster of sequences to one gene. For example, for *RBP4* there is one UniGene entry with the UniGene accession number Hs.50223. This UniGene entry includes a list of all the GenBank entries, including ESTs, that correspond to the *RBP4* gene. The UniGene entry also includes mapping information, homologies, and expression information (i.e., a list of the tissues from which cDNA libraries were generated that contain ESTs corresponding to the RBP gene).

Entrez Gene now has about 40,000 human gene entries (as of November 2008).

UniGene and Entrez Gene have features in common, such as links to OMIM, homologs, and mapping information. They both show RefSeq accession numbers. There are four main differences between UniGene and Entrez Gene:

1. UniGene has detailed expression information; the regional distributions of cDNA libraries from which particular ESTs have been sequenced are listed.

2. UniGene lists ESTs corresponding to a gene, allowing one to study them in detail.
3. Entrez Gene may provide a more stable description of a particular gene; as described above, UniGene entries may be collapsed as genome-sequencing efforts proceed.
4. Entrez Gene has fewer entries than UniGene, but these entries are better curated.

### Entrez Gene and HomoloGene

The HomoloGene database provides groups of annotated proteins from a set of completely sequenced eukaryotic genomes. Proteins are compared (by blastp; see Chapter 4), placed in groups of homologs, and then the protein alignments are matched to the corresponding DNA sequences. This allows distance metrics to be calculated such as Ka/Ks, the ratio of nonsynonymous to synonymous mutations (see Chapter 7). You can find a HomoloGene entry for a gene/protein of interest by following a link on the Entrez Gene page.

A search of HomoloGene with the term hemoglobin results in dozens of matches for myoglobin, alpha globin, and beta globin. By clicking on the beta globin group one gains access to a list of proteins with RefSeq accession numbers from human, chimpanzee, dog, mouse, and chicken. The pairwise alignment scores (see Chapter 3) are summarized and linked to, and the sequences can be displayed as a multiple sequence alignment (Chapter 6), or in the FASTA format.

HomoloGene is available by clicking All Databases from the NCBI home page, or at ► http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene. Release 53 (March 2007) has over 170,000 groups. We will define homologs in Chapter 3.

## Access to Information: Protein Databases

In many cases you are interested in obtaining protein sequences. The Entrez Protein database at NCBI consists of translated coding regions from GenBank as well as sequences from external databases (the Protein Information Resource [PIR], SWISS-PROT, Protein Research Foundation [PRF], and the Protein Data Bank [PDB]). The EBI also provides information on proteins via these major databases. We will next explore ways to obtain protein data through UniProt, an authoritative and comprehensive protein database.

EBI offers access to over a dozen different protein databases, listed at ► http://www.ebi.ac.uk/Databases/protein.html.

### UniProt

The Universal Protein Resource (UniProt) is the most comprehensive, centralized protein sequence catalog (UniProt Consortium, 2009). Formed as a collaborative effort in 2002, it consists of a combination of three key databases. (1) Swiss-Prot is considered the best-annotated protein database, with descriptions of protein structure and function added by expert curators. (2) The translated EMBL (TrEMBL) Nucleotide Sequence Database Library provides automated (rather than manual) annotations of proteins not in Swiss-Prot. It was created because of the vast number of protein sequences that have become available through genome sequencing projects. (3) PIR maintains the Protein Sequence Database, another protein database curated by experts.

UniProt is organized in three database layers. (1) The UniProt Knowledgebase (UniProtKB) is the central database that is divided into the manually annotated UniProtKB/Swiss-Prot and the computationally annotated UniProtKB/TrEMBL.

The European Bioinformatics Institute (EBI) in Hinxton and the Swiss Institute of Bioinformatics (SIB) in Geneva created Swiss-Prot and TrEMBL. PIR is a division of the National Biomedical Research Foundation (► http://pir.georgetown.edu/) in Washington, D.C. PIR was founded by Margaret Dayhoff, whose work is described in Chapter 3. The UniProt web site is ► http://www.uniprot.org. It contains over 7 million entries (release 14.4, November 2008).

To access UniProt from EBI, visit ► http://www.ebi.ac.uk/uniprot/. To access UniProt from ExPASy, visit ► http://www.expasy.org/sprot/.

(2) The UniProt Reference Clusters (UniRef) offer nonredundant reference clusters based on UniProtKB. UniRef clusters are available with members sharing at least 50%, 90%, or 100% identity. (3) The UniProt Archive, UniParc, consists of a stable, nonredundant archive of protein sequences from a wide variety of sources (including model organism databases, patent offices, RefSeq, and Ensembl).

You can access UniProt directly from its website, or from EBI or ExPASy.

## The Sequence Retrieval System at ExPASy

One of the most useful resources available to obtain protein sequences and associated data is provided by ExPASy, the Expert Protein Analysis System. The ExPASy server is a major resource for proteomics-related analysis tools, software, and databases. In addition to providing access to the UniProt database, ExPASy serves as a portal for the Sequence Retrieval System (SRS). The query page has four rectangular boxes (Fig. 2.10). Each has an associated pull-down menu, and as a default condition each says "AllText." In the first box, type "retinol-binding." (Note that queries should consist of one word.) In the second box, type "human," change the corresponding pull-down menu to "organism," then click "do query." You see 10 entries listed. Click the link in which we are interested (SWISS_PROT: RETB_HUMAN P02753).

ExPASy is a proteomics server of the Swiss Institute of Bioinformatics (► http://www.expasy.ch/), another portal from which the Sequence Retrieval System (SRS) is accessed. From ► http://www.expasy.ch/srs5/, click "Start a new SRS session," then click "continue." SRS was created by Lion Biosciences, and a list of several dozen publicly available SRS servers is at ► http://downloads.lionbio.co.uk/publicsrs.html.

An output consists of a SwissProt record. This provides very useful, well-organized information, including alternative names and accession numbers; literature links; functional data and information about cellular localization; links to GenBank and other database records for both the RBP protein and gene; and links to many databases such as OMIM, InterPro, Pfam, Prints, GeneCards, PROSITE, and two-dimensional protein gel databases. We will describe these resources later (Chapters 6 and 10). The record includes features; note that by clicking on any of the linked features, you can see the protein sequence with that feature highlighted in color. While we have mentioned several key ways to acquire sequence data, there are dozens of other useful servers. As an example, the Protein Information Resource (PIR) provides access to sequences (Wu et al., 2002). PIR is especially useful for its efforts to annotate functional information on proteins.

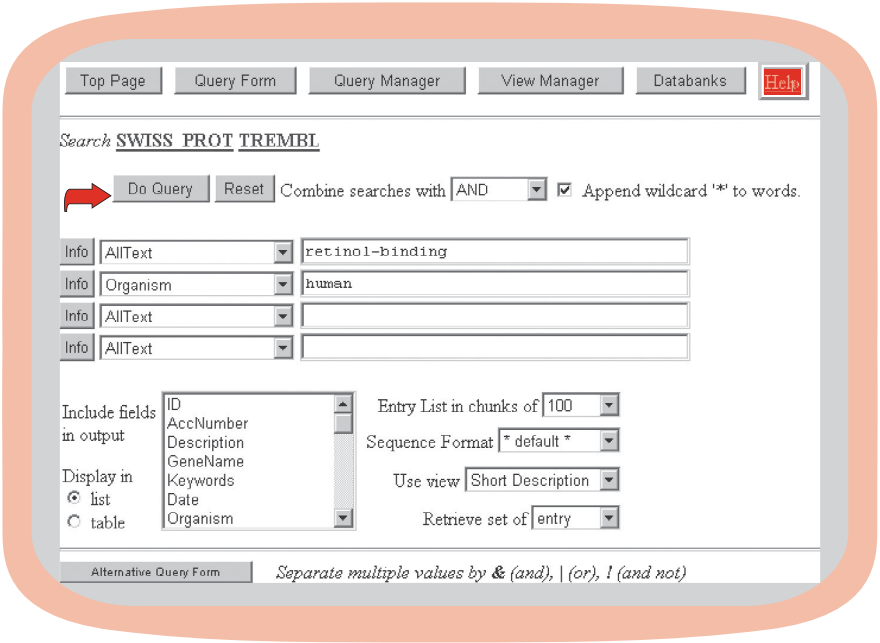



*FIGURE 2.10. Format of a query at the Sequence Retrieval System (SRS) of the Expert Protein Analysis System (ExPASy) (► http://www.expasy.ch/srs5/). This website provides one of the most useful resources for protein analysis. You can also access the SRS through other sites such as the European Bioinformatics Institute (► http://srs6.ebi.ac.uk/).*

# Access to Information: The Three Main Genome Browsers

Genome browsers are databases with a graphical interface that presents a representation of sequence information and other data as a function of position across the chromosomes. We will focus on viral, prokaryotic, and eukaryotic chromosomes in Chapters 14 to 19. Genome browsers have emerged as an essential tool for organizing information about genomes. We will now briefly introduce the three principal genome browsers and describe how they may be used to acquire information about a gene or protein of interest.

Genomes are analyzed over time in assemblies (see Chapter 13). The main human genome browsers share the same underlying assemblies, and differ in the ways they annotate and present information. NCBI Build 36 (November, 2005) is an example of a human assembly.

## The Map Viewer at NCBI

The NCBI Map Viewer includes chromosomal maps (both physical maps and genetic maps; see Chapter 16) for a variety of organisms, including metazoans (animals), fungi, and plants. Map Viewer allows text-based queries (e.g., "beta globin") or sequence-based queries (e.g., BLAST; see Chapter 4). For each genome, four levels of detail are available: (1) the home page of an organism; (2) the genome view, showing ideograms (representations of the chromosomes); (3) the map view, allowing you to view regions at various levels of resolution; and (4) the sequence view, displaying sequence data as well as annotation of interest such as the location of genes.

The Map Viewer is accessed from the main page of NCBI or via ► http://www.ncbi.nlm.nih.gov/mapview/. Records in Entrez Gene, Entrez Nucleotide, and Entrez Protein also provide direct links to the Map Viewer.

## The University of California, Santa Cruz (UCSC) Genome Browser

The UCSC browser currently supports the analysis of three dozen vertebrate and invertebrate genomes, and it is perhaps the most widely used genome browser for human and other prominent organisms such as mouse. The Genome Browser provides graphical views of chromosomal locations at various levels of resolution (from several base pairs up to hundreds of millions of base pairs spanning an entire chromosome). Each chromosomal view is accompanied by horizontally oriented annotation tracks. There are hundreds of available tracks in categories such as mapping and sequencing, phenotype and disease associations, genes, expression, comparative genomics, and genomic variation. These annotation tracks offer the Genome Browser tremendous depth and flexibility. The Genome Browser has a complementary, interconnected Table Browser that provides tabular output of information.

The UCSC genome browser is available from the UCSC bioinformatics site at ► http://genome.ucsc.edu. You can see examples of it in Figs. 5.17, 5.20, 6.10, 8.8, 12.8, 16.4, and 9.20.

As an example of how to use the browser, go the UCSC bioinformatics site, click Genome Browser, set the clade (group) to Vertebrate, the genome to human, the assembly to March 2006 (or any other build date), and under "position or search term" type beta globin (Fig. 2.11a). Click submit and you will see a list of known genes and a RefSeq gene entry for beta globin on chromosome 11 (Fig. 2.11b). By following this RefSeq link you will view the beta globin gene (spanning about 1600 base pairs) on chromosome 11, and can perform detailed analyses of the beta globin gene (including neighboring regulatory elements), the messenger RNA (see Chapter 8), and the protein (Fig. 2.11c).

## The Ensembl Genome Browser

The Ensembl project offers a series of comprehensive websites for a variety of eukaryotic organisms (Hubbard et al., 2007). The project's goals are to automatically analyze and annotate genome data (see Chapter 13) and to present genomic data via its

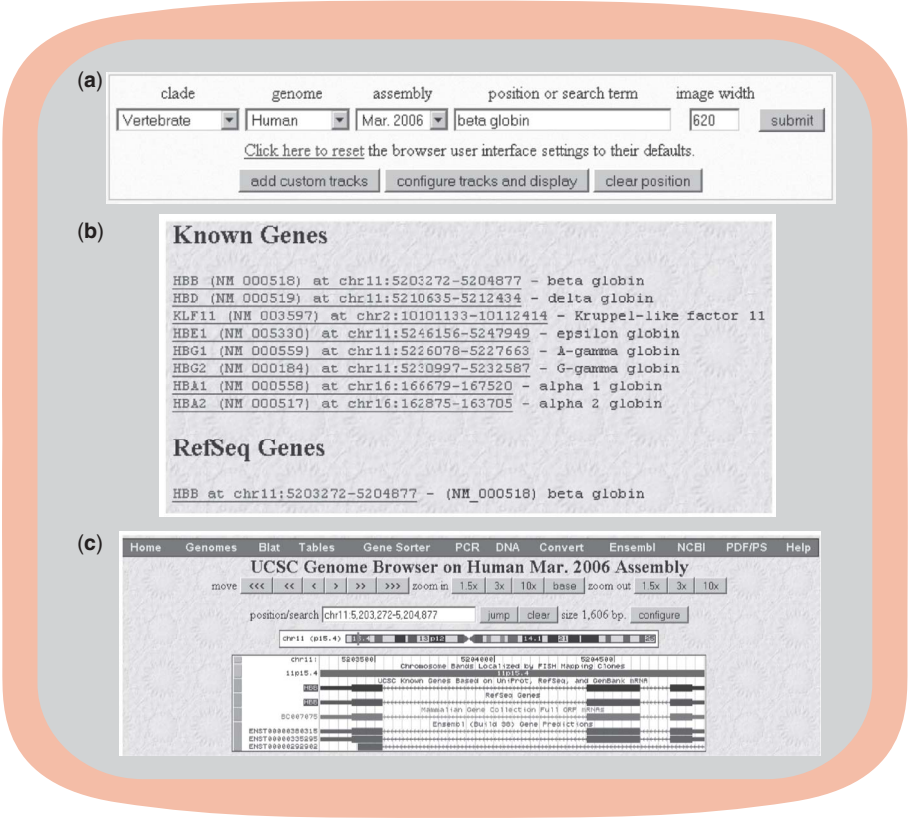



*FIGURE 2.11. Using the UCSC Genome Browser. (a) One can select from dozens of organisms (mostly vertebrates) and assemblies, then enter a query such as "beta globin" (shown here) or an accession number or chromosomal position. (b) By clicking submit, a list of known genes as well as RefSeq genes is displayed. (c) Following the link to the RefSeq gene for beta globin, a browser window is opened showing 1606 base pairs on human chromosome 11. A series of horizontal tracks is displayed including a list of RefSeq genes and Ensembl gene predictions; exons are displayed as thick bars, and arrows indicate the direction of transcription (from right to left, toward the telomere or end of the short arm of chromosome 11). See ► http://genome.ucsc.edu.*

Ensembl (►http://www.ensembl.org) is supported by EMBL and the EBI (►http://www.ebi.ac.uk/) in cooperation with the Wellcome Trust Sanger Institute (WTSI; ►http://www.sanger.ac.uk/). Ensembl focuses on vertebrate genomes, although its genome browser format is being adopted for the analysis of many additional eukaryotic genomes.

web browser. Ensembl is in some ways comparable in scope to the UCSC Genome Browser, although the two offer distinct resources.

We can begin to explore Ensembl from its home page by selecting *Homo sapiens* and doing a text search for "hbb," the gene symbol for beta globin. This yields a link to the beta globin protein and gene; we will return to the Ensembl resource in later chapters. This entry contains a large number of features relevant to HBB, including identifiers, the DNA sequence, and convenient links to many other database resources.

## Examples of How to Access Sequence Data

We will next explore two practical problems in accessing data: the human immunodeficiency virus-1 (HIV-1) pol protein, and human histones. Each presents distinct challenges.

### HIV *pol*

We explore bioinformatics approaches to HIV-1 in detail in Chapter 14 on viruses.

Consider reverse transcriptase, the RNA-dependent DNA polymerase of HIV-1 (Frankel and Young, 1998). The gene-encoding reverse transcriptase is called *pol* (for polymerase). How do you obtain its DNA and protein sequence?

As of November 2008 there are about 250,000 entries in Entrez Nucleotide for the query "hiv-1."

From the home page of NCBI enter "hiv-1" (do not use quotation marks; the use of capital letters is optional). All Entrez databases are searched. Under the Nucleotide category, there are several hundred thousand entries. Click Nucleotide to see these entries. Over 800 entries have RefSeq identifiers; while this narrows the search considerably, there are still too many matches to easily find HIV-1 pol. One reason for the large number of entries in Entrez Nucleotide is that the HIV-1 genome has been

resequenced thousands of times in efforts to identify variants. Another reason for the many hits is that entries for a variety of organisms, including mouse and human, refer to HIV-1 and thus are listed in the output. Performing a search with the query "hiv-1 pol" further reduces the number of matches, but there are still several thousand.

A useful alternate strategy is to limit the search to the organism you are interested in. Begin the search again from the home page of NCBI by clicking "Taxonomy Browser" (along the top bar), and entering Hiv-1. Next follow the link to the taxonomy page specific to HIV-1 (Fig. 2.12). Here you will find the taxonomy identifier for HIV-1; each organism or group in GenBank (e.g., kingdom, phylum, order, genus, species) is assigned a unique identifier. Also, there is an extremely useful table of links to Entrez records. By clicking on the link to Entrez Nucleotide (Fig. 2.11, right side), you will find all the records of sequences from HIV-1, but no records from any other organisms. There is now only one RefSeq entry (NC_001802). This entry refers to the 9181 bases that constitute HIV-1, encoding just nine genes including gag-pol. Given the thousands of HIV-1 pol variants that exist, this example highlights the usefulness of the RefSeq project, allowing the research community to have a common reference sequence to explore.

As an alternative strategy, from the Entrez table on the HIV-1 taxonomy page one can link to the single Entrez Genome record for HIV-1, and find a table of the nine genes (and nine proteins) encoded by the genome. Each of these nine Entrez Genome records contains detailed information on the genes; in the case of gag-pol, there are seven separate RefSeq entries, including one for the gag-pol precursor (NP_057849, 1435 amino acids in length) and one for the mature HIV-1 pol protein (NP_789740, 995 amino acids).

Note that other NCBI databases are not appropriate for finding the sequence of a viral reverse transcriptase: UniGene does not incorporate viral records, while OMIM is limited to human entries. UniGene and OMIM, however, do have links to genes that are related to HIV, such as eukaryotic reverse transcriptases.

We will see that BLAST searches (Chapter 4) can be limited by any Entrez query; you can enter the taxonomy identifier into a BLAST search to restrict the output to any organism or taxonomic group of interest.

From the Entrez Genome or other Entrez pages, try exploring the various options under the Display pull-down menu. For example, for the Entrez Genome entry for NC_001802 you can display a convenient protein table; from Entrez Nucleotide or Entrez Protein you can select Graph to obtain a schematic view of the HIV-1 genome and the genes and proteins it encodes.

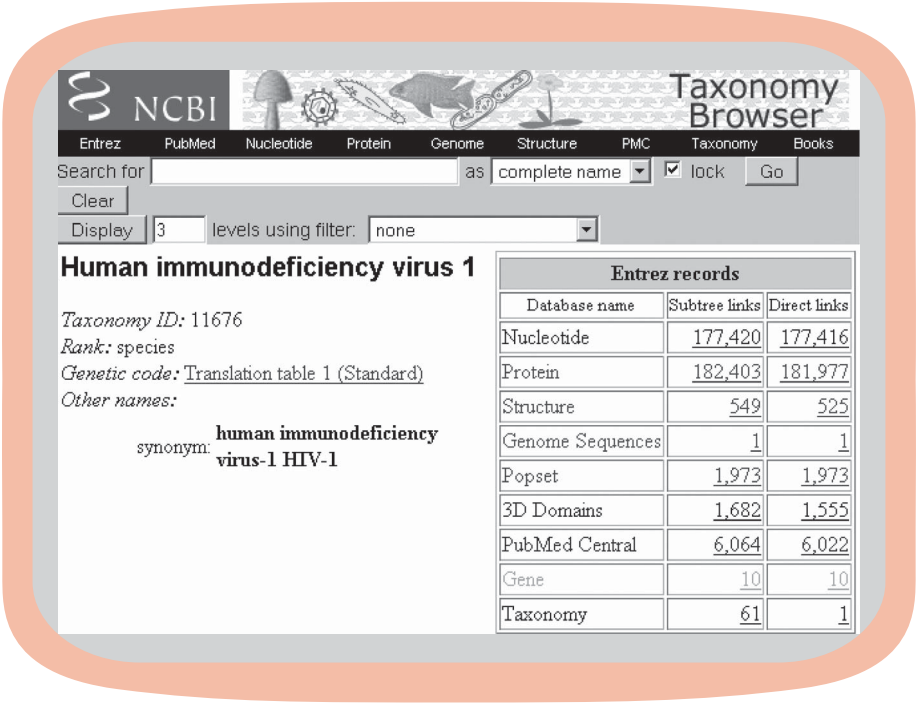



*FIGURE 2.12. The entry for human immunodeficiency virus 1 (HIV-1) at the NCBI Taxonomy Browser displays information about the genus and species as well as a variety of links to Entrez records. By following these links, one can obtain a list of proteins, genes, DNA sequences, structures, or other data types that are restricted to this organism. This can be a useful strategy to find a protein or gene from a particular organism (e.g., a species or subspecies of interest), excluding data from all other species. By following the Entrez Genome Sequences link, one can access a list of nine known HIV-1 protein-coding genes.*

In a separate approach, one can obtain the HIV-1 reverse transcriptase sequence from SRS. Select the SwissProt database to search. In the four available dialog boxes, set one row to "organism" and "HIV-1," then set another row to "AllText" and "reverse." Upon clicking "Do query," a list of several dozen entries is returned; many of these are identified as fragments and may be ignored. One entry is SWISS_PROT:POL_ HV1A2 (SWISS-PROT accession P03369), a protein of 1437 amino acids. Following the SwissProt link, one finds the "NiceProt" for this database entry. This information includes entry and modification dates, names of this protein and synonyms, references (with PubMed links), comments (including a brief functional description), cross-references to over a dozen other useful databases, a keyword listing, features such as predicted secondary structure, and finally, the amino acid sequence in the single-letter amino acid code and the predicted molecular weight of the protein. For this case, the gene encodes a protein as an unprocessed precursor that is further cleaved to generate many smaller proteins, including matrix protein p17, capsid protein p24, nucleocapsid protein p7, a viral protease, a reverse transcriptase/ribonuclease H multifunctional protein, and an integrase. These features are clearly described in the UniProtKB/Swiss-Prot entry for P03369.

## Histones

By clicking the Details tab on an Entrez Protein search, you can see that the command is interpreted as "txid9606[Organism:exp] AND histone[All Fields]". The Boolean operator AND is included between search terms by default.

The biological complexity of proteins can be astonishing, and accessing information about some proteins can be extraordinarily challenging. Histones are among the most familiar proteins by name. They are small proteins (12 to 20 kilodaltons) that are localized to the nucleus where they interact with DNA. There are five major histone subtypes as well as additional variant forms; the major forms serve as core histones (the H2A, H2B, H3, and H4 families) which ~147 base pairs of DNA wrap around, and linker histones (the H1 family). Suppose you want to inspect a typical human histone for the purpose of understanding the properties of a representative gene and its corresponding protein. A challenge is that there are currently 80,000 histone entries in Entrez Protein (November 2008). Restricting the output to human histone proteins (using the command "txid9606[Organism:exp] histone") there are currently 5000 human histone proteins, of which 1200 have RefSeq accession numbers. Some of these are histone deacetylases and histone acetyltransferases; by expanding the query to "txid9606[Organism:exp] AND histone[All Fields] NOT deacetylase NOT acetyltransferase" there are 800 proteins with RefSeq accession numbers. There are many additional strategies for limiting Entrez searches (Box 2.2).

The Histone Sequence Database is available at ► http://research.nhgri.nih.gov/histones/ (Sullivan et al., 2002). It was created by David Landsman, Andy Baxevanis, and colleagues at the National Human Genome Research Institute.

You can find links to a large collection of specialized databases at ► http://www.expasy.org/links.html, the Life Science Directory at the ExPASy (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics (SIB).

How can the search be further pursued? (1) You may select a histone at random and study it although you may not know whether it is representative. (2) There are specialized, expert-curated databases available online for many genes, proteins, diseases, and other molecular features of interest. The Histone Sequence Database (Sullivan et al., 2002) shows that the human genome has about 86 histone genes, including a cluster of 68 adjacent genes on chromosome 6p. This information is useful to understand the scope of the family. (3) There are databases of protein families, including Pfam and InterPro. We will introduce these in Chapters 6 (multiple sequence alignment) and 10 (proteomics). Such databases offer succinct descriptions of protein and gene families and can orient you toward identifying representative members.

# Access to Biomedical Literature

The NLM website is ► http://www.nlm.nih.gov/.

The NLM is the world's largest medical library. In 1971 the NLM created MEDLINE (Medical Literature, Analysis, and Retrieval System Online), a

## BOX 2-2
## Tips for Using Entrez Databases

The Boolean operators AND, OR, and NOT must be capitalized. By default, AND is assumed to connect two terms; subject terms are automatically combined.

You can perform a search of a specific phrase by adding quotation marks. This may potentially restrict the output, so it is a good idea to repeat a search with and without quotation marks.

Boolean operators are processed from left to right. If you add parentheses, the enclosed terms will be processed as a unit rather than sequentially. A search of Entrez Gene with the query "globin AND promoter OR enhancer" yields 4800 results; however, by adding parentheses, the query "globin AND (promoter or enhancer)" yields just 70 results.

If you are interested in obtaining results from a particular organism (or from any taxonomic group such as the primates or viruses), try beginning with TaxBrowser to select the organism first. See Fig. 2–11 for a detailed explanation. Adding the search term human[ORGN] will restrict the output to human. Alternatively, you can use the taxonomy identifier for human, 9606, as follows: txid9606[Organism:exp]

A variety of limiters can be added. In Entrez Protein, the search 500000:999999[Molecular weight] will return proteins having a molecular weight from 500,000 to 1 million daltons. If you would like to see proteins between 10,000 and 50,000 daltons that I have worked on, enter 010000:050000[Molecular weight] pevsner j (or, equivalently, 010000[MOLWT]: 050000[MOLWT] AND pevsner j[Author]).

By truncating a query with an asterisk, you can search for all records that begin with a particular text string. For example, a search of Entrez Nucleotide with the query "globin" returns 5800 results; querying with "glob*" returns 8.2 million results. These include entries with the species *Chaetomium globosum* or the word global.

Keep in mind that any Entrez query can be applied to a BLAST search to restrict its output (Chapter 4).

bibliographic database. MEDLINE currently contains over 18 million references to journal articles in the life sciences with citations from over 4300 biomedical journals in 70 countries. Free access to MEDLINE is provided on the World Wide Web through PubMed (► http://www.ncbi.nlm.nih.gov/PubMed/), which is developed by NCBI. While MEDLINE and PubMed both provide bibliographic citations, PubMed also contains links to online full-text journal articles. PubMed also provides access and links to the integrated molecular biology databases maintained by NCBI. These databases contain DNA and protein sequences, genome-mapping data, and three-dimensional protein structures.

MEDLINE is also accessible through the SRS at the European Bioinformatics Institute via ► http://srs.ebi.ac.uk/. A PubMed tutorial is offered at ► http://www.nlm.nih.gov/bsd/pubmed_tutorial/m1001.html. The growth of MEDLINE is described at ► http://www.nlm.nih.gov/bsd/medline_growth.html. Despite the multinational contributions to MEDLINE, the percentage of articles written in English has risen from 59% at its inception in 1966 to 92% in the year 2008 (► http://www.nlm.nih.gov/bsd/medline_lang_distr.html).

## PubMed Central and Movement toward Free Journal Access

The biomedical research community has steadily increased access to literature information. Groups such as the Association of Research Libraries (ARL) monitor the migration of publications to an electronic form. Thousands of journals are currently available online. Increasingly, online versions of articles include supplementary material such as molecular data (e.g., the sequence of complete

The National Library of Medicine also offers access to PubMed through NLM Gateway (► http://gateway.nlm.nih.gov). This comprehensive service includes access to a variety of NLM databases not offered through PubMed, such as meeting abstracts and a medical encyclopedia.

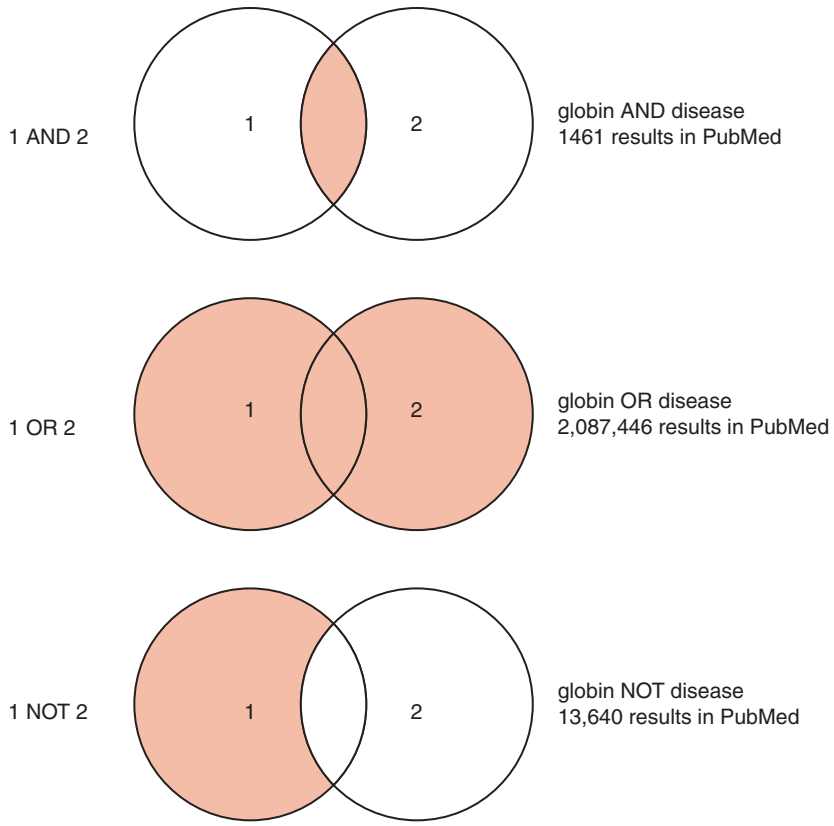
genomes, or gene expression data) or videotapes illustrating an article. PubMed Central provides a central repository for biological literature (Roberts, 2001). All these articles have been peer reviewed and published simultaneously in another journal. As of 2008, publications resulting from research funded by the NIH, Wellcome Trust, and Medical Research Council must be made freely available in PubMed Central.

## Example of PubMed Search: RBP

The ARL website is ► http://www.arl.org/index.shtml.

A search of PubMed for information about "RBP" yields 1700 entries. Box 2.3 describes the basics of using Boolean operators in PubMed. There are many additional ways to limit this search. Press "limits" and try applying features such as restricting the output to articles that are freely available through PubMed Central.

### BOX 2-3
### Venn Diagrams of Boolean Operators AND, OR, and NOT for Hypothetical Search Terms 1 and 2




The AND command restricts the search to entries that are both present in a query. The OR command allows either one or both of the terms to be present. The NOT command excludes query results. The shaded areas represent search queries that are retrieved. Examples are provided for the queries "lipocalin" or "retinol-binding protein" in PubMed. The Boolean operators affect the searches as indicated.

The Medical Subject Headings (MeSH) browser provides a convenient way to focus or expand a search. MeSH is a controlled vocabulary thesaurus containing 25,000 descriptors (headings). From PubMed, click "MeSH Database" on the left sidebar and enter "retinol-binding protein." The result suggests a series of possibly related topics. By adding MeSH terms, a search can be focused and structured according to the specific information you seek. Lewitter (1998) and Fielding and Powell (2002) discuss strategies for effective MEDLINE searches, such as avoiding inconsistencies in MeSH terminology and finding a balance between sensitivity (i.e., finding relevant articles) and specificity (i.e., excluding irrelevant citations). For example, for a subject that is not well indexed, it is helpful to combine a text keyword with a MeSH term. It can also be helpful to use truncations; for example, the search "therap*" introduces a wildcard that will retrieve variations such as therapy, therapist, and therapeutic. Figure 2.13 provides an example of sensitivity and specificity in a PubMed search for articles on hemoglobin.

The MeSH website at NLM is ► http://www.nlm.nih.gov/mesh/meshhome.html; you can also access MeSH via the NCBI website including its PubMed page.

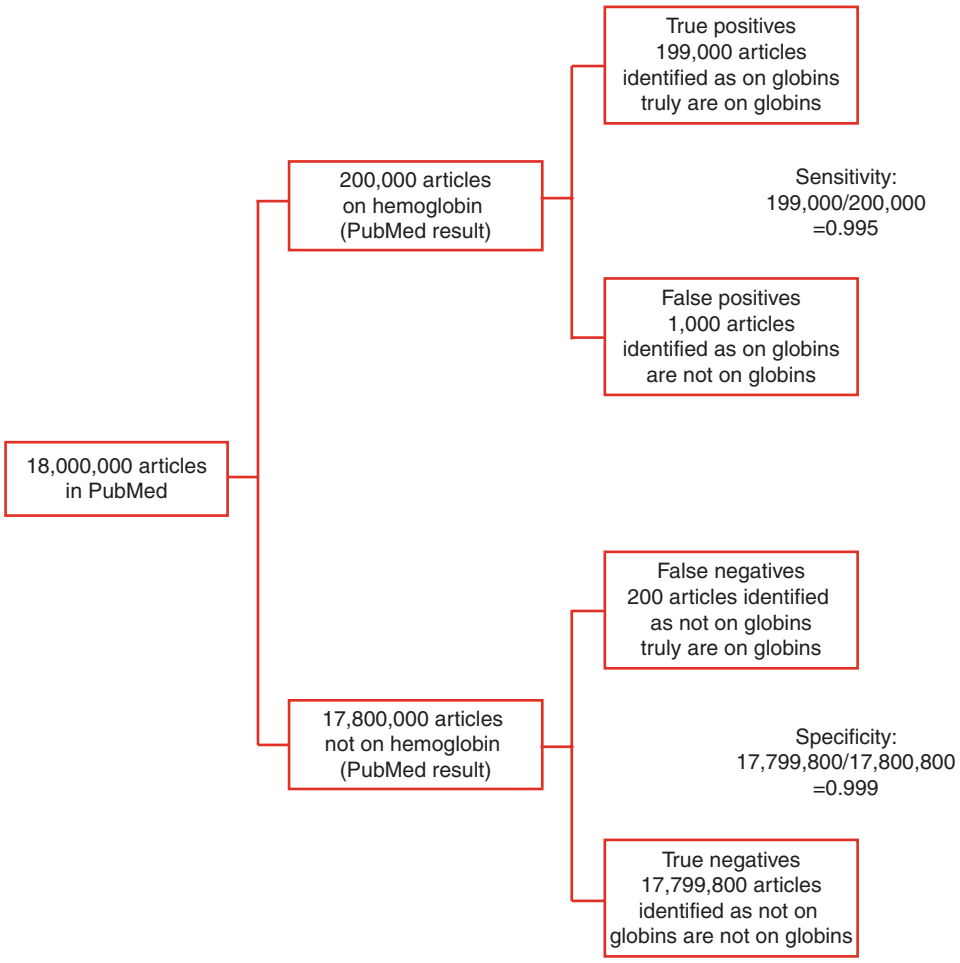



*FIGURE 2.13. Sensitivity and specificity in a database search. We will describe sensitivity and specificity in Chapter 3 (see Fig. 3.27) but can begin thinking about those concepts in terms of a hypothetical search of PubMed for hemoglobin. Each search of a database yields results that are reported (positives) or not (negatives). According to some "gold standard" or objective measure of the truth, these results may be true positives (e.g., a search for globins does return literature citations on globins) or false positives (e.g., a search for glob* returns information about the species C. globosum but those citations are irrelevant to globins). The sensitivity is defined as the proportion of true positives relative to true plus false positives. There also will be many negative results (lower portion of figure). These may include true negatives (e.g., articles that do not describe globins and are not included in the search results) and false negatives (e.g., articles that do discuss globins but are not part of the search results; this might occur if the title and abstract do not mention globins but the body of the article does). Specificity may be defined as the proportion of true negative results divided by the sum of true negative and false positive results.*

## Perspective

Bioinformatics is a young, emerging field whose defining feature is the accumulation of biological information in databases. The three major DNA databases—GenBank, EMBL, and DDBJ—are adding several million new sequences each year as well as billions of nucleotides. Beginning in 2008, terabases (thousands of gigabases) of DNA sequence are arriving.

In this chapter, we described ways to find information on the DNA and/or protein sequence of globins, RBP4, and the HIV *pol* gene. In addition to the three major databases, a variety of additional resources are available on the web. Increasingly, there is no single correct way to find information—many approaches are possible. Moreover, resources such as those described in this chapter—NCBI, ExPASy, EBI/EMBL, and Ensembl—are closely interrelated, providing links between the databases.

## Pitfalls

There are many pitfalls associated with the acquisition of both sequence and literature information. In any search, the most important first step is to define your goal: for example, decide whether you want protein or DNA sequence data. A common difficulty that is encountered in database searches is receiving too much information; this problem can be addressed by learning how to generate specific searches with appropriate limits.

## Web Resources

You can visit the website for this book (▶ http://www.bioinfbook.org) to find many of the URLs, organized by chapter. The Wiley-Blackwell website for this book is http://www.wiley.com/go/pevsnerbioinformatics.

## Discussion Questions

[2-1] What categories of errors occur in databases? How are these errors assessed?

[2-2] How is quality control maintained in GenBank, given that thousands of individual investigators submit data?

## Problems

[2-1] In this chapter we explored histones as an example of a protein that can be challenging to study because it is part of a large gene family. Another challenging example is ubiquitin. How many ubiquitins are there in the human genome, and what is the sequence of a prototypical (that is, representative) ubiquitin?

[2-2] How many human proteins are bigger than 300,000 daltons? *Hints:* Try to first limit your search to human by using TaxBrowser. Then follow the link to Entrez Protein, where all the results will be limited to human. Enter a command in the format xxxxxx:yyyyyy[molwt] to restrict the output to a certain number of daltons; for example, 002000:010000[molwt] will select proteins of molecular weight 2,000 to 10,000.

[2-3] You are interested in learning about genes involved in breast cancer. Which genes have been implicated? What are the DNA and protein accession numbers for several of these genes? Try all of these approaches: PubMed, Entrez, OMIM, and SRS at ExPASy.

[2-4] An ATP (adenosine triphosphate) binding cassette (ABC) is an example of a common protein domain that is found in many so-called ABC transporter proteins. However, you are not familiar with this motif and would like to learn more. Approximately

how many human proteins have ABC domains? Approximately how many bacterial proteins have ABC domains? Which of the resources you used in problem 2.3 is most useful in providing you a clear definition of an ABC motif ? (We will discuss additional resources to solve this problem in Chapter 10.)

[2-5] Find the accession number of a lipocalin protein (e.g., retinol-binding protein, lactoglobulin, any bacterial lipocalin, glycodelin, or odorant-binding protein). First, use Entrez, then UniGene, then OMIM. Which approach is most effective? What is the function of this protein?

[2-6] Three prominent tools for *text*-based searching of molecular information are:

- the National Center for Biotechnology Information's PubMed, Entrez, and OMIM tools (► http://www.ncbi.nlm.nih.gov),
- the European Bioinformatics Institute (EBI) Sequence Retrieval System (SRS) (► http://srs.ebi.ac.uk) or its related SRS site (► http://www.expasy.ch/srs5/), and
- DBGET, the GenomeNet tool of Kyoto University, and the University of Tokyo (► http://www.genome.ad.jp/dbget/dbget2.html) literature database LitDB.

You are interested in learning more about West Nile virus. What happens when you use that query to search each of these three resources?

[2-7] You would like to know what articles about viruses have been published in the journal *BMC Bioinformatics*. Do this search using PubMed.

## SELF-TEST QUIZ

[2-1] Which of the following is a RefSeq accession number corresponding to an mRNA?

(a) J01536

(b) NM_15392

(c) NP_52280

(d) AAB134506

[2-2] Approximately how many human clusters are currently in UniGene?

(a) About 8,000

(b) About 25,000

(c) About 100,000

(d) About 300,000

[2-3] You have a favorite gene, and you want to determine in what tissues it is expressed. Which one of the following resources is likely the most direct route to this information?

(a) UniGene

(b) Entrez

(c) PubMed

(d) PCR

[2-4] Is it possible for a single gene to have more than one UniGene cluster?

(a) Yes

(b) No

[2-5] Which of the following databases is derived from mRNA information?

(a) dbEST

(b) PBD

(c) OMIM

(d) HTGS

[2-6] Which of the following databases can be used to access text information about human diseases?

(a) EST

(b) PBD

(c) OMIM

(d) HTGS

[2-7] What is the difference between RefSeq and GenBank?

(a) RefSeq includes publicly available DNA sequences submitted from individual laboratories and sequencing projects.

(b) GenBank provides nonredundant curated data.

(c) GenBank sequences are derived from RefSeq.

(d) RefSeq sequences are derived from GenBank and provide nonredundant curated data.

[2-8] If you want literature information, what is the best website to visit?

(a) OMIM

(b) Entrez

(c) PubMed

(d) PROSITE

[2-9] Compare the use of Entrez and ExPASy to retrieve information about a protein sequence.

(a) Entrez is likely to yield a more comprehensive search because GenBank has more data than EMBL.

(b) The search results are likely to be identical because the underlying raw data from GenBank and EMBL are the same.

(c) The search results are likely to be comparable, but the SwissProt record from ExPASy will offer a different output format with distinct kinds of information.

## Suggested Reading

Bioinformatics databases are evolving extremely rapidly. Each January, the first issue of the journal *Nucleic Acids Research* includes nearly 100 brief articles on databases. These include descriptions of NCBI (Wheeler et al., 2007), GenBank (Benson et al., 2009), and EMBL (Cochrane et al., 2008).

## References


Adams, M. D., et al. Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* **252**, 1651–1656 (1991).

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

Altschul, S. F., et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

Beach, E. F. Beccari of Bologna. The discoverer of vegetable protein. *J. Hist. Med.* **16**, 354–373 (1961).

Bentley, D. R., et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–39 (2008).

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. GenBank. *Nucl. Acids Res.* **37**, D26–D31 (2009).

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370 (2003).

Boguski, M. S., Lowe, T. M., and Tolstoshev, C. M. dbEST—database for "expressed sequence tags." *Nat. Genet.* **4**, 332–333 (1993).

Brooksbank, C., Camon, E., Harris, M. A., Magrane, M., Martin, M. J., Mulder, N., O'Donovan, C., Parkinson, H., Tuli, M. A., Apweiler, R., Birney, E., Brazma, A., Henrick, K., Lopez, R., Stoesser, G., Stoehr, P., and Cameron, G. The European Bioinformatics Institute's data resources. *Nucleic Acids Res.* **31**, 43–50 (2003).

Cochrane, G., et al. Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. *Nucleic Acids Res.* **36**, D5–D12 (2008).

Fielding, A. M., and Powell, A. Using Medline to achieve an evidence-based approach to diagnostic clinical biochemistry. *Ann. Clin. Biochem.* **39**, 345–350 (2002).

Fleischmann, R. D., et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512 (1995).

Frankel, A. D., and Young, J. A. HIV-1: Fifteen proteins and an RNA. *Annu. Rev. Biochem.* **67**, 1–25 (1998).

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514-D517 (2005).

Hubbard, T. J., et al. Ensembl 2007. *Nucleic Acids Res.* **35**, D610–D617 (2007).

Kouranov, A., Xie, L., de la Cruz, J., Chen, L., Westbrook, J., Bourne, P. E., and Berman, H. M. The RCSB PDB information portal for structural genomics. *Nucleic Acids Res.* **34**, D302–D305 (2006).

Kulikova, T., et al. EMBL Nucleotide Sequence Database in 2006. *Nucleic Acids Res.* **35**, D16–D20 (2007).

Lewitter, F. Text-based database searching. *Bioinformatics: A Trends Guide* **19**, 3–5 (1998).

Ley, T. J., et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66–72 (2008).

Maglott, D. R., Katz, K. S., Sicotte, H., and Pruitt, K. D. NCBI's LocusLink and RefSeq. *Nucleic Acids Res.*, **28**, 126–128 (2000).

Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. Entrez Gene: Gene-centered information at NCBI. *Nucleic Acids Res.* **35**, D26–D31 (2007).

Miyazaki, S., Sugawara, H., Ikeo, K., Gojobori, T., and Tateno, Y. DDBJ in the stream of various biological data. *Nucleic Acids Res.* **32**, D31–D34 (2004).

Olson, M., Hood, L., Cantor, C., and Botstein, D. A common language for physical mapping of the human genome. *Science* **245**, 1434–1435 (1989).

Pruitt, K. D., Tatusova, T., Klimke, W., and Maglott, D. R. NCBI Reference Sequences: current status, policy and new initiatives. *Nucl. Acids Res.* **37**, D32–D36 (2009).

Roberts, R. J. PubMed Central: The GenBank of the published literature. *Proc. Natl. Acad. Sci. USA* **98**, 381–382 (2001).

Sullivan, S., Sink, D. W., Trout, K. L., Makalowska, I., Taylor, P. M., Baxevanis, A. D., and Landsman, D. The Histone Database. *Nucleic Acids Res.* **30**, 341–342 (2002).

UniProt Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **35** (Database issue), D193–D197 (2007).

UniProt Consortium. The Universal Protein Resource (UniProt) 2009. *Nucl. Acids Res.* **37**, D169–D174 (2009).

Wang, J., et al. The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).

Wheeler, D. L., et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **35** (Database issue), D5–D12 (2007).

Wu, C. H., et al. The Protein Information Resource: An integrated public resource of functional annotation of proteins. *Nucleic Acids Res.* **30**, 35–37 (2002).

### Adrenocorticotropin (ACTH)

The complete amino acid sequences are known for corticotropins isolated from the anterior pituitary glands of three different species, pig, beef, and sheep. The structure of sheep ACTH was discussed in the last chapter, and the sequences shown in Table 9 include only those areas of the three molecules where differences are to be found. Although some difference between the content of amide nitrogen groups has been reported for the three species, these are not included in the figure since it has not been possible to rule out, with certainty, the possibility that these variations are due, in part, to the rigors of the isolation and purification techniques employed.

TABLE 9
Variations in Amino Acid Sequences Among Different Preparations of ACTH

| | | Residue No. | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Preparation | Species | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 |
| β-Corticotropin | sheep, beef[a] | Ala. | Gly. | Glu. | Asp. | Asp. | Glu | Ala. | Ser. | Glu.$NH_2$ |
| Corticotropin A | pig | Asp. | Gly. | Ala. | Glu. | Asp. | Glu | Leu. | Ala. | Glu |

[a] Identity with sheep hormone not absolutely certain but very probable as judged from the nearly complete sequence analysis by J. S. Dixon and C. H. Li (personal communication to the author).

Two points are of particular interest in regard to the sequences shown. First, the corticotropins of sheep and beef are identical and differ from that of the pig. This finding is consonant with the closer phylogenetic relationship of sheep and cows to each other than of either to pigs. Second, chemical differences are found only in that portion of the ACTH molecule which has been shown to be unessential for hormonal activity. Genetic mutations leading to such differences might, therefore, not be expected to impose significant disadvantages in terms of survival, and these genes could become established in the gene pools of the species.

### Melanotropin (MSH)

Melanotropin, like the other hormones considered in this chapter, is a typically chordate polypeptide. Indeed, the demonstration of melanocyte-stimulating activity in extracts of tunicates constitutes an



*Pairwise alignment involves matching two protein or DNA sequences. The first proteins that were sequenced include insulin (by Frederick Sanger and colleagues; see fig. 7.1) and globins. This figure is from The Molecular Basis of Evolution by the Nobel laureate Christian Anfinsen (1959, p. 153). It shows the results of a pairwise alignment of a portion of adrenocorticotropic hormone (ACTH) from sheep or cow (top) with that of pig (below). Such alignments, performed manually, led to the realization that amino acid sequences of proteins reflect the phylogenetic relatedness of different species. Furthermore, pairwise alignments reveal the portions of a protein that may be important for its biological function. Used with permission.*

# 3

# Pairwise Sequence Alignment

## Introduction

One of the most basic questions about a gene or protein is whether it is related to any other gene or protein. Relatedness of two proteins at the sequence level suggests that they are homologous. Relatedness also suggests that they may have common functions. By analyzing many DNA and protein sequences, it is possible to identify domains or motifs that are shared among a group of molecules. These analyses of the relatedness of proteins and genes are accomplished by aligning sequences. As we complete the sequencing of many organisms' genomes, the task of finding out how proteins are related within an organism and between organisms becomes increasingly fundamental to our understanding of life.

Two genes (or proteins) are homologous if they have evolved from a common ancestor.

In this chapter we will introduce pairwise sequence alignment. We will adopt an evolutionary perspective in our description of how amino acids (or nucleotides) in two sequences can be aligned and compared. We will then describe algorithms and programs for pairwise alignment.

### Protein Alignment: Often More Informative Than DNA Alignment

Given the choice of aligning a DNA sequence or the sequence of the protein it encodes, it is often more informative to compare protein sequences. There are several reasons for this. Many changes in a DNA sequence (particularly at

*Bioinformatics and Functional Genomics, Second Edition*. By Jonathan Pevsner

the third position of a codon) do not change the amino acid that is specified. Furthermore, many amino acids share related biophysical properties (e.g., lysine and arginine are both basic amino acids). The important relationships between related (but mismatched) amino acids in an alignment can be accounted for using scoring systems (described in this chapter). DNA sequences are less informative in this regard. Protein sequence comparisons can identify homologous sequences from organisms that last shared a common ancestor over 1 billion years ago (BYA) (e.g., glutathione transferases) (Pearson, 1996). In contrast, DNA sequence comparisons typically allow lookback times of up to about 600 million years ago (MYA).

When a nucleotide coding sequence is analyzed, it is often preferable to study its translated protein. In Chapter 4 (on BLAST searching), we will see that we can move easily between the worlds of DNA and protein. For example, the tblastn tool from the National Center for Biotechnology Information (NCBI) BLAST website allows one to search with a protein sequence for related proteins derived from a DNA database (see Chapter 4). This query option is accomplished by translating each DNA sequence into all of the six proteins that it potentially encodes.

Nevertheless, in many cases it is appropriate to compare nucleotide sequences. This comparison can be important in confirming the identity of a DNA sequence in a database search, in searching for polymorphisms, in analyzing the identity of a cloned cDNA fragment, or in many other applications.

## Definitions: Homology, Similarity, Identity

Let us consider the globin family of proteins. We will begin with human myoglobin (accession number NP_005359) and beta globin (accession number NP_000509) as two proteins that are distantly but significantly related. The accession numbers are obtained from Entrez Gene (Chapter 2). Myoglobin and the hemoglobin chains (alpha, beta, and other) are thought to have diverged some 600 million years ago, near the time the vertebrate and insect lineages diverged.

Some researchers use the term *analogous* to refer to proteins that are not homologous, but share some similarity by chance. Such proteins are presumed to have not descended from a common ancestor.

Two sequences are *homologous* if they share a common evolutionary ancestry. There are no degrees of homology; sequences are either homologous or not (Reeck et al., 1987; Tautz, 1998). Homologous proteins almost always share a significantly related three-dimensional structure. Myoglobin and beta globin have very similar structures as determined by x-ray crystallography (Fig. 3.1). When two sequences are homologous, their amino acid or nucleotide sequences usually share significant identity. Thus, while homology is a qualitative inference (sequences are homologous or not), identity and similarity are quantities that describe the relatedness of sequences. Notably, two molecules may be homologous without sharing statistically significant amino acid (or nucleotide) identity. For example, in the globin family, all the members are homologous, but some have sequences that have diverged so greatly that they share no recognizable sequence identity (e.g., human beta globin and human neuroglobin share only 22% amino acid identity). Perutz, Kendrew and others demonstrated that individual globin chains share the same overall shape as myoglobin (see Ingram, 1963), even though the myoglobin and alpha globin proteins share only about 26% amino acid identity. In general, three-dimensional structures diverge much more slowly than amino acid sequence identity between two proteins

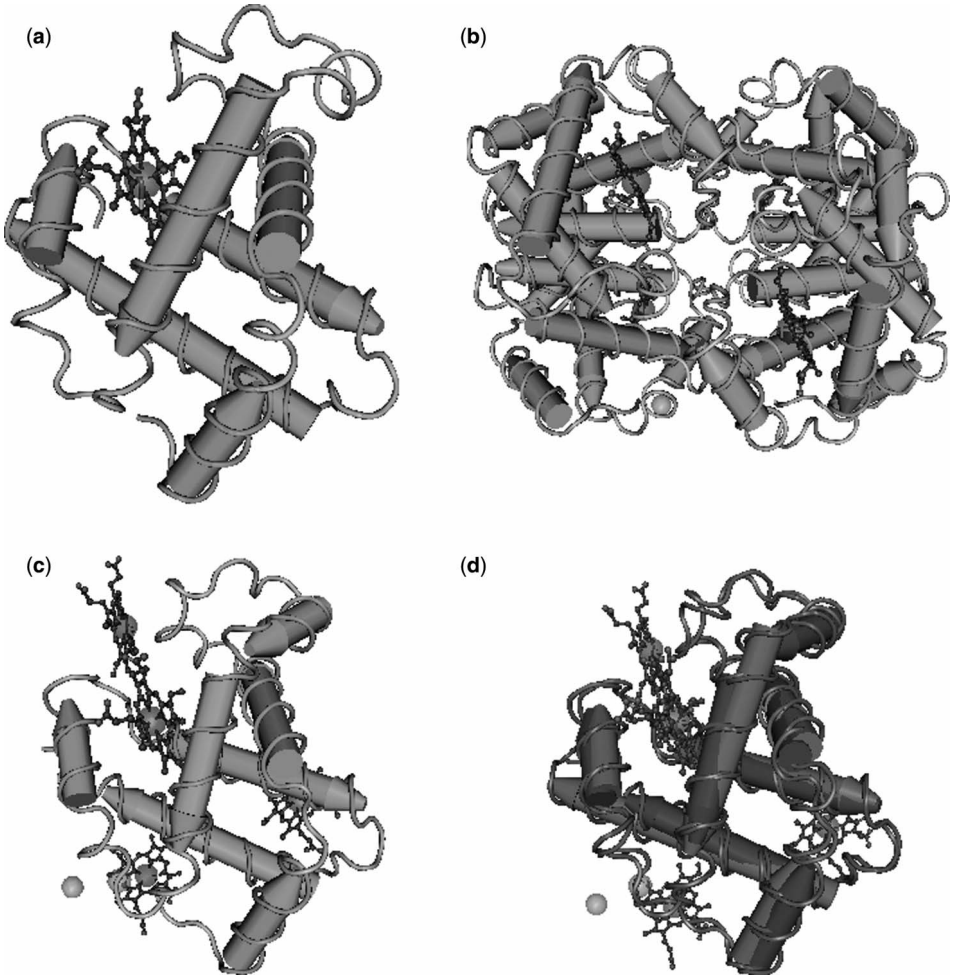



*FIGURE 3.1. Three-dimensional structures of (a) myoglobin (accession 2MM1), (b) the tetrameric hemoglobin protein (2H35), (c) the beta globin subunit of hemoglobin, and (d) myoglobin and beta globin superimposed. The images were generated with the program Cn3D (see Chapter 11). These proteins are homologous (descended from a common ancestor), and they share very similar three-dimensional structures. However, pairwise alignment of these proteins' amino acid sequences reveals that the proteins share very limited amino acid identity.*

(Chothia and Lesk, 1986). Recognizing this type of homology is an especially challenging bioinformatics problem.

Proteins that are homologous may be orthologous or paralogous. *Orthologs* are homologous sequences in different species that arose from a common ancestral gene during speciation. Figure 3.2 shows a tree of myoglobin orthologs. There is a human myoglobin gene and a rat gene. Humans and rodents diverged about 80 MYA (see Chapter 18), at which time a single ancestral myoglobin gene diverged by speciation. Orthologs are presumed to have similar biological functions; in this example, human and rat myoglobins both transport oxygen in muscle cells. *Paralogs* are homologous sequences that arose by a mechanism such as gene duplication. For example, human alpha-1 globin (NP_000549) is paralogous to alpha-2 globin (NP_000508); indeed, these two proteins share 100% amino acid identity. Human alpha-1 globin and beta globin are also paralogs (as are all the proteins shown in Fig. 3.3). All of the globins have distinct properties, including regional distribution in the body, developmental timing of gene expression, and abundance. They are all thought to have distinct but related functions as oxygen carrier proteins.

You can see the protein sequences used to generate Figs. 3.2 and 3.3 in web documents 3.1 and 3.2 at ▶ http://www.bioinfbook.org/chapter3.

In general when we consider other paralogous families they are presumed to share common functions. Consider the lipocalins: all are about 20 kilodalton proteins that have a hydrophobic binding pocket that is thought to be used to transport a hydrophobic ligand. Members include retinol binding protein (a retinol transporter), apolipoprotein D (a cholesterol transporter), and odorant-binding protein (an odorant transporter secreted from a nasal gland).

We thus define homologous genes within the same organism as paralogous. But consider further the case of globins. Human α-globin and β-globin are paralogs, as are mouse α-globin and mouse β-globin. Human α-globin and mouse α-globin are orthologs. What is the relation of human α-globin to mouse β-globin? These could be considered paralogs, because α-globin and β-globin originate from a gene duplication event rather than from a speciation event. However, they are not paralogs because they do not occur in the same species. It may thus be most appropriate to simply call them "homologs," reflecting their descent from a common ancestor. Fitch (1970, p. 113) notes that phylogenies require the study of orthologs (see also Chapter 7).

Richard Owen (1804–1892) was one of the first biologists to use the term homology. He defined homology as "the same organ in different animals under every variety of form and function" (Owen, 1843, p. 379). Charles Darwin (1809–1882) also discussed homology in the sixth edition of *The Origin of Species by means of Natural Selection or, The Preservation of Favoured Races in the Struggle for Life* (1872). He wrote: "That relation between parts which results from their development from corresponding embryonic parts, either in different animals, as in the case of the arm of man, the foreleg of a quadruped, and the wing of a bird; or in the same individual, as in the case of the fore and hind legs in quadrupeds, and the segments or rings and their appendages of which the body of a worm, a centipede, &c., is composed. The latter is called serial homology. The parts which stand in such a relation to each other are said to be homologous, and one such part or organ is called the homologue of the other. In different plants the parts of the flower are homologous, and in general these parts are regarded as homologous with leaves."

Walter M. Fitch (1970, p. 113) defined these terms. He wrote:

> there should be two subclasses of homology. Where the homology is the result of gene duplication so that both copies have descended side by side during the history of an organism (for example, α and β hemoglobin) the genes should be called paralogous (para = in parallel). Where the homology is the result of speciation so that the history of the gene reflects the history of the species (for example α hemoglobin in man and mouse) the genes should be called orthologous (ortho = exact).

Notably, orthologs and paralogs do not necessarily have the same function. We will provide various definitions of gene and protein function in Chapter 10. Later we will explore genomes across the tree of life (Chapters 13 to 19). In all genome sequencing projects, orthologs and paralogs are identified based on database searches. Two DNA (or protein) sequences are defined as homologous based on achieving significant alignment scores, as discussed below and in Chapter 4. However, homologous proteins do not necessarily share the same function.

We can assess the relatedness of any two proteins by performing a *pairwise alignment*. In this procedure, we place the two sequences directly next to each other. One practical way to do this is through the NCBI pairwise BLAST tool (Tatusova and Madden, 1999) (Fig. 3.4). Perform the following steps:

1. Choose the protein BLAST program and select "BLAST 2 sequences" for our comparison of two proteins. An alternative is to select blastn (for "BLAST nucleotides") for DNA–DNA comparison.
2. Enter the sequences or their accession numbers. Here we use the sequence of human beta globin in the fasta format, and for myoglobin we use the accession number (Fig. 3.4).
3. Select any optional parameters.
   - You can choose from five scoring matrices: BLOSUM62, BLOSUM45, BLOSUM80, PAM70, and PAM30. Select PAM250.
   - You can change the gap creation penalty and gap extension penalty.
   - For blastn searches you can change reward and penalty values.
   - There are other parameters you can change, such as word size, expect value, filtering, and dropoff values. We will discuss these more in Chapter 4.
4. Click "BLAST." The output includes a pairwise alignment using the single-letter amino acid code (Fig. 3.5a).

Note that the fasta format uses the single-letter amino acid code; those abbreviations are shown in Box 3.1.

It is extremely difficult to align proteins by visual inspection. Also, if we allow gaps in the alignment to account for deletions or insertions in the two sequences, the number of possible alignments rises exponentially. Clearly, we will need a computer algorithm to perform an alignment (see Box 3.2). In the pairwise alignments shown in Fig. 3.5a, beta globin is on top (on the line labeled query) and myoglobin is below (on the subject line). An intermediate row indicates the presence of *identical* amino acids in the alignment. For example, notice that near the beginning of the alignment the residues WGKV are identical between the two proteins. We can count the total number of identical residues; in this case, the two proteins share

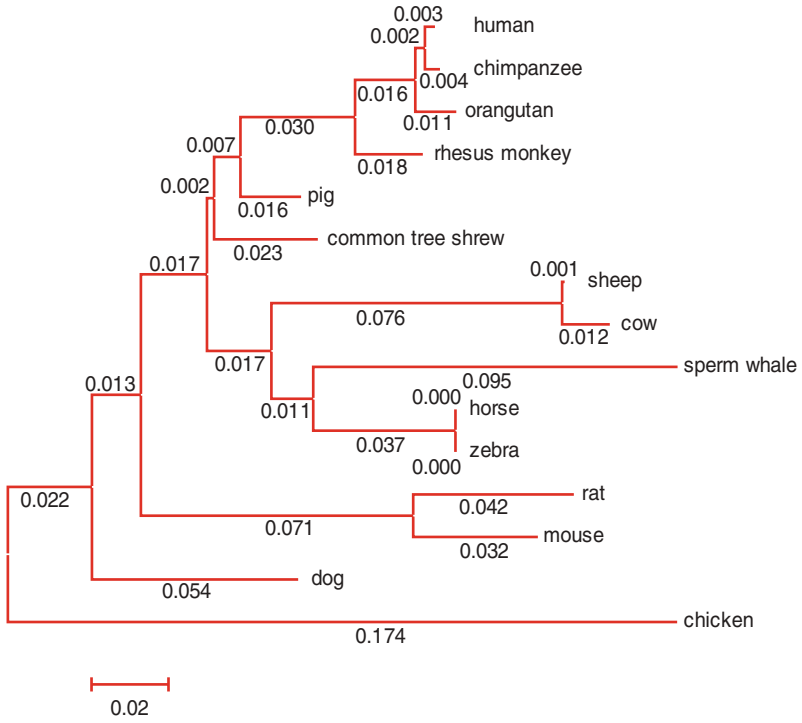



*FIGURE 3.2. A group of myoglobin orthologs, visualized by multiply aligning the sequences (Chapter 6) then creating a phylogenetic tree by neighbor-joining (Chapter 7). The accession numbers and species names are as follows: human, NP_005359 (*Homo sapiens*); chimpanzee, XP_001156591 (*Pan troglodytes*); orangutan, P02148 (*Pongo pygmaeus*); rhesus monkey, XP_001082347 (*Macaca mulatta*); pig, NP_999401 (*Sus scrofa*); common tree shrew, P02165 (*Tupaia glis*); horse, P68082 (*Equus caballus*); zebra, P68083 (*Equus burchellii*); dog, XP_850735 (*Canis familiaris*); sperm whale, P02185 (*Physeter catodon*); sheep, P02190 (*Ovis aries*); rat, NP_067599 (*Rattus norvegicus*); mouse, NP_038621 (*Mus musculus*); cow, NP_776306 (*Bos taurus*); chicken_XP_416292 (*Gallus gallus*). The sequences are shown in web document 3.1 (► http://www.bioinfbook.org/chapter3). In this tree, sequences that are more closely related to each other are grouped closer together. Note that as entire genomes continue to be sequenced (Chapters 13 to 19), the number of known orthologs will grow rapidly for most families of orthologous proteins.*

25% identity (37 of 145 aligned residues). Identity is the extent to which two amino acid (or nucleotide) sequences are invariant. Note that this particular alignment is called *local* because only a subset of the two proteins is aligned: the first and last few amino acid residues of each protein are not displayed. A global pairwise alignment includes all residues of both sequences.

Another aspect of this pairwise alignment is that some of the aligned residues are similar but not identical; they are related to each other because they share similar biochemical properties. *Similar* pairs of residues are structurally or functionally related. For example, on the first row of the alignment we can find threonine and serine (T and S connected by a + sign in Fig. 3.5a); nearby we can see a leucine and a valine residue that are aligned. These are *conservative substitutions*. Amino acids with similar properties include the basic amino acids (K, R, H), acidic amino acids (D, E), hydroxylated amino acids (S, T), and hydrophobic amino acids (W, F, Y, L, I, V, M, A). Later in this chapter we will see how scores are assigned to aligned amino acid residues.

You can access the pairwise BLAST program at the NCBI blast site, ► http://www.ncbi.nlm.nih.gov/BLAST/. We discuss various options for using the Basic Local Alignment Search Tool (BLAST) in Chapter 4. We discuss global and local alignments below.

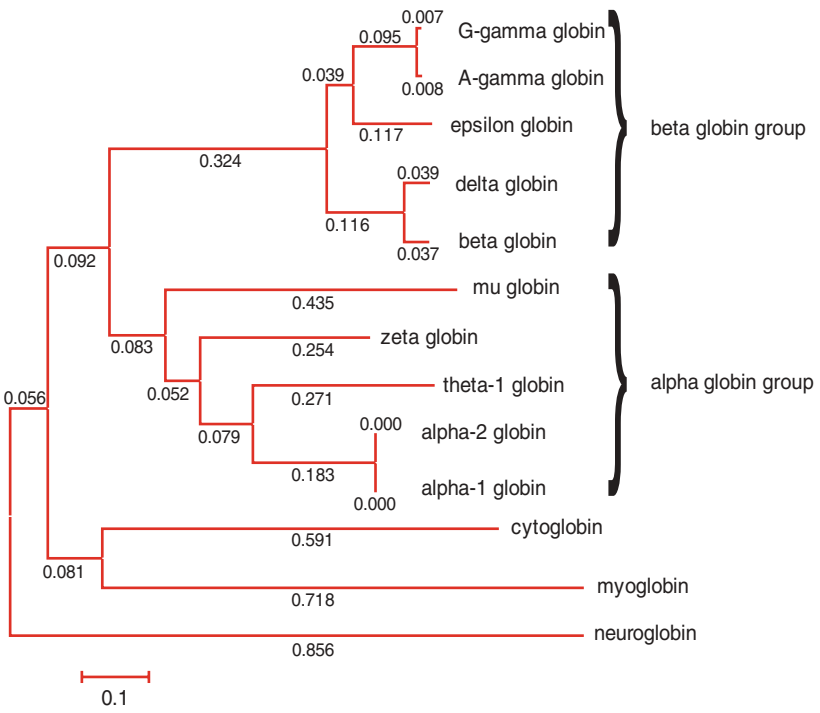



*FIGURE 3.3. Paralogous human globins: Each of these proteins is human, and each is a member of the globin family. This unrooted tree was generated using the neighbor-joining algorithm in MEGA (see Chapter 7). The proteins and their RefSeq accession numbers (also shown in web document 3.2) are delta globin (NP_000510), G-gamma globin (NP_000175), beta globin (NP_000509), A-gamma globin (NP_000550), epsilon globin (NP_005321), zeta globin (NP_005323), alpha-1 globin (NP_000549), alpha-2 globin (NP_000508), theta-1 globin (NP_005322), hemoglobin mu chain (NP_001003938), cytoglobin (NP_599030), myoglobin (NP_005359), and neuroglobin (NP_067080). A Poisson correction model was used (see Chapter 7).*

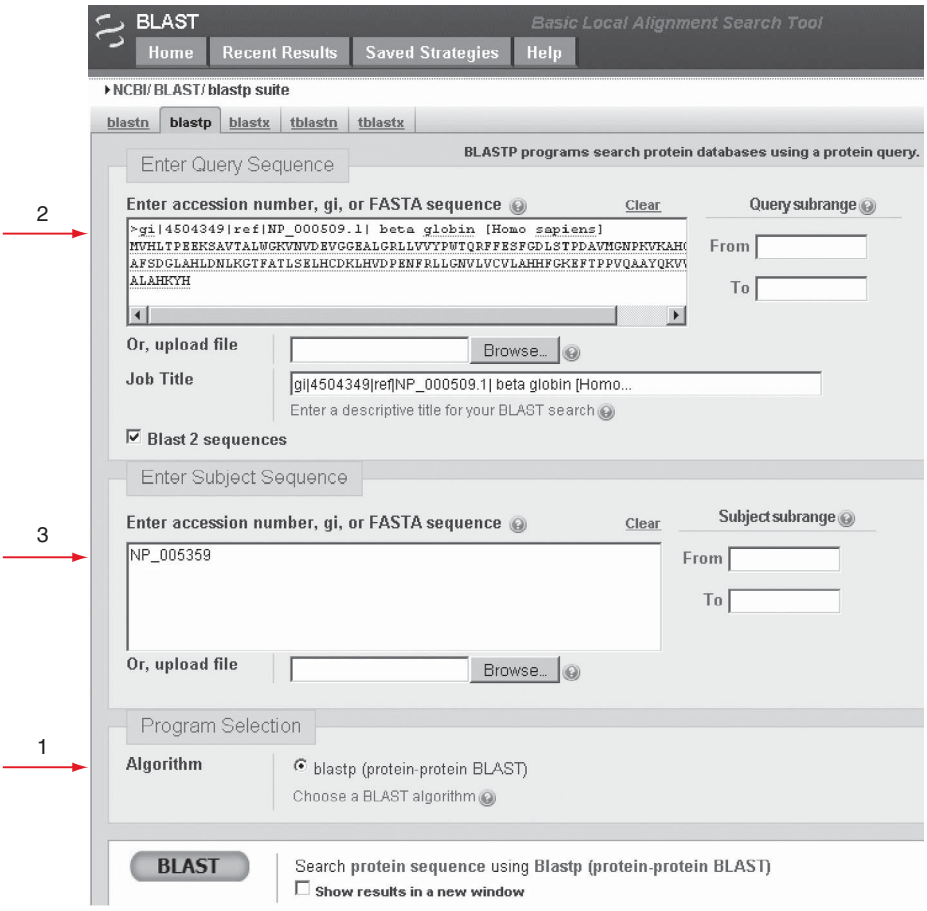



*FIGURE 3.4. The BLAST program at the NCBI website allows the comparison of two DNA or protein sequences. Here the program is set to blastp for the comparison of two proteins (arrow 1). Human beta globin (NP_000509) is input in the fasta format (arrow 2), while human myoglobin (NP_005359) is input as an accession number (arrow 3).*

**(a)**

```
 Score = 43.9 bits (102),  Expect = 1e-09, Method: Composition-based stats.
 Identities = 37/145 (25%), Positives = 57/145 (39%), Gaps = 2/145 (1%)

                                  ▼                   ▼
Query  4    LTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  61
            L+  E   V  +WGKV  D    G E L RL   +P T   F+ F  L + D +  +  +
Sbjct  3    LSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDL  62

Query  62   KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK  121
            K HG  VL A    L    + +     L++ H  K  +  +    +  ++ VL
Sbjct  63   KKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPG  122

Query  122  EFTPPVQAAYQKVVAGVANALAHKY  146
            +F    Q A  K +      +A  Y
Sbjct  123  DFGADAQGAMNKALELFRKDMASNY  147
```

**(b)**

```
 Score = 18.1 bits (35),  Expect = 0.015, Method: Composition-based stats.
 Identities = 11/24 (45%), Positives = 12/24 (50%), Gaps = 2/24 (8%)

Query  12  VTALWGKVNVD--EVGGEALGRLL  33
           V  +WGKV  D    G E L RL
Sbjct  11  VLNVWGKVEADIPGHGQEVLIRLF  34
```

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| match | 4 11 5 6 4 | 6 | 6 5 4 | 5 4 | sum of matches: +60 |
| mismatch | -1 1 -2 | 0 0 | -2 -3 | -2 0 | -4 0 | sum of mismatches: -13 |
| gap open | | | -11 | | | sum of gap penalties: -12 |
| gap extend | | | -1 | | | |
| | | | | | | total raw score: 60 - 13 - 12 = 35 |

*FIGURE 3.5. Pairwise alignment of human beta globin (the "query") and myoglobin (the "subject"). Panel (a) shows the alignment from the search shown in Fig. 3.4. Note that this alignment is local (i.e., the entire lengths of each protein are not compared), and there are many positions of identity between the two sequences (indicated with amino acids intervening between the query and subject lines). The alignment contains an internal gap (indicated by two dashes). Panel (b) illustrates how raw scores are calculated, using the result of a separate search with just amino acids 10–34 of HBB (corresponding to the region between the arrowheads in panel a). The raw score is 35; this represents the sum of the match scores (from a BLOSUM62 matrix in this case), the mismatch scores, the gap opening penalty (set to −11 for this search), and the gap extension penalty (set to −1).*

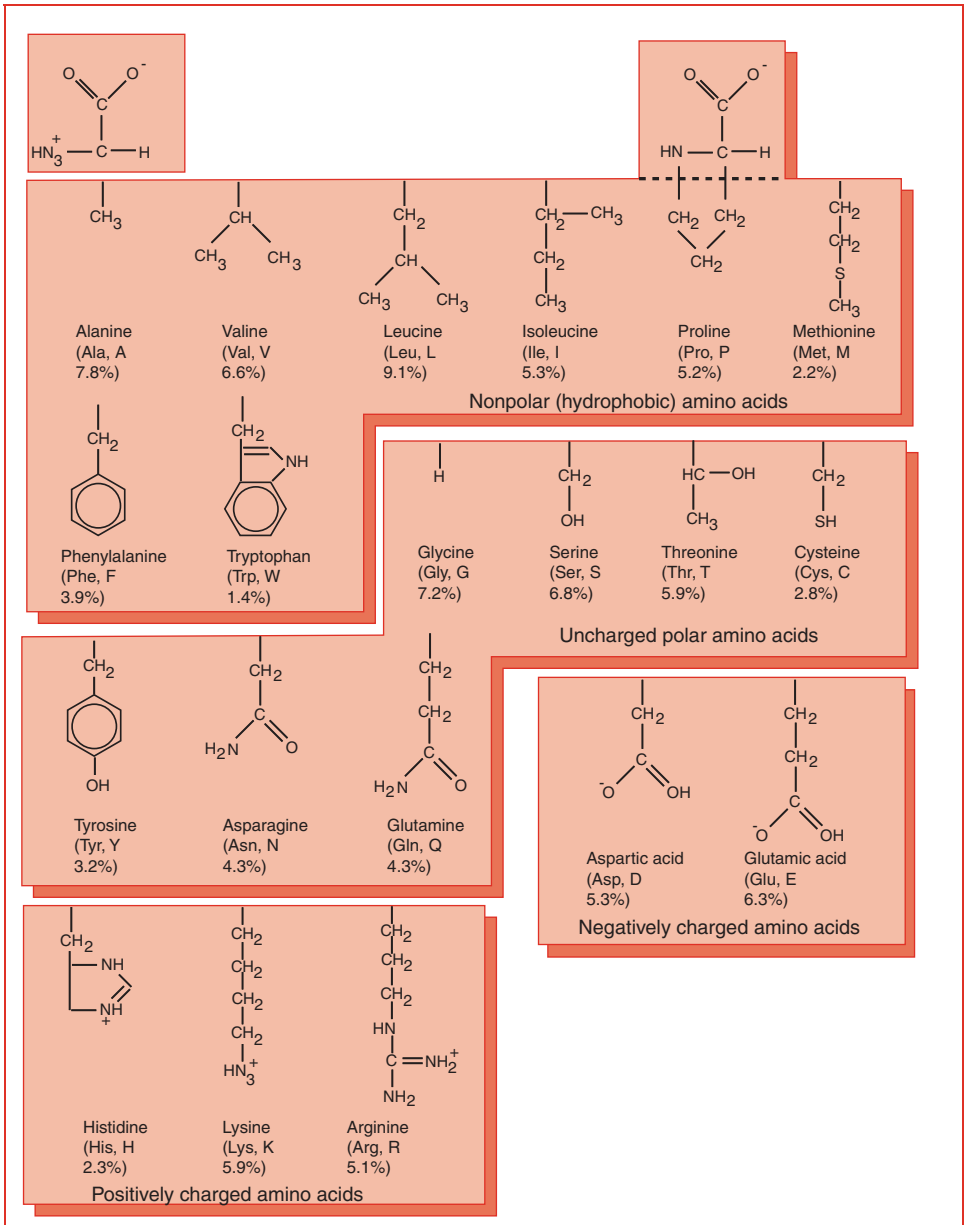
In the pairwise alignment of a segment of HBB and myoglobin, you can see that each pair of residues is assigned a score that is relatively high for matches, and often negative for mismatches.

The *percent similarity* of two protein sequences is the sum of both identical and similar matches. In Fig. 3.5a, there are 57 aligned amino acid residues that are similar. In general, it is more useful to consider the identity shared by two protein sequences, rather than the similarity, because the similarity measure may be based on a variety of definitions of how related (similar) two amino acid residues are to each other.

Two proteins could have similar structures due to convergent evolution. Molecular evolutionary studies are essential (based on sequence analyses) to assess this possibility.

In summary, pairwise alignment is the process of lining up two sequences to achieve maximal levels of identity (and maximal levels of conservation in the case of amino acid alignments). The purpose of a pairwise alignment is to assess the degree of similarity and the possibility of homology between two molecules. We may say that two proteins share, for example, 25% amino acid identity and 39% similarity. If the amount of sequence identity is sufficient, then the two sequences are probably homologous. It is never correct to say that two proteins share a certain percent homology, because they are either homologous or not. Similarly, it is not appropriate to describe two sequences as "highly homologous"; instead one can say that they share a high degree of similiarity. We will discuss the statistical significance of sequence alignments below, including the use of expect values to assess whether an alignment of two sequences is likely to have occurred by chance (Chapter 4).

## Box 3.1
## Structures and One- and Three-Letter Abbreviations of Twenty Common Amino Acids




It is very helpful to memorize these abbreviations and to become familiar with the physical properties of the amino acids. The percentages refer to the relative abundance of each amino acid in proteins.

Such analyses provide evidence to assess the hypothesis that two proteins are homologous. Ultimately the strongest evidence to determine whether two proteins are homologous comes from structural studies in combination with evolutionary analyses.

### Box 3.2
### Algorithms and Programs

An *algorithm* is a procedure that is structured in a computer program (Sedgewick, 1988). For example, there are many algorithms used for pairwise alignment. A computer *program* is a set of instructions that uses an algorithm (or multiple algorithms) to solve a task. For example, the BLAST program (Chapters 3 to 5) uses a set of algorithms to perform sequence alignments. Other programs that we introduce in Chapter 7 use algorithms to generate phylogenetic trees.

Computer programs are essential to solve a variety of bioinformatics problems because millions of operations may need to be performed. The algorithm used by a program provides the means by which the operations of the program are automated. Throughout this book, note how many hundreds of programs have been developed using many hundreds of different algorithms. Each program and algorithm is designed to solve a specific task. An algorithm that is useful to compare one protein sequence to another may not work in a comparison of one sequence to a database of 10 million protein sequences.

Why is it that an algorithm that is useful for comparing two sequences cannot be used to compare millions of sequences? Some problems are so inherently complex that an exhaustive analysis would require a computer with enormous memory or the problem would take an unacceptably long time to complete. A *heuristic algorithm* is one that makes approximations of the best solution without exhaustively considering every possible outcome. The 13 proteins in Fig. 3.2 can be arranged in a tree over a billion distinct ways (see Chapter 7)—and finding the optimal tree is a problem that a heuristic algorithm can solve in a second.

## Gaps

Pairwise alignment is useful as a way to identify mutations that have occurred during evolution and have caused divergence of the sequences of the two proteins we are studying. The most common mutations are *substitutions*, *insertions*, and *deletions*. In protein sequences, substitutions occur when a mutation results in the codon for one amino acid being changed into that for another. This results in the alignment of two nonidentical amino acids, such as serine and threonine. Insertions and deletions occur when residues are added or removed and are typically represented by dashes that are added to one or the other sequence. Insertions or deletions (even those just one character long) are referred to as *gaps* in the alignment.

In our alignment of human beta globin and myoglobin there is one gap (Fig. 3.5a, between the D and E residues of the query). Gaps can occur at the ends of the proteins or in the middle. Note that one of the effects of adding gaps is to make the overall length of each alignment exactly the same. The addition of gaps can help to create an alignment that models evolutionary changes that have occurred. In a typical scoring scheme there are two gap penalties: one for creating a gap (−11 in the example of Fig. 3.5b) and one for each additional residue that a gap extends (−1 in Fig. 3.5b).

## Pairwise Alignment, Homology, and Evolution of Life

If two proteins are homologous, they share a common ancestor. Generally, we observe the sequence of proteins (and genes) from organisms that are extant. We

It is possible to infer the sequence of the common ancestor (see Chapter 7).

can compare myoglobins from species such as human, horse, and chicken, and see that the sequences are homologous (Fig. 3.2). This implies that an ancestral organism had a myoglobin gene and lived sometime before the divergences of the lineages that gave rise to human and chicken (over 300 MYA; see Chapter 18). Descendants of that ancestral organism include many vertebrate species. The study of homologous protein (or DNA) sequences by pairwise alignment involves an investigation of the evolutionary history of that protein (or gene).

For a brief overview of the time scale of life on Earth, see Fig. 3.6 (refer to Chapter 13 for a more detailed discussion). The divergence of different species is established through the use of many sources of data, especially the fossil record. Fossils of prokaryotes have been discovered in rocks 3.5 billion years old or even older (Schopf, 2002). Fossils of methane-producing archaea, representative of a second domain of life, are found in rocks over 3 billion years old. The other main domain of life, the eukaryotes, emerged soon after. In the case of globins, in addition to the vertebrate proteins represented in Fig. 3.2, there are plant globins that must have shared a common ancestor with the metazoan (animal) globins some 1.5 billion years ago. There are also many bacterial and archaeal globins suggesting that the globin family arose earlier than two billion years ago.

Databases such as Pfam (Chapter 6) and COGS (Chapter 15) summarize the phylogenetic distribution of gene/protein families across the tree of life.

As we examine a variety of homologous protein sequences, we can observe a wide range of conservation between family members. Some are very ancient and well conserved, such as the enzyme glyceraldehyde-3-phosphate dehydrogenase (GAPDH). A multiple sequence alignment, which is essentially a series of pairwise alignments between a group of proteins, reveals that GAPDH orthologs are extraordinarily well conserved (Fig. 3.7). Such highly conserved proteins may have any degree of representation across the tree of life, from being present in most known species to only a select few.

The GAPDH sequences used to generate Fig. 3.7 and the kappa casein sequences used to generate fig. 3.8 are shown in web documents 3.3 and 3.4 at ▶ http://www.bioinfbook.org/chapter3.

Orthologous kappa caseins from various species provide an example of a less well-conserved family (Fig. 3.8). Some columns of residues in this alignment are perfectly conserved among the selected species, but most are not, and many gaps needed to be introduced. Several positions at which four or even five different residues occur in an aligned column are indicated.

We can see from the preceding examples that pairwise sequence alignment between any two proteins can exhibit widely varying amounts of conservation. We will next examine how the information in such alignments can be used to decide how to quantitate the relatedness of any two proteins.

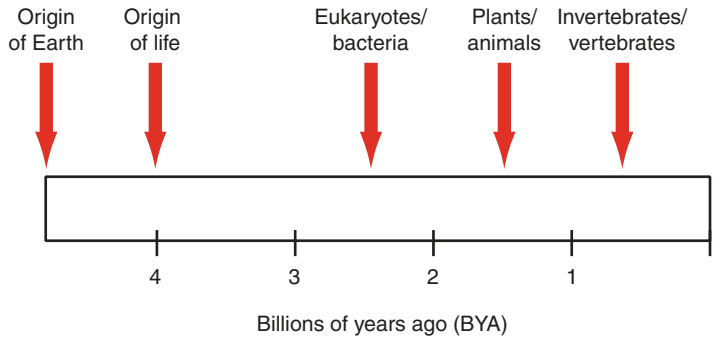



*FIGURE 3.6. Overview of the history of life on Earth. See Chapter 13 for details. Gene/protein sequences are analyzed in the context of evolution: Which organisms have orthologous genes? When did these organisms evolve? How related are human and bacterial globins?*

```
NP_002037.2          164   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGALQNII   207
XP_001162057.1       164   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGALQNII   207
NP_001003142.1       162   IHDHFGIVEGLMTTVHAITATQKTVDGPSGKMWRDGRGAAQNII   205
XP_893121.1          168   IHDNFGIMEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII   211
XP_576394.1          162   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII   205
NP_058704.1          162   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII   205
XP_001070653.1       162   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII   205
XP_001062726.1       162   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII   205
NP_989636.1          162   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII   205
NP_525091.1          161   INDNFEIVEGLMTTVHATTATQKTVDGPSGKLWRDGRGAAQNII   204
XP_318655.2          161   INDNFGILEGLMTTVHATTATQKTVDGPSGKLWRDGRGAAQNII   204
NP_508535.1          170   INDNFGIIEGLMTTVHAVTATQKTVDGPSGKLWRDGRGAGQNII   213
NP_595236.1          164   INDTFGIEEGLMTTVHATTATQKTVDGPSKKDWRGGRGASANII   207
NP_011708.1          162   INDAFGIEEGLMTTVHSLTATQKTVDGPSHKDWRGGRTASGNII   205
XP_456022.1          161   INDEFGIDEALMTTVHSITATQKTVDGPSHKDWRGGRTASGNII   204
NP_001060897.1       166   IHDNFGIIEGLMTTVHAITATQKTVDGPSSKDWRGGRAASFNII   209
```

*FIGURE 3.7. Multiple sequence alignment of a portion of the glyceraldehyde 3-phosphate dehydrogenase (GAPDH) protein from thirteen organisms:* Homo sapiens *(human),* Pan troglodytes *(chimpanzee),* Canis lupus *(dog),* Mus musculus *(mouse),* Rattus norvegicus *(rat; four variants),* Gallus gallus *(chicken),* Drosophila melanogaster *(fruit fly),* Anopheles gambiae *(mosquito),* Caenorhabditis elegans *(worm),* Schizosaccharomyces pombe *(fission yeast),* Saccharomyces cerevisiae *(baker's yeast),* Kluyveromyces lactis *(a fungus), and* Oryza sativa *(rice). Columns in the alignment having even a single amino acid change are indicated with arrowheads. The accession numbers are given in the figure. The alignment was created by searching HomoloGene at NCBI with the term gapdh. The full alignment is given in Web Document 3.3 at ► http://www.bioinfbook.org/chapter3.*

```
mouse     AIPNPSFLAMPTNENQDNTAIPTIDPITPIVST--PVPTM------ESIVNTVANPEAST
rabbit    S--HPFFMAILPNKMQDKAVTPTTNTIAAVEPT--PIPTT------EPVVSTEVIAEASP
sheep     PHPHLSFMAIPPKKDQDKTEIPAINTIASAEPTVHSTPTT------EAVVNAVDNPEASS
cattle    PHPHLSFMAIPPKKNQDKTEIPTINTIASGEPT--STPTT------EAVESTVATLEDSP
pig       PRPHASFIAIPPKKNQDKTAIPAINSIATVEPT--IVPATEPIVNAEPIVNAVVTPEASS
human     PNLHPSFIAIPPKKIQDKIIIPTINTIATVEPT--PAPAT------EPTVDSVVTPEAFS
horse     PCPHPSFIAIPPKKLQEITVIPKINTIATVEPT--PIPTP------EPTVNNAVIPDASS
          .  :  *:*: .:: *:    *  :.*:.  .*    *:       *.  .     :  .
```

*FIGURE 3.8. Multiple sequence alignment of seven kappa caseins, representing a protein family that is relatively poorly conserved. Only a portion of the entire alignment is shown. Note that just eight columns of residues are perfectly conserved (indicated with asterisks), and gaps of varying length form part of the alignment. In several columns, there are four different aligned amino acids (arrowheads); in two instances there are five different residues (double arrowheads). The sequences were aligned with MUSCLE 3.6 (see Chapter 6) and were human (NP_005203), equine (Equus caballus; NP_001075353), pig (Sus scrofa NP_001004026), ovine (Ovis aries NP_001009378), rabbit (Oryctolagus cuniculus P33618), bovine (Bos taurus NP_776719), and mouse (Mus musculus NP_031812). The full alignment is available as web document 3.3 at ► http://www.bioinfbook.org/chapter3.*

## Scoring Matrices

When two proteins are aligned, what scores should they be assigned? For the alignment of beta globin and myoglobin in Fig. 3.5a there were specific scores for matches and mismatches; how were they derived? Margaret Dayhoff (1978) provided a model of the rules by which evolutionary change occurs in proteins. We will now examine the Dayhoff model, which provides the basis of a quantitative scoring system for pairwise alignments. This system accounts for scores between any proteins, whether they are closely or distantly related. We will then describe the BLOSUM matrices of

Steven Henikoff and Jorja G. Henikoff (1992). Next, we will discuss the two main kinds of pairwise sequence algorithms, global and local. Many database searching methods such as BLAST (Chapters 4 and 5) depend in some form on the evolutionary insights of the Dayhoff model.

The Dayhoff (1978) reference is to the *Atlas of Protein Sequence and Structure*, a book with 25 chapters (and various coauthors) describing protein families. The 1966 version of the *Atlas* described the sequences of just several dozen proteins (cytochromes c, other respiratory proteins, globins, some enzymes such as lysozyme and ribonucleases, virus coat proteins, peptide hormones, kinins, and fibrinopeptides). The 1978 edition included about 800 protein sequences.

## Dayhoff Model: Accepted Point Mutations

Dayhoff and colleagues considered the problem of how to assign scores to aligned amino acid residues. Their approach was to catalog thousands of proteins and compare the sequences of closely related proteins in many families. They considered the question of which specific amino acid substitutions are observed to occur when two homologous protein sequences are aligned. They defined an *accepted point mutation* as a replacement of one amino acid in a protein by another residue that has been accepted by natural selection. Accepted point mutation is abbreviated *PAM* (which is easier to pronounce than APM). An amino acid change that is accepted by natural selection occurs when (1) a gene undergoes a DNA mutation such that it encodes a different amino acid and (2) the entire species adopts that change as the predominant form of the protein.

Which point mutations are accepted in protein evolution? Intuitively, conservative replacements such as serine for threonine would be most readily accepted. In order to determine all possible changes, Dayhoff and colleagues examined 1572 changes in 71 groups of closely related proteins (Box 3.3). Thus, their definition of "accepted" mutations was based on empirically observed amino acid substitutions. Their approach involved a phylogenetic analysis: rather than comparing two amino acid residues directly, they compared them to the inferred common ancestor of those sequences (Fig. 3.9 and Box 3.4).

Dayhoff et al. (1972) focused on proteins sharing 85% or more identity. Thus, they could construct their alignments with a high degree of confidence. Later in this chapter, we will see how the Needleman and Wunsch algorithm (described in 1970) permits the optimal alignment of protein sequences.

For the PAM1 matrix, the proteins have undergone 1% change (that is, 1 accepted point mutation per 100 amino acid residues). The results are shown in Fig. 3.10, which describes the frequency with which any amino acid pairs $i, j$ are aligned. Inspection of this table reveals which substitutions are unlikely to occur (for example, cysteine and tryptophan have noticeably few substitutions), while others such as asparagine and serine tolerate replacements quite commonly. Today, we could generate a table like this with vastly more data (refer to Fig. 2.1 and the explosive growth of GenBank). Several groups have produced updated versions of the PAM matrices (Gonnet et al., 1992; Jones et al., 1992). Nonetheless the findings from 1978 are essentially correct.

The main goal of Dayhoff's approach was to define a set of scores for the comparison of aligned amino acid residues. By comparing two aligned proteins, one can then tabulate an overall score, taking into account identities as well as mismatches, and also applying appropriate penalties for gaps. A scoring matrix defines scores for the interchange of residues $i$ and $j$. It is given by the probability $q_{i,j}$ of aligning original amino acid residue $j$ with replacement residue $i$ relative to the likelihood of observing residues $i$ by chance ($p_i$). The scoring matrix further incorporates a logarithm to generate log-odds scores. For the Dayhoff matrices, this takes the following form:

$$s_{i,j} = 10 \times \log\left(\frac{q_{i,j}}{p_i}\right) \tag{3.1}$$

Here the score $s_{i,j}$ refers to the score for aligning any two residues (including an amino acid with itself) along the length of a pairwise alignment. The probability $q_{i,j}$ is the

## Box 3.3
## Dayhoff's Protein Superfamilies

Dayhoff (1978, p. 3) and colleagues studied 34 protein "superfamilies" grouped into 71 phylogenetic trees. These proteins ranged from some that are very well conserved (e.g., histones and glutamate dehydrogenase; see Fig. 3.7) to others that have a high rate of mutation acceptance (e.g., immunoglobulin [Ig] chains and kappa casein; see Fig. 3.8). Protein families were aligned (compare Fig. 3.7); then they counted how often any one amino acid in the alignment was replaced by another. Here is a partial list of the proteins they studied, including the rates of mutation acceptance. For a more detailed list, see Table 11.1. There is a range of almost 400-fold between the families that evolve fastest and slowest, but within a given family the rate of evolution (measured in PAMs per unit time) varies only two- to threefold between species. Used with permission.

| Protein | PAMs per 100 million years |
|---|---|
| Immunoglobulin (Ig) kappa chain C region | 37 |
| Kappa casein | 33 |
| Epidermal growth factor | 26 |
| Serum albumin | 19 |
| Hemoglobin alpha chain | 12 |
| Myoglobin | 8.9 |
| Nerve growth factor | 8.5 |
| Trypsin | 5.9 |
| Insulin | 4.4 |
| Cytochrome *c* | 2.2 |
| Glutamate dehydrogenase | 0.9 |
| Histone H3 | 0.14 |
| Histone H4 | 0.10 |

observed frequency of substitution for each pair of amino acids. The values for $q_{ij}$ are called the "target frequencies," and they are estimated in reference to a particular amount of evolutionary change. For example, in a comparison of human beta globin versus the closely related chimpanzee beta globin, the likelihood of any particular residue matching another in a pairwise alignment is extremely high, while in a comparison of human beta globin and a bacterial globin the likelihood of a match is low. If in a particular comparison of closely related proteins an aligned serine were to change to a threonine 5% of the time, then that target frequency $q_{S,T}$ would be 0.05. If in a different comparison of differently related proteins serine were to change to threonine more often, say 40% of the time, then that target frequency $q_{S,T}$ would be 0.4.

Equation 3.1 describes an odds ratio (Box 3.5). For the numerator, Dayhoff et al. (1972) considered an entire spectrum of models for evolutionary change in determining target frequencies. We begin with the PAM1 matrix, which describes substitutions that occur in very closely related proteins. For the denominator of Equation 3.1, $p_i p_j$, is the probability of amino acid residues $i$ and $j$ occurring by chance. We will


*FIGURE 3.9. Dayhoff's approach to determining amino acid substitutions. Panel (a) shows a partial multiple sequence alignment of human alpha-1 globin, beta globin, delta globin, and myoglobin. Four columns in which alpha-1 globin and myoglobin have different amino acid residues are indicated in red. For example, A is aligned with G (arrow). Panel (b) shows a phylogenetic tree that shows the four extant sequences (labeled 1 to 4), as well as two internal nodes that represent the ancestral sequences (labeled 5 and 6). The inferred ancestral sequences were identified by maximum parsimony analysis using the software PAUP (Chapter 7), and are displayed in panel (a). From this analysis it is apparent that at each of the columns labeled in red, there was not a direct interchange of two amino acids between alpha-1 globin and myoglobin. Instead, an ancestral residue diverged. For example, the arrow in panel (a) indicates an ancestral glutamate that evolved to become alanine or glycine, but it would not be correct to suggest that alanine had been converted directly to glycine.*

## Box 3.4
## A Phylogenetic Approach to Aligning Amino Acids

Dayhoff and colleagues did not compare the probability of one residue mutating directly into another. Instead, they constructed phylogenetic trees using parsimony analysis (see Chapter 7). Then, they described the probability that two aligned residues derived from a common ancestral residue. With this approach, they could minimize the confounding effects of multiple substitutions occurring in an aligned pair of residues. As an example, consider an alignment of the four human proteins alpha-1 globin, beta globin, delta globin, and myoglobin. A direct comparison of alpha-1 globin to myoglobin would suggest several amino acid replacements, such as ala ↔ gly, asn ↔ leu, lys ↔ leu, and ala ↔ val (Fig. 3.9a, residues highlighted in red). However, a phylogenetic analysis of these four proteins results in the estimation of internal nodes that represent ancestral sequences. In Fig. 3.9b the external nodes (corresponding to the four existing proteins) are labeled, as are internal nodes 5 and 6, which correspond to inferred ancestral sequences. In one of the four cases that are highlighted in Fig. 3.9a, the ancestral sequences suggest that a glu residue changed to ala and gly in alpha-1 globin and myoglobin, but ala and gly never directly interchanged (Fig. 3.9a, arrow). Thus, the Dayhoff approach was more accurate by taking an evolutionary perspective.

In a further effort to avoid the complicating factor of multiple substitutions occurring in alignments of protein families, Dayhoff et al. also focused on using multiple sequence alignments of closely related proteins. Thus, for example, their analysis of globins considered the alpha globins and beta globins separately.

| | A Ala | R Arg | N Asn | D Asp | C Cys | Q Gln | E Glu | G Gly | H His | I Ile | L Leu | K Lys | M Met | F Phe | P Pro | S Ser | T Thr | W Trp | Y Tyr | V Val |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | | | | | | | | | | | | | | | | | | | | |
| R | 30 | | | | | | | | | | | | | | | | | | | |
| N | 109 | 17 | | | | | | | | | | | | | | | | | | |
| D | 154 | 0 | 532 | | | | | | | | | | | | | | | | | |
| C | 33 | 10 | 0 | 0 | | | | | | | | | | | | | | | | |
| Q | 93 | 120 | 50 | 76 | 0 | | | | | | | | | | | | | | | |
| E | 266 | 0 | 94 | 831 | 0 | 422 | | | | | | | | | | | | | | |
| G | 579 | 10 | 156 | 162 | 10 | 30 | 112 | | | | | | | | | | | | | |
| H | 21 | 103 | 226 | 43 | 10 | 243 | 23 | 10 | | | | | | | | | | | | |
| I | 66 | 30 | 36 | 13 | 17 | 8 | 35 | 0 | 3 | | | | | | | | | | | |
| L | 95 | 17 | 37 | 0 | 0 | 75 | 15 | 17 | 40 | 253 | | | | | | | | | | |
| K | 57 | 477 | 322 | 85 | 0 | 147 | 104 | 60 | 23 | 43 | 39 | | | | | | | | | |
| M | 29 | 17 | 0 | 0 | 0 | 20 | 7 | 7 | 0 | 57 | 207 | 90 | | | | | | | | |
| F | 20 | 7 | 7 | 0 | 0 | 0 | 0 | 17 | 20 | 90 | 167 | 0 | 17 | | | | | | | |
| P | 345 | 67 | 27 | 10 | 10 | 93 | 40 | 49 | 50 | 7 | 43 | 43 | 4 | 7 | | | | | | |
| S | 772 | 137 | 432 | 98 | 117 | 47 | 86 | 450 | 26 | 20 | 32 | 168 | 20 | 40 | 269 | | | | | |
| T | 590 | 20 | 169 | 57 | 10 | 37 | 31 | 50 | 14 | 129 | 52 | 200 | 28 | 10 | 73 | 696 | | | | |
| W | 0 | 27 | 3 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 13 | 0 | 0 | 10 | 0 | 17 | 0 | | | |
| Y | 20 | 3 | 36 | 0 | 30 | 0 | 10 | 0 | 40 | 13 | 23 | 10 | 0 | 260 | 0 | 22 | 23 | 6 | | |
| V | 365 | 20 | 13 | 17 | 33 | 27 | 37 | 97 | 30 | 661 | 303 | 17 | 77 | 10 | 50 | 43 | 186 | 0 | 17 | |
| | A Ala | R Arg | N Asn | D Asp | C Cys | Q Gln | E Glu | G Gly | H His | I Ile | L Leu | K Lys | M Met | F Phe | P Pro | S Ser | T Thr | W Trp | Y Tyr | V Val |

*FIGURE 3.10. Numbers of accepted point mutations, multiplied by 10, in 1572 cases of amino acid substitutions from closely related protein sequences. This figure is modified from Dayhoff (1978, p. 346). Amino acids are presented alphabetically according to the three-letter code. Notice that some substitutions are very commonly accepted (such as V and I or S and T). Other amino acids, such as C and W, are rarely substituted by any other residue. Used with permission.*

## Box 3.5
## Statistical Concept: The Odds Ratio

Dayhoff et al. (1972) developed their scoring matrix by using odds ratios. The mutation probability matrix has elements $M_{ij}$ that give the probability that amino acid $j$ changes to amino acid $i$ in a given evolutionary interval. The normalized frequency $f_i$ gives the probability that amino acid $i$ will occur at that given amino acid position by chance. The relatedness odds matrix in Equation 3.1 may also be expressed as follows:

$$R_{ij} = \frac{M_{ij}}{f_i}$$

Here, $R_{ij}$ is the relatedness odds ratio. Equation 3.1 may also be represented:

$$\text{Probability of an authentic alignment} = \frac{p(\text{aligned} \mid \text{authentic})}{p(\text{aligned} \mid \text{random})}$$

The right side of this equation can be read, "the probability of an alignment given that it is authentic (i.e. the substitution of amino acid $j$ with amino acid $i$) divided by the probability that the alignment occurs given that it happened by chance. An *odds ratio* can be any positive ratio. The *probability* that an event will occur is the fraction of times it is expected to be observed over many trials; probabilities have values ranging from 0 to 1. Odds and probability are closely related concepts. A probability of 0 corresponds to an odds of 0; a probability of 0.5 corresponds to an odds of 1.0; a probability of 0.75 corresponds to odds of 75:25 or 3. Odds and probabilities may be converted as follows:

$$odds = \frac{probability}{1 - probability} \quad and \quad probability = \frac{odds}{1 + odds}$$

next explain how they calculated these values, resulting in the creation of an entire series of scoring matrices.

Dayhoff et al. calculated the relative mutabilities of the amino acids (Table 3.1). This simply describes how often each amino acid is likely to change over a short evolutionary period. (We note that the evolutionary period in question is short because this analysis involves protein sequences that are closely related to each other.) To calculate relative mutability, they divided the number of times each amino acid was observed to mutate by the overall frequency of occurrence of that amino acid. Table 3.2 shows the frequency with which each amino acid is found.

Why are some amino acids more mutable than others? The less mutable residues probably have important structural or functional roles in proteins, such that the consequence of replacing them with any other residue could be harmful to the organism. (We will see in Chapter 20 that many human diseases, from cystic fibrosis to the autism-related Rett syndrome to hemoglobinopathies, can be caused by a single amino acid substitution in a protein.) Conversely, the most mutable amino acids—asparagine, serine, aspartic acid, and glutamic acid—have functions in proteins that are easily assumed by other residues. The most common substitutions seen in Fig. 3.10 are glutamic acid for aspartic acid (both are acidic), serine for

You can look up a recent estimate of the frequency of occurrence of each amino acid at the SwissProt website ► http://www.expasy.ch/sprot/relnotes/relstat.html. From the UniProtKB/Swiss-Prot protein knowledgebase (release 51.7), the amino acid composition of all proteins is shown in web document 3.5 (► http://www.bioinfbook.org/chapter3).

**TABLE 3-1** Relative Mutabilities of Amino Acids

| | | | |
|---|---|---|---|
| Asn | 134 | His | 66 |
| Ser | 120 | Arg | 65 |
| Asp | 106 | Lys | 56 |
| Glu | 102 | Pro | 56 |
| Ala | 100 | Gly | 49 |
| Thr | 97 | Tyr | 41 |
| Ile | 96 | Phe | 41 |
| Met | 94 | Leu | 40 |
| Gln | 93 | Cys | 20 |
| Val | 74 | Trp | 18 |

The value of alanine is arbitrarily set to 100.
*Source*: From Dayhoff (1978). Used with permission.

**TABLE 3-2** Normalized Frequencies of Amino Acid

| | | | |
|---|---|---|---|
| Gly | 0.089 | Arg | 0.041 |
| Ala | 0.087 | Asn | 0.040 |
| Leu | 0.085 | Phe | 0.040 |
| Lys | 0.081 | Gln | 0.038 |
| Ser | 0.070 | Ile | 0.037 |
| Val | 0.065 | His | 0.034 |
| Thr | 0.058 | Cys | 0.033 |
| Pro | 0.051 | Tyr | 0.030 |
| Glu | 0.050 | Met | 0.015 |
| Asp | 0.047 | Trp | 0.010 |

These values sum to 1. If the 20 amino acids were equally represented in proteins, these values would all be 0.05 (i.e., 5%); instead, amino acids vary in their frequency of occurrence
*Source*: From Dayhoff (1978). Used with permission.

alanine, serine for threonine (both are hydroxylated), and isoleucine for valine (both are hydrophobic and of a similar size).

The substitutions that occur in proteins can also be understood with reference to the genetic code (Box 3.6). Observe how common amino acid substitutions tend to require only a single nucleotide change. For example, aspartic acid is encoded by GAU or GAC, and changing the third position to either A or G causes the codon to encode a glutamic acid. Also note that four of the five least mutable amino acids (tryptophan, cysteine, phenylalanine, and tyrosine) are specified by only one or two codons. A mutation of any of the three bases of the W codon is guaranteed to change that amino acid. The low mutability of this amino acid suggests that substitutions are not tolerated by natural selection. Of the eight least mutable amino acids (Table 3.1), only one (leucine) is specified by six codons, and only two (glycine and proline) are specified by four codons. The others are specified by one or two codons.

## PAM1 Matrix

Dayhoff and colleagues next used the data on accepted mutations (Fig. 3.10) and the probabilities of occurrence of each amino acid to generate a *mutation probability*

## Box 3.6
## The Standard Genetic Code

| First nucleotide | Second nucleotide: T | Second nucleotide: C | Second nucleotide: A | Second nucleotide: G | Third nucleotide |
|---|---|---|---|---|---|
| T | TTT Phe 171 | TCT Ser 147 | TAT Tyr 124 | TGT Cys 99 | T |
| | TTC Phe 203 | TCC Ser 172 | TAC Tyr 158 | TGC Cys 119 | C |
| | TTA Leu 73 | TCA Ser 118 | TAA Ter 0 | TGA Ter 0 | A |
| | TTG Leu 125 | TCG Ser 45 | TAG Ter 0 | TGG Trp 122 | G |
| C | CTT Leu 127 | CCT Pro 175 | CAT His 104 | CGT Arg 47 | T |
| | CTC Leu 187 | CCC Pro 197 | CAC His 147 | CGC Arg 107 | C |
| | CTA Leu 69 | CCA Pro 170 | CAA Gln 121 | CGA Arg 63 | A |
| | CTG Leu 392 | CCG Pro 69 | CAG Gln 343 | CGG Arg 115 | G |
| A | ATT Ile 165 | ACT Thr 131 | AAT Asn 174 | AGT Ser 121 | T |
| | ATC Ile 218 | ACC Thr 192 | AAC Asn 199 | AGC Ser 191 | C |
| | ATA Ile 71 | ACA Thr 150 | AAA Lys 248 | AGA Arg 113 | A |
| | ATG Met 221 | ACG Thr 63 | AAG Lys 331 | AGG Arg 110 | G |
| G | GTT Val 111 | GCT Ala 185 | GAT Asp 230 | GGT Gly 112 | T |
| | GTC Val 146 | GCC Ala 282 | GAC Asp 262 | GGC Gly 230 | C |
| | GTA Val 72 | GCA Ala 160 | GAA Glu 301 | GGA Gly 168 | A |
| | GTG Val 288 | GCG Ala 74 | GAG Glu 404 | GGG Gly 160 | G |

In this table, the 64 possible codons are depicted along with the frequency of codon utilization and the three-letter code of the amino acid that is specified. There are four bases (A, C, G, U) and three bases per codon, so there are $4^3 = 64$ codons.

Several features of the genetic code should be noted. Amino acids may be specified by one codon (M, W), two codons (C, D, E, F, H, K, N, Q, Y), three codons (I), four codons (A, G, P, T, V), or six codons (L, R, S). UGA is rarely read as a selenocysteine (abbreviated sec, and the assigned single-letter abbreviations is U).

For each block of four codons that are grouped together, one is often used dramatically less frequently. For example, for F, L, I, M, and V (i.e., codons with a U in the middle, occupying the first column of the genetic code), adenine is used relatively infrequently in the third-codon position. For codons with a cytosine in the middle position, guanine is strongly underrepresented in the third position.

Also note that in many cases mutations cause a conservative change (or no change at all) in the amino acid. Consider threonine (ACX). Any mutation in the third position causes no change in the specified amino acid, because of "wobble." If the first nucleotide of any threonine codon is mutated from A to U, the conservative replacement to a serine occurs. If the second nucleotide C is mutated to a G, a serine replacement occurs. Similar patterns of conservative substitution can be seen along the entire first column of the genetic code, where all of the residues are hydrophobic, and for the charged residues D, E and K, R as well.

Codon usage varies between organisms and between genes within organisms. Note also that while this is the standard genetic code, some organisms use

alternate genetic codes. A group of two dozen alternate genetic codes are listed at the NCBI Taxonomy website, ► http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/. As an example of a nonstandard code, vertebrate mitochondrial genomes use AGA and AGG to specify termination (rather than arg in the standard code), ATA to specify met (rather than ile), and TGA to specify trp (rather then termination).

*Source*: Adapted from the International Human Genome Sequencing Consortium (2001), fig. 34. Used with permission.

*matrix M* (Fig. 3.11). Each element of the matrix $M_{ij}$ shows the probability that an original amino acid $j$ (see the columns) will be replaced by another amino acid $i$ (see the rows) over a defined evolutionary interval. In the case of Fig. 3.11 the interval is one PAM, which is defined as the unit of evolutionary divergence in which 1% of the amino acids have been changed between the two protein sequences. Note that the evolutionary interval of this PAM matrix is defined in terms of percent amino acid divergence and not in units of years. One percent divergence of protein sequence may occur over vastly different time frames for protein families that undergo substitutions at different rates.

Examination of Fig. 3.11 reveals several important features. The highest scores are distributed in a diagonal from top left to bottom right. The values in each column sum to 100%. The value 98.67 at the top left indicates that when the original sequence consists of an alanine there is a 98.67% chance that the replacement amino acid will also be an alanine over an evolutionary distance of one PAM. There is a 0.28% chance that it will be changed to serine. The most mutable amino acid (from Table 3.1), asparagine, has only a 98.22% chance of remaining unchanged; the least mutable amino acid, tryptophan, has a 99.76% chance of remaining the same.

For each original amino acid, it is easy to observe the amino acids that are most likely to replace it if a change should occur. These data are very relevant to pairwise sequence alignment because they will form the basis of a scoring system (described below) in which reasonable amino acid substitutions in an alignment are rewarded while unlikely substitutions are penalized. These concepts are also relevant to database searching algorithms such as BLAST (Chapters 4 and 5) which depend on rules to score the relatedness of molecular sequences.

Almost all molecular sequence data are obtained from extant organisms. We can infer ancestral sequences, as described in Box 3.4 and Chapter 7. But in general, for an aligned pair of residues $i, j$ we do not know which one mutated into the other. Dayhoff and colleagues used the assumption that accepted amino acid mutations are undirected, that is, they are equally likely in either direction. In the PAM1 matrix, the close relationship of the proteins makes it unlikely that the ancestral residue is entirely different than both of the observed, aligned residues.

## PAM250 and Other PAM Matrices

The PAM1 matrix was based on the alignment of closely related protein sequences, all of which were at least 85% identical within a protein family. We are often interested in exploring the relationships of proteins that share far less than 85% amino acid identity. We can accomplish this by constructing probability matrices for proteins that share any degree of amino acid identity. Consider closely related proteins, such as the GAPDH proteins shown in Fig. 3.7. A mutation from one residue to another

| | A Ala | R Arg | N Asn | D Asp | C Cys | Q Gln | E Glu | G Gly | H His | I Ile | L Leu | K Lys | M Met | F Phe | P Pro | S Ser | T Thr | W Trp | Y Tyr | V Val |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 98.67 | 0.02 | 0.09 | 0.10 | 0.03 | 0.08 | 0.17 | 0.21 | 0.02 | 0.06 | 0.04 | 0.02 | 0.06 | 0.02 | 0.22 | 0.35 | 0.32 | 0.00 | 0.02 | 0.18 |
| R | 0.01 | 99.13 | 0.01 | 0.00 | 0.01 | 0.10 | 0.00 | 0.00 | 0.10 | 0.03 | 0.01 | 0.19 | 0.04 | 0.01 | 0.04 | 0.06 | 0.01 | 0.08 | 0.00 | 0.01 |
| N | 0.04 | 0.01 | 98.22 | 0.36 | 0.00 | 0.04 | 0.06 | 0.06 | 0.21 | 0.03 | 0.01 | 0.13 | 0.00 | 0.01 | 0.02 | 0.20 | 0.09 | 0.01 | 0.04 | 0.01 |
| D | 0.06 | 0.00 | 0.42 | 98.59 | 0.00 | 0.06 | 0.53 | 0.06 | 0.04 | 0.01 | 0.00 | 0.03 | 0.00 | 0.00 | 0.01 | 0.05 | 0.03 | 0.00 | 0.00 | 0.01 |
| C | 0.01 | 0.01 | 0.00 | 0.00 | 99.73 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.05 | 0.01 | 0.00 | 0.03 | 0.02 |
| Q | 0.03 | 0.09 | 0.04 | 0.05 | 0.00 | 98.76 | 0.27 | 0.01 | 0.23 | 0.01 | 0.03 | 0.06 | 0.04 | 0.00 | 0.06 | 0.02 | 0.02 | 0.00 | 0.00 | 0.01 |
| E | 0.10 | 0.00 | 0.07 | 0.56 | 0.00 | 0.35 | 98.65 | 0.04 | 0.02 | 0.03 | 0.01 | 0.04 | 0.01 | 0.00 | 0.03 | 0.04 | 0.02 | 0.00 | 0.01 | 0.02 |
| G | 0.21 | 0.01 | 0.12 | 0.11 | 0.01 | 0.03 | 0.07 | 99.35 | 0.01 | 0.00 | 0.01 | 0.02 | 0.01 | 0.01 | 0.03 | 0.21 | 0.03 | 0.00 | 0.00 | 0.05 |
| H | 0.01 | 0.08 | 0.18 | 0.03 | 0.01 | 0.20 | 0.01 | 0.00 | 99.12 | 0.00 | 0.01 | 0.01 | 0.00 | 0.02 | 0.03 | 0.01 | 0.01 | 0.01 | 0.04 | 0.01 |
| I | 0.02 | 0.02 | 0.03 | 0.01 | 0.02 | 0.01 | 0.02 | 0.00 | 0.00 | 98.72 | 0.09 | 0.02 | 0.21 | 0.07 | 0.00 | 0.01 | 0.07 | 0.00 | 0.01 | 0.33 |
| L | 0.03 | 0.01 | 0.03 | 0.00 | 0.00 | 0.06 | 0.01 | 0.01 | 0.04 | 0.22 | 99.47 | 0.02 | 0.45 | 0.13 | 0.03 | 0.01 | 0.03 | 0.04 | 0.02 | 0.15 |
| K | 0.02 | 0.37 | 0.25 | 0.06 | 0.00 | 0.12 | 0.07 | 0.02 | 0.02 | 0.04 | 0.01 | 99.26 | 0.20 | 0.00 | 0.03 | 0.08 | 0.11 | 0.00 | 0.01 | 0.01 |
| M | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 | 0.05 | 0.08 | 0.04 | 98.74 | 0.01 | 0.00 | 0.01 | 0.02 | 0.00 | 0.00 | 0.04 |
| F | 0.01 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.02 | 0.08 | 0.06 | 0.00 | 0.04 | 99.46 | 0.00 | 0.02 | 0.01 | 0.03 | 0.28 | 0.00 |
| P | 0.13 | 0.05 | 0.02 | 0.01 | 0.01 | 0.08 | 0.03 | 0.02 | 0.05 | 0.01 | 0.02 | 0.02 | 0.01 | 0.01 | 99.26 | 0.12 | 0.04 | 0.00 | 0.00 | 0.02 |
| S | 0.28 | 0.11 | 0.34 | 0.07 | 0.11 | 0.04 | 0.06 | 0.16 | 0.02 | 0.02 | 0.01 | 0.07 | 0.04 | 0.03 | 0.17 | 98.40 | 0.38 | 0.05 | 0.02 | 0.02 |
| T | 0.22 | 0.02 | 0.13 | 0.04 | 0.01 | 0.03 | 0.02 | 0.02 | 0.01 | 0.11 | 0.02 | 0.08 | 0.06 | 0.01 | 0.05 | 0.32 | 98.71 | 0.00 | 0.02 | 0.09 |
| W | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.01 | 0.00 | 99.76 | 0.01 | 0.00 |
| Y | 0.01 | 0.00 | 0.03 | 0.00 | 0.03 | 0.00 | 0.01 | 0.00 | 0.04 | 0.01 | 0.01 | 0.00 | 0.00 | 0.21 | 0.00 | 0.01 | 0.01 | 0.02 | 99.45 | 0.01 |
| V | 0.13 | 0.02 | 0.01 | 0.01 | 0.03 | 0.02 | 0.02 | 0.03 | 0.03 | 0.57 | 0.11 | 0.01 | 0.17 | 0.01 | 0.03 | 0.02 | 0.10 | 0.00 | 0.02 | 99.01 |

*FIGURE 3.11. The PAM1 mutation probability matrix. From Dayhoff (1978, p. 348, fig. 82). The original amino acid j is arranged in columns (across the top), while the replacement amino acid i is arranged in rows. Used with permission.*

is a relatively rare event, and a scoring system used to align two such closely related proteins should reflect this. In the PAM1 mutation probability matrix (Fig. 3.11) some substitutions such as tryptophan to threonine are so rare that they are never observed in the data set. But next consider two distantly related proteins, such as the kappa caseins shown in Fig. 3.8. Here, substitutions are likely to be very common. PAM matrices such as PAM100 or PAM250 were generated to reflect the kinds of amino acid substitutions that occur in distantly related proteins.

How are PAM matrices other than PAM1 derived? Dayhoff et al. multiplied the PAM1 matrix by itself, up to hundreds of times, to obtain other PAM matrices (see Box 3.7). Thus they extrapolated from the PAM1 matrix.

To make sense of what different PAM matrices mean, consider the extreme cases. When PAM equals zero, the matrix is a unit diagonal (Fig. 3.12), because no amino acids have changed. PAM can be extremely large (e.g., PAM greater than 2000, or the matrix can even be multiplied against itself an infinite number of times). In the resulting PAM∞ matrix there is an equal likelihood of any amino acid being present and all the values consist of rows of probabilities that approximate the background probability for the frequency of occurrence of each amino acid (Fig. 3.12, lower panel). We described these background frequencies in Table 3.2.

The PAM250 matrix is of particular interest (Fig. 3.13). It is produced when the PAM1 matrix is multiplied against itself 250 times, and it is one of the common matrices used for BLAST searches of databases (Chapter 4). This matrix applies

## Box 3.7
## Matrix Multiplication

A matrix is an orderly array of numbers. An example of a matrix with rows $i$ and columns $j$ is:

$$\begin{bmatrix} 1 & 2 & 4 \\ 2 & 0 & -3 \\ 4 & -3 & 6 \end{bmatrix}$$

In a symmetric matrix, such as the one above, $a_{ij} = a_{ji}$. This means that all the corresponding nondiagonal elements are equal. Matrices may be added, subtracted, or manipulated in a variety of ways. Two matrices can be multiplied together provided that the number of columns in the first matrix $M_1$ equals the number of rows in the second matrix $M_2$. Following is an example of how to multiply $M_1$ by $M_2$.

Successively multiply each row of $M_1$ by each column of $M_2$:

$$M_1 = \begin{bmatrix} 3 & 4 \\ 0 & 2 \end{bmatrix} \quad M_2 = \begin{bmatrix} 5 & -2 \\ 2 & 1 \end{bmatrix}$$

$$M_{12} = \begin{bmatrix} (3)(5)+(4)(2) & (3)(-2)+(4)(1) \\ (0)(5)+(2)(2) & (0)(-2)+(2)(1) \end{bmatrix} = \begin{bmatrix} 23 & -2 \\ 4 & 2 \end{bmatrix}$$

If you want to try matrix multiplication yourself, enter the PAM1 mutation probability matrix of Fig. 3.11 into a program such as MATLAB® (Mathworks), divide each value by 10,000, and multiply the matrix times itself 250 times. You will get the PAM250 matrix of Fig. 3.13.

original amino acid

replacement amino acid

| PAM0 | A | R | N | D | C | Q | E | G |
|---|---|---|---|---|---|---|---|---|
| A | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| R | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 0 |
| N | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 |
| D | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 |
| C | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 |
| Q | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 |
| E | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0 |
| G | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100 |

original amino acid

replacement amino acid

| PAM∞ | A | R | N | D | C | Q | E | G |
|---|---|---|---|---|---|---|---|---|
| A | 8.7 | 8.7 | 8.7 | 8.7 | 8.7 | 8.7 | 8.7 | 8.7 |
| R | 4.1 | 4.1 | 4.1 | 4.1 | 4.1 | 4.1 | 4.1 | 4.1 |
| N | 4.0 | 4.0 | 4.0 | 4.0 | 4.0 | 4.0 | 4.0 | 4.0 |
| D | 4.7 | 4.7 | 4.7 | 4.7 | 4.7 | 4.7 | 4.7 | 4.7 |
| C | 3.3 | 3.3 | 3.3 | 3.3 | 3.3 | 3.3 | 3.3 | 3.3 |
| Q | 3.8 | 3.8 | 3.8 | 3.8 | 3.8 | 3.8 | 3.8 | 3.8 |
| E | 5.0 | 5.0 | 5.0 | 5.0 | 5.0 | 5.0 | 5.0 | 5.0 |
| G | 8.9 | 8.9 | 8.9 | 8.9 | 8.9 | 8.9 | 8.9 | 8.9 |

*FIGURE 3.12. Portion of the matrices for a zero PAM value (PAM0; upper panel) or for an infinite PAM∞ value (lower panel). At PAM∞ (i.e., if the PAM1 matrix is multiplied against itself an infinite number of times), all the entries in each row converge on the normalized frequency of the replacement amino acid (see Table 3.2). A PAM2000 matrix has similar values that tend to converge on these same limits. In a PAM2000 matrix, the proteins being compared are at an extreme of unrelatedness. In constrast, at PAM0, no mutations are tolerated, and the residues of the proteins are perfectly conserved.*

to an evolutionary distance where proteins share about 20% amino acid identity. Compare this matrix to the PAM1 matrix (Fig. 3.11) and note that much of the information content is lost. The diagonal from top left to bottom right tends to contain higher values than elsewhere in the matrix, but not in the dramatic fashion of the PAM1 matrix. As an example of how to read the PAM250 matrix, if the original amino acid is an alanine, there is just a 13% chance that the second sequence will also have an alanine. In fact, there is a nearly equal probability (12%) that the alanine will have been replaced by a glycine. For the least mutable amino acids, tryptophan and cysteine, there is more than a 50% probability that those residues will remain unchanged at this evolutionary distance.

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 13 | 6 | 9 | 9 | 5 | 8 | 9 | 12 | 6 | 8 | 6 | 7 | 7 | 4 | 11 | 11 | 11 | 2 | 4 | 9 |
| R | 3 | 17 | 4 | 3 | 2 | 5 | 3 | 2 | 6 | 3 | 2 | 9 | 4 | 1 | 4 | 4 | 3 | 7 | 2 | 2 |
| N | 4 | 4 | 6 | 7 | 2 | 5 | 6 | 4 | 6 | 3 | 2 | 5 | 3 | 2 | 4 | 5 | 4 | 2 | 3 | 3 |
| D | 5 | 4 | 8 | 11 | 1 | 7 | 10 | 5 | 6 | 3 | 2 | 5 | 3 | 1 | 4 | 5 | 5 | 1 | 2 | 3 |
| C | 2 | 1 | 1 | 1 | 52 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 3 | 2 | 1 | 4 | 2 |
| Q | 3 | 5 | 5 | 6 | 1 | 10 | 7 | 3 | 7 | 2 | 3 | 5 | 3 | 1 | 4 | 3 | 3 | 1 | 2 | 3 |
| E | 5 | 4 | 7 | 11 | 1 | 9 | 12 | 5 | 6 | 3 | 2 | 5 | 3 | 1 | 4 | 5 | 5 | 1 | 2 | 3 |
| G | 12 | 5 | 10 | 10 | 4 | 7 | 9 | 27 | 5 | 5 | 4 | 6 | 5 | 3 | 8 | 11 | 9 | 2 | 3 | 7 |
| H | 2 | 5 | 5 | 4 | 2 | 7 | 4 | 2 | 15 | 2 | 2 | 3 | 2 | 2 | 3 | 3 | 2 | 2 | 3 | 2 |
| I | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 10 | 6 | 2 | 6 | 5 | 2 | 3 | 4 | 1 | 3 | 9 |
| L | 6 | 4 | 4 | 3 | 2 | 6 | 4 | 3 | 5 | 15 | 34 | 4 | 20 | 13 | 5 | 4 | 6 | 6 | 7 | 13 |
| K | 6 | 18 | 10 | 8 | 2 | 10 | 8 | 5 | 8 | 5 | 4 | 24 | 9 | 2 | 6 | 8 | 8 | 4 | 3 | 5 |
| M | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 2 | 3 | 2 | 6 | 2 | 1 | 1 | 1 | 1 | 1 | 2 |
| F | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 3 | 5 | 6 | 1 | 4 | 32 | 1 | 2 | 2 | 4 | 20 | 3 |
| P | 7 | 5 | 5 | 4 | 3 | 5 | 4 | 5 | 5 | 3 | 3 | 4 | 3 | 2 | 20 | 6 | 5 | 1 | 2 | 4 |
| S | 9 | 6 | 8 | 7 | 7 | 6 | 7 | 9 | 6 | 5 | 4 | 7 | 5 | 3 | 9 | 10 | 9 | 4 | 4 | 6 |
| T | 8 | 5 | 6 | 6 | 4 | 5 | 5 | 6 | 4 | 6 | 4 | 6 | 5 | 3 | 6 | 8 | 11 | 2 | 3 | 6 |
| W | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 55 | 1 | 0 |
| Y | 1 | 1 | 2 | 1 | 3 | 1 | 1 | 1 | 3 | 2 | 2 | 1 | 2 | 15 | 1 | 2 | 2 | 3 | 31 | 2 |
| V | 7 | 4 | 4 | 4 | 4 | 4 | 4 | 5 | 4 | 15 | 10 | 4 | 10 | 5 | 5 | 5 | 7 | 2 | 4 | 17 |

*FIGURE 3.13. The PAM250 mutation probability matrix. From Dayhoff (1978, p. 350, fig. 83). At this evolutionary distance, only one in five amino acid residues remains unchanged from an original amino acid sequence (columns) to a replacement amino acid (rows). Note that the scale has changed relative to Fig. 3.11, and the columns sum to 100. Used with permission.*

## From a Mutation Probability Matrix to a Log-Odds Scoring Matrix

Our goal in studying PAM matrices is to derive a scoring system so that we can assess the relatedness of two sequences. When we perform BLAST searches (Chapters 4 and 5) or pairwise alignments, we employ a scoring matrix, but it is not in the form we have described so far. The PAM250 mutation probability matrix (Fig. 3.13) is useful because it describes the frequency of amino acid replacements between distantly related proteins. We next need to convert the elements of a PAM mutation probability matrix into a scoring matrix, also called a log-odds matrix or relatedness odds matrix.

The cells in a log-odds matrix consist of scores as defined in Equation 3.1 above. The target frequencies $q_{ij}$ are derived from a mutation probability matrix, such as those shown in Figs. 3.11 (for PAM1) and 3.13 (for PAM250). These values consist of positive numbers that sum to 1. The background frequencies $p_i p_j$ reflect the independent probabilities of each amino acid $i, j$ occurring in this position. Its values were given in Table 3.2.

For this scoring system Dayhoff and colleagues took 10 times the base 10 logarithm of the odds ratio (Equation 3.1). Using the logarithm here is helpful because it allows us to sum the scores of the aligned residues when we perform an overall alignment of two sequences. (If we did not take the logarithm, we would need to multiply the ratios at all the aligned positions, and this is computationally more cumbersome.)

A log-odds matrix for PAM250 is shown in Fig. 3.14. The values have been rounded off to the nearest integer. Try using Equation 3.1 to make sure you understand how the mutation probability matrix (Fig. 3.13) is converted into the log-odds scoring matrix (Fig. 3.14). As an example, to determine the score assigned to two aligned tryptophan residues, the PAM250 mutation probability matrix value is 0.55 (Fig. 3.13), and the normalized frequency of tryptophan is 0.010 (Table 3.2). Thus,

$$S_{(tryptophan,tryptophan)} = 10 \times \log_{10}\left(\frac{0.55}{0.01}\right) = +17.4 \qquad (3.2)$$

Note that this scoring matrix is symmetric, in contrast to the mutation probability matrix in Fig. 3.13. In a comparison of two sequences it does not matter which is given first. In problem [3-6] of this chapter we will calculate the likelihood of changing cys to glu, then of changing glu to cys.

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 2 | | | | | | | | | | | | | | | | | | | |
| R | −2 | 6 | | | | | | | | | | | | | | | | | | |
| N | 0 | 0 | 2 | | | | | | | | | | | | | | | | | |
| D | 0 | −1 | 2 | 4 | | | | | | | | | | | | | | | | |
| C | −2 | −4 | −4 | −5 | 12 | | | | | | | | | | | | | | | |
| Q | 0 | 1 | 1 | 2 | −5 | 4 | | | | | | | | | | | | | | |
| E | 0 | −1 | 1 | 3 | −5 | 2 | 4 | | | | | | | | | | | | | |
| G | 1 | −3 | 0 | 1 | −3 | −1 | 0 | 5 | | | | | | | | | | | | |
| H | −1 | 2 | 2 | 1 | −3 | 3 | 1 | −2 | 6 | | | | | | | | | | | |
| I | −1 | −2 | −2 | −2 | −2 | −2 | −2 | −3 | −2 | 5 | | | | | | | | | | |
| L | −2 | −3 | −3 | −4 | −6 | −2 | −3 | −4 | −2 | −2 | 6 | | | | | | | | | |
| K | −1 | 3 | 1 | 0 | −5 | 1 | 0 | −2 | 0 | −2 | −3 | 5 | | | | | | | | |
| M | −1 | 0 | −2 | −3 | −5 | −1 | −2 | −3 | −2 | 2 | 4 | 0 | 6 | | | | | | | |
| F | −3 | −4 | −3 | −6 | −4 | −5 | −5 | −5 | −2 | 1 | 2 | −5 | 0 | 9 | | | | | | |
| P | 1 | 0 | 0 | −1 | −3 | 0 | −1 | 0 | 0 | −2 | −3 | −1 | −2 | −5 | 6 | | | | | |
| S | 1 | 0 | 1 | 0 | 0 | −1 | 0 | 1 | −1 | −1 | −3 | 0 | −2 | −3 | 1 | 2 | | | | |
| T | 1 | −1 | 0 | 0 | −2 | −1 | 0 | 0 | −1 | 0 | −2 | 0 | −1 | −3 | 0 | 1 | 3 | | | |
| W | −6 | 2 | −4 | −7 | −8 | −5 | −7 | −7 | −3 | −5 | −2 | −3 | −4 | 0 | −6 | −2 | −5 | 17 | | |
| Y | −3 | −4 | −2 | −4 | 0 | −4 | −4 | −5 | 0 | −1 | −1 | −4 | −2 | 7 | −5 | −3 | −3 | 0 | 10 | |
| V | 0 | −2 | −2 | −2 | −2 | −2 | −2 | −1 | −2 | 4 | 2 | −2 | 2 | −1 | −1 | −1 | 0 | −6 | −2 | 4 |

*FIGURE 3.14. Log-odds matrix for PAM250. High PAM values (e.g., PAM250) are useful for aligning very divergent sequences. A variety of algorithms for pairwise alignment, multiple sequence alignment, and database searching (e.g., BLAST) allow you to select an assortment of PAM matrices such as PAM250, PAM70, and PAM30.*