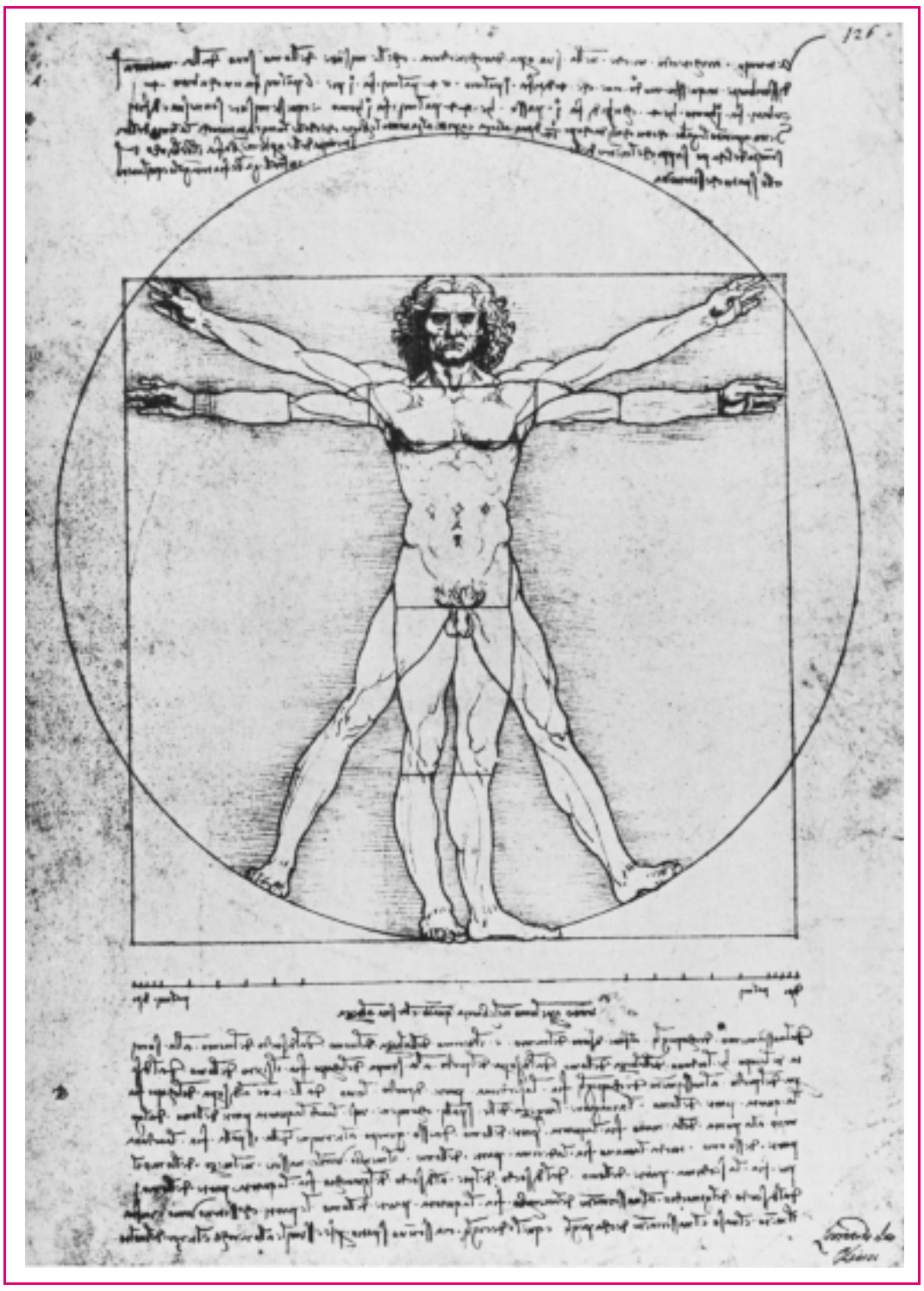
Part I

Analyzing DNA, RNA, and Protein Sequences in Databases



Leonardo da Vinci (1452–1519) drew the human body in 1490 based on the writings of Vitruvius. This drawing symbolizes Leonardo's quest to unify his art, science and engineering. Leonardo himself is a symbol for the effort to maximize human potential by understanding as many aspects of the human experience as possible. He attempted to study the human body from mathematical principles. The text accompanying this figure reads in part, "If you open the legs so as to reduce the stature by one-fourteenth, and open and raise your arms so that your middle fingers touch the line through the top of the head, know that the center of the extremities of the outspread limbs will be the umbilicus, and the space between the legs will make and equilateral triangle." We can use this image as a symbol to think about the efforts of bioinformatics and genomics to understand all of human biology, from molecular sequences to behavior.

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://grants.nih.gov/grants/bistic/CompuBioDef.pdf. For additional definitions of bioinformatics and functional genomics, see Boguski (1994), Luscombe et al. (2001), Ideker et al. (2001), and Goodman (2002).

For definitions of functional genomics, see http://bip.weizmann.ac.il/mb/functional_genomics.html.

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 translated into protein. Functional

Bioinformatics and Functional Genomics by Jonathan Pevsner
ISBN 0-471-21004-8

P1: FHB/SPH P2: FHB/SPH QC: UKS/FHB T1: UKS
PB347-01 PB347/Pevsner.cls August 5, 2003 14:38 Char Count= 0

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

The aim of this book is to explain both the theory and practice of bioinformatics. 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.

Organization of The Book

There are three main sections of the book. The first part 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].

The second part of the book describes functional genomics approaches to RNA and protein. The central dogma of biology states that DNA is transcribed into RNA then translated into protein. We will examine gene expression, including a description of the emerging technology of DNA microarrays (Chapters 6 and 7). We then consider proteins from the perspective of protein families, the analysis of individual proteins, protein structure, and multiple sequence alignment (Chapters 8–10). The relationships of protein and DNA sequences that are multiply aligned can be visualized in phylogenetic trees (Chapter 11). Chapter 11 thus introduces the subject of molecular evolution.

Since 1995, the genomes have been sequenced for several hundred bacteria and archaea as well as fungi, animals, and plants. The third section of the book covers genome analysis. Chapter 12 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 13) and bacteria and archaea (Chapter 14; these are two of the three main branches of life). Next we examine a variety of eukaryotes (from fungi to primates; Chapters 15 and 16) and then the human genome (Chapter 17). Finally, we explore bioinformatic approaches to human disease (Chapter 18).

Bioinformatics: The Big Picture

We can summarize the entire field of bioinformatics 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 (mRNA) transcripts, 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 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

P1: FHB/SPH P2: FHB/SPH QC: UKS/FHB T1: UKS
PB347-01 PB347/Pevsner.cls August 5, 2003 14:38 Char Count= 0

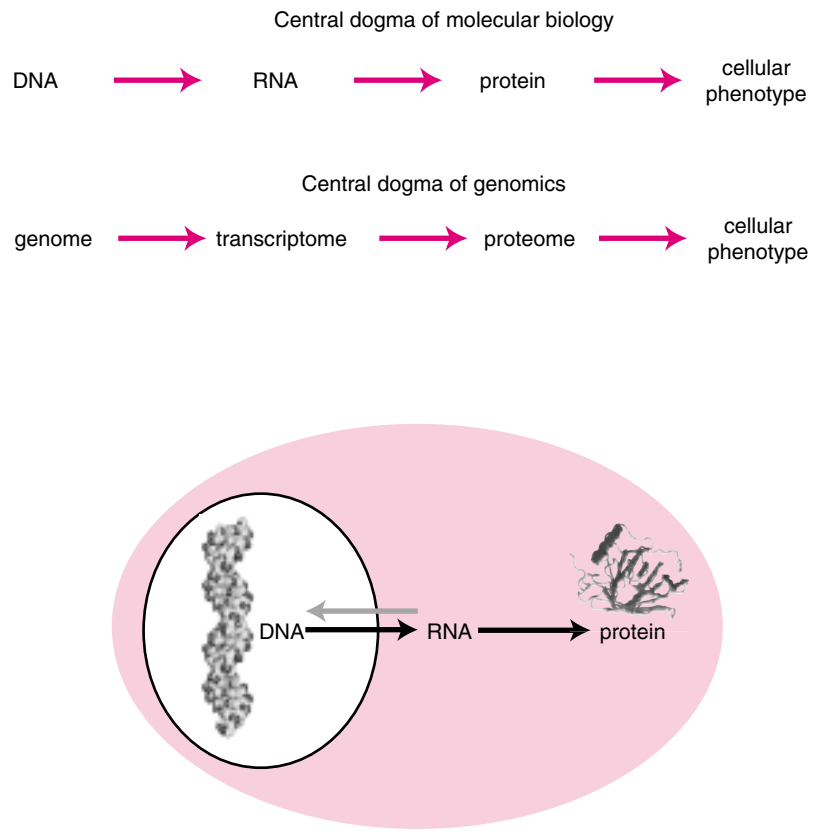


FIGURE 1.1. *The first perspective of the field of bioinformatics is the cell. Bioinformatics has emerged as a discipline as biology has become transformed by the emergence of molecular sequence data. Databases such as the European Molecular Biology Laboratory (EMBL), GenBank, and the DNA Database of Japan (DDBJ) serve as repositories for 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.*

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/proteins.

From the cell we can focus on individual organisms, which represents the 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

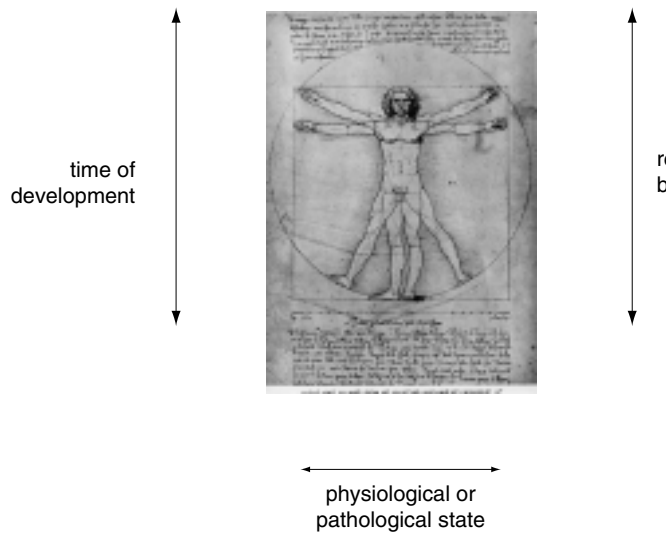


FIGURE 1.2. *The second perspective of bioinformatics is the organism. Broadening our view from the level of the cell to the organism, we can consider the individual's genome (collection of genes), including the genes that are expressed as RNA transcripts and the protein products. Thus, for an individual 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.*

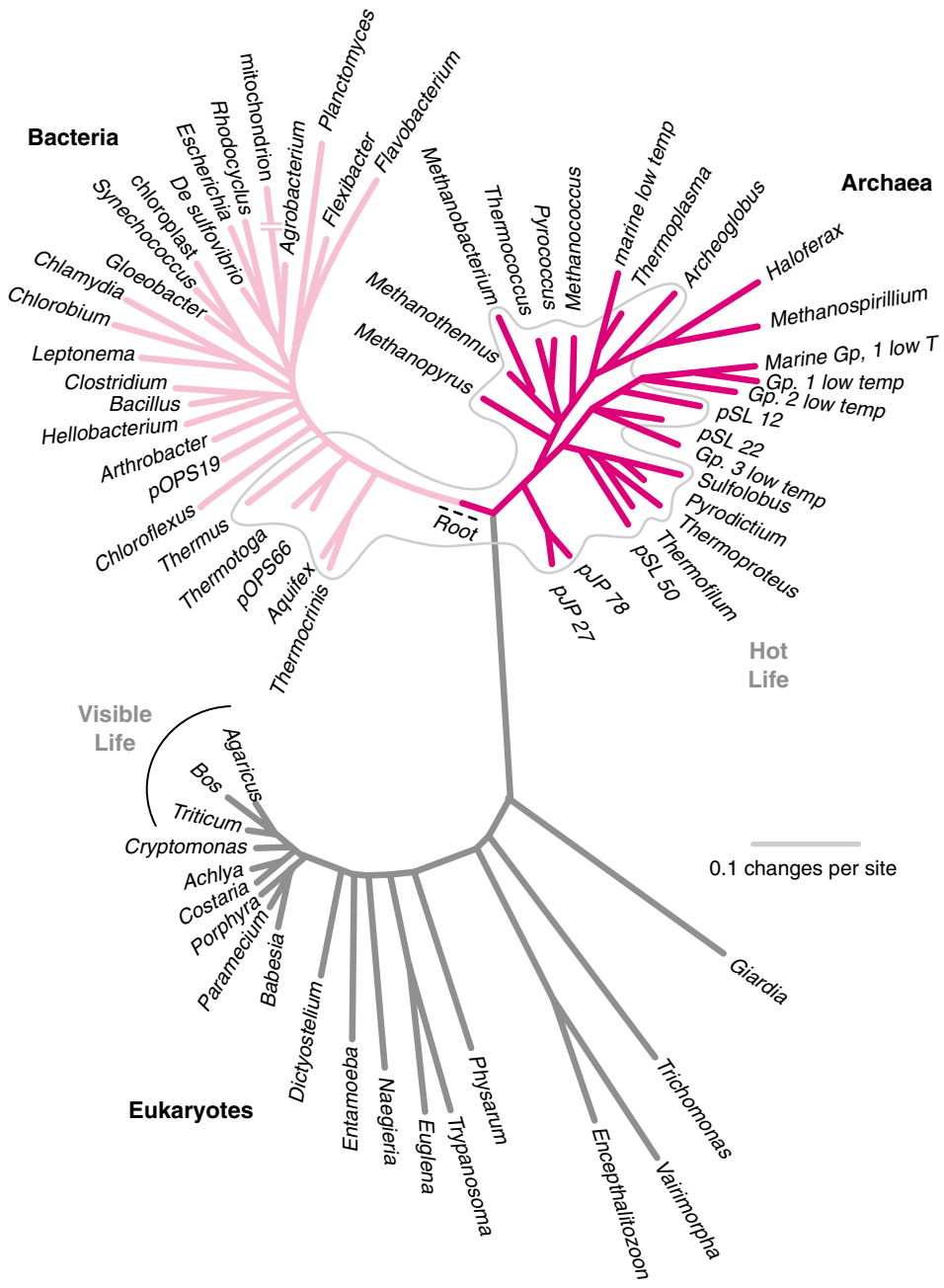


FIGURE 1.3. *The third perspective of the field of bioinformatics is represented by the tree of life. The scope of bioinformatics includes all of life on Earth, including the three major branches of bacteria, archaea, and eukaryotes. Viruses, which exist on the borderline of the definition of life, are 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 Pace, 1997.) Used with permission.*

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 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 12). 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 sequence from

P1: FHB/SPH P2: FHB/SPH QC: UKS/FHB T1: UKS
PB347-01 PB347/Pevsner.cls August 5, 2003 14:38 Char Count= 0

over 100,000 different organisms. The complete genome sequences of several hundred organisms will soon become available. 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.

Figure 1.4 on the following page presents the contents of this book in the context of the three perspectives of bioinformatics.

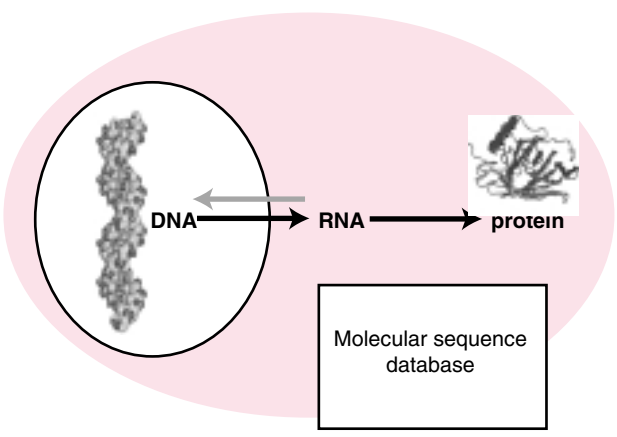
A Consistent Example: Retinol-Binding Protein

Throughout this book we will focus on the example of a gene and its corresponding protein product: 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 (Chapters 2–11).
- 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 even 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.
- The gene expression levels of some lipocalins are dramatically regulated (Chapters 6 and 7).
- 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 9).
- Some lipocalins have been implicated in human disease (Chapter 18).

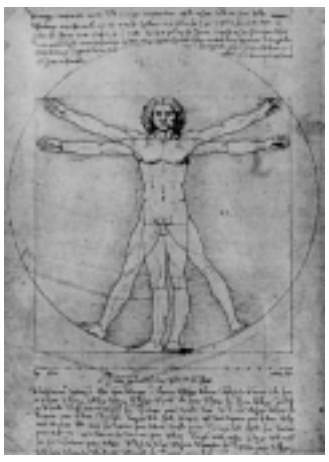
Another molecule we will introduce is the *pol* (polymerase) gene of human immunodeficiency virus 1 (HIV-1). HIV presents one of the greatest public health challenges in the world today. Over 42 million people are infected as of the end of the year 2002 and over 16 million people have died. The HIV-1 genome encodes just nine proteins, including *pol* (Frankel and Young, 1998). We will examine *pol* throughout the book because the properties of this gene, its protein products, and the HIV-1 genome are distinct from the lipocalins.

- The *pol* gene is a multidomain protein: it is a single polypeptide with several structurally and functionally distinct domains. The *pol* gene encodes a protein of 1003 amino acids with reverse transcriptase activity (that is, an



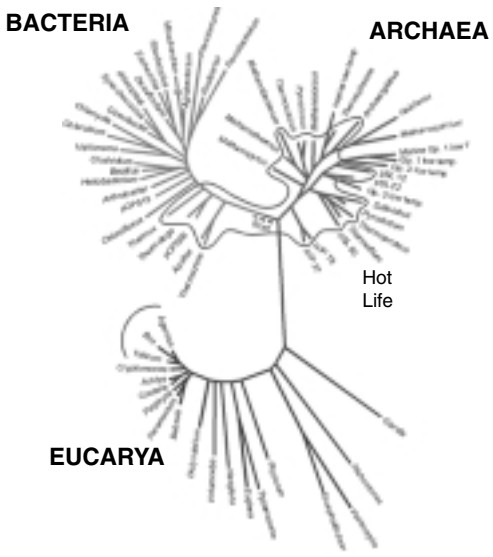
Part 1: Analyzing DNA, RNA, and protein sequences

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



Part 2: Genome-wide analysis of RNA and protein

Chapter 6: Gene expression
Chapter 7: Microarrays
Chapter 8: Protein analysis and protein families
Chapter 9: Protein structure
Chapter 10: How to multiply align sequences
Chapter 11: How to view multiply aligned sequences as phylogenetic trees



Part 3: Genome analysis

Chapter 12: The tree of life
Chapter 13: Viruses
Chapter 14: Prokaryotes
Chapters 15 and 16: Eukaryotes
Chapter 17: The human genome
Chapter 18: Human disease

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

P1: FHB/SPH P2: FHB/SPH QC: UKS/FHB T1: UKS
PB347-01 PB347/Pevsner.cls August 5, 2003 14:38 Char Count= 0

RNA-dependent DNA polymerase). It is also an aspartyl protease, and it has integrase activity. These multiple activities are typical of multidomain proteins.

- The modular nature of the pol protein affects our ability to perform database searches (Chapters 4 and 5) and multiple sequence alignments (Chapters 8 and 10).
- The *pol* gene incorporates substitutions extremely rapidly. A typical individual infected by HIV may have over a million variants of *pol*. The study of the evolution of *pol* complements our study of the lipocalins (Chapter 11).
- As a viral protein, our study of pol gives us the opportunity to learn how to access bioinformatics resources relevant to studying viruses (Chapter 13). Database searches with pol will help emphasize how to restrict searches to particular domains of the tree of life.

Organization of The Chapters

The website for this book (http://www.bioinfbook.org) contains about 1000 URLs, organized by chapter.)

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.

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 algorithm that is 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.

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 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.

Suggestions For Students and Teachers: Web Exercises and Find-a-Gene

Often, students of bioinformatics have a particular research area of interest such as a gene, a physiological process, a disease, or a genome. It is hoped that by studying RBP4 and other specific proteins and genes throughout this book, students

can simultaneously apply the principles of bioinformatics to their own research questions.

In teaching a course on bioinformatics at Johns Hopkins, it has been helpful to complement lectures with computer labs. All the websites described in this book are freely available on the World Wide Web, and many of the software packages are free for academic use.

Another feature of the Johns Hopkins 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 Chapter 5 (see Fig. 5.17). 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.

Teaching bioinformatics is notable for the diversity of students learning this new discipline. Each chapter provides background on the subject matter. For more advanced students, several 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 (Table 1.1). In Chapter 2 we begin with the National Center for Biotechnology

TABLE 1-1 Three Primary Bioinformatics Web Servers That Serve as Centralized Repositories for DNA and Protein Sequence Data

These will be introduced in Chapter 2

Resource	Description	URL[a]
DNA Data Bank of Japan (DDBJ)	Associated with the Center for Information Biology	http://www.ddbj.nig.ac.jp/
European Bioinformatics Institute (EBI)	Maintains the EMBL database	http://www.ebi.ac.uk/
National Center for Biotechnology Information (NCBI)	Maintains GenBank	http://www.ncbi.nlm.nih.gov/

[a] Uniform Resource Locator.

P1: FHB/SPH P2: FHB/SPH QC: UKS/FHB T1: UKS
PB347-01 PB347/Pevsner.cls August 5, 2003 14:38 Char Count= 0

TABLE 1-2 Additional Bioinformatics Web Servers

Each website contains access to dozens of software tools, research projects, literature references, and other information relevant to bioinformatics

Resource	Description	URL
Centre for Molecular and Biomolecular Informatics	From the University of Nijmegen	http://www.cmbi.kun.nl/
ExPASy (Expert Protein Analysis System)	Proteomics server of the Swiss Institute of Bioinformatics	http://www.expasy.org/
GENESTREAM	Institut de Génétique Humaine, Montpellier	http://www2.igh.cnrs.fr/
GenomeNet	In Kyoto	http://www.genome.ad.jp/
INFOBIOGEN	In Montpellier	http://www.infobiogen.fr/page_accueil_en.html
Oak Ridge National Laboratory (ORNL)	In Tennessee	http://compbio.ornl.gov/
Protein Information Resource (PIR)	A Division of National Biomedical Research Foundation	http://pir.georgetown.edu/
The Wellcome Trust Sanger Institute	A genome research center in Cambridge	http://www.sanger.ac.uk/
The Institute for Genomic Research (TIGR)	In Rockville, Maryland	http://www.tigr.org/

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. A general theme of the discipline of bioinformatics is that many databases are closely interconnected.

Throughout the chapters of this book we will introduce several hundred additional websites that are relevant to bioinformatics. Table 1.2 lists several additional servers that offer databases as well as many programs for the analysis of biological sequences. Table 1.3 lists several additional sites that offer links to bioinformatics resources. We present them now for those who wish to explore the types of bioinformatics resources that are currently available.

TABLE 1-3 Bioinformatics Sites with Useful Links

Websites that provide links to bioinformatics resources

Resource	Description	URL
Amos' WWW links page	From ExPASy	http://www.expasy.ch/alinks.html
DBCAT, The Public Catalog of Databases	From INFOBIOGEN	http://www.infobiogen.fr/services/dbcat/
European Molecular Biology Network	Various European nodes	http://www.embnet.org/
Human Genome Most Used Links	From Los Alamos National Laboratories	http://www-ls.lanl.gov/HGhotlist.html

P1: FHB/SPH P2: FHB/SPH QC: UKS/FHB T1: UKS
PB347-01 PB347/Pevsner.cls August 5, 2003 14:38 Char Count= 0

Suggested Reading

Overviews of the field of bioinformatics have been written by Mark Gerstein and colleagues (Luscombe et al., 2001) and Claverie et al. (2001). 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.

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).

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).

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).

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

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).

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).