

Part I

Analyzing DNA, RNA, and Protein Sequences in Databases

COPYRIGHTED MATERIAL

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 (C₄₀₀H₂₆₀N₁₀₀O₁₂₀), with varying amounts of phosphorus and sulfur. Justus Liebig (1803–1873) believed that the composition of protein was C₄₈H₃₆N₆O₁₄. 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).

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.”

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

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).

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

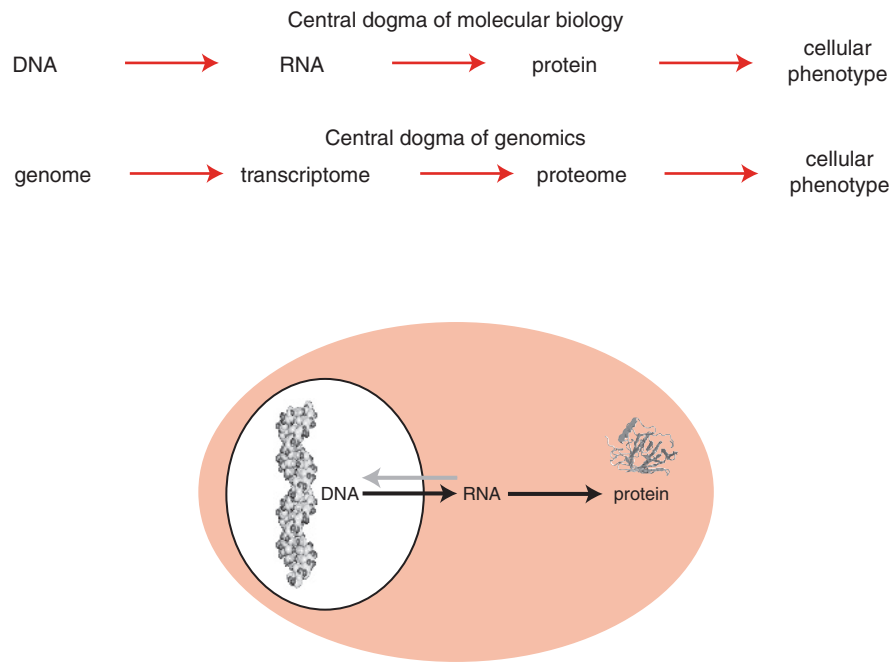


FIGURE 1.1. The first perspective of the field of bioinformatics is the cell. Bioinformatics has emerged as a discipline as biology has become transformed by the emergence of molecular sequence data. Databases such as the European Molecular Biology Laboratory (EMBL), GenBank, and the DNA Database of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

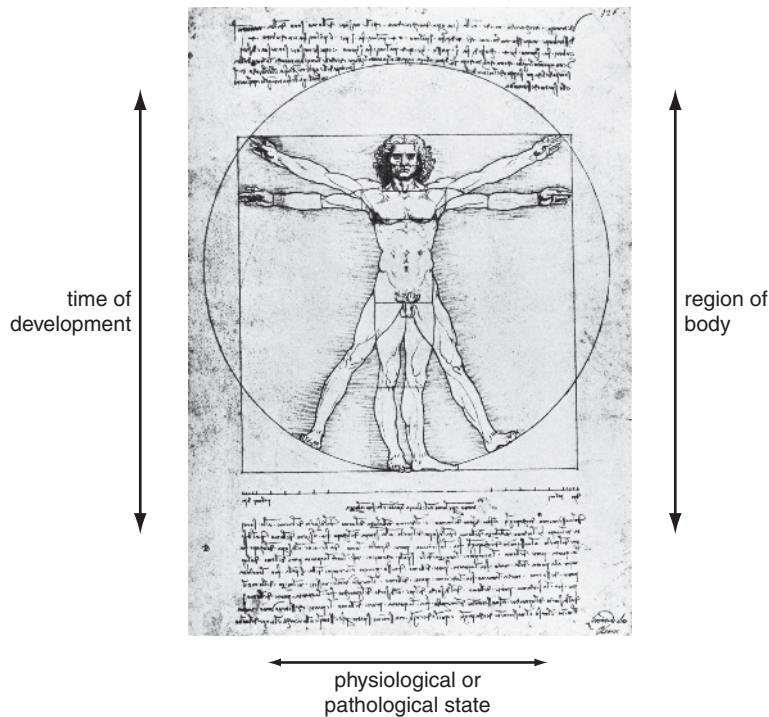


FIGURE 1.2. The second perspective of bioinformatics is the organism. Broadening our view from the level of the cell to the organism, we can consider the individual's genome (collection of genes), including the genes that are expressed as RNA transcripts and the protein products. Thus, for an individual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

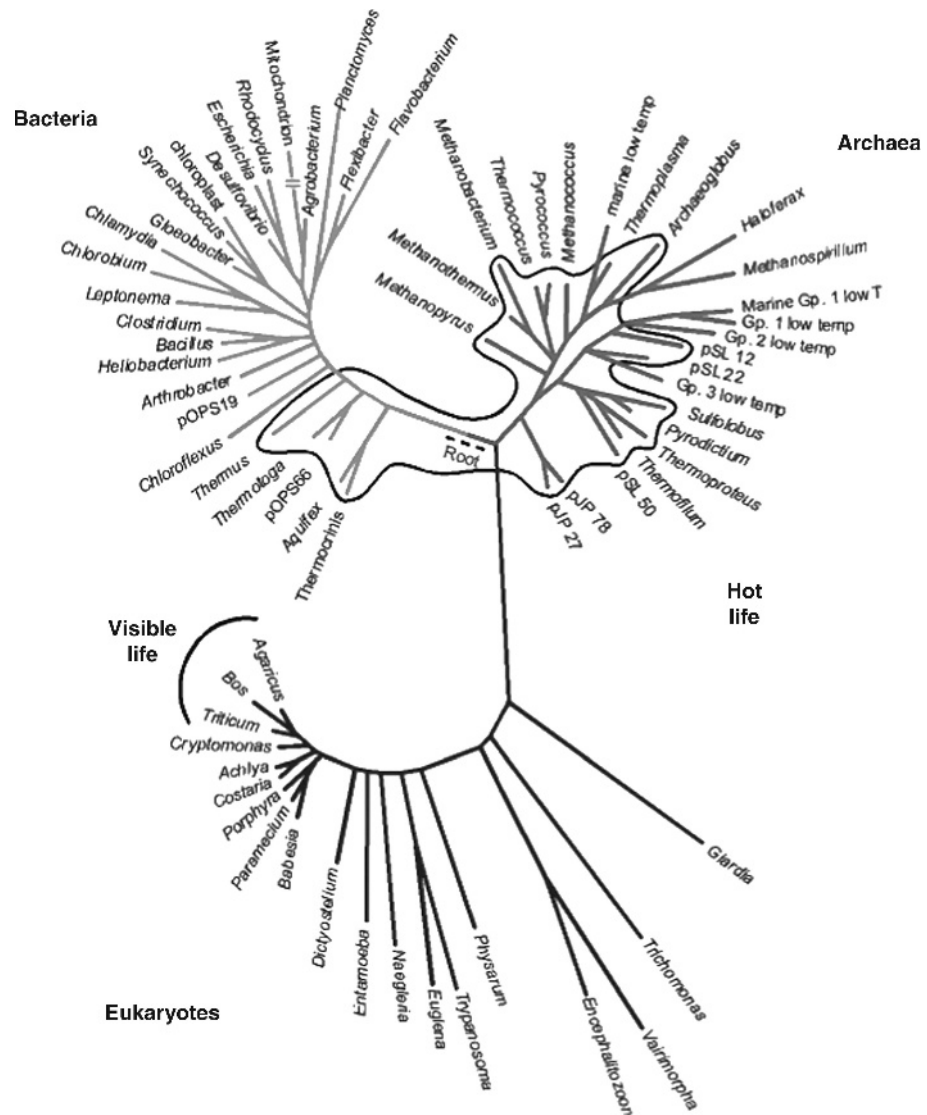


FIGURE 1.3. The third perspective of the field of bioinformatics is represented by the tree of life. The scope of bioinformatics includes all of life on Earth, including the three major branches of bacteria, archaea, and eukaryotes. Viruses, which exist on the borderline of the definition of life, are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

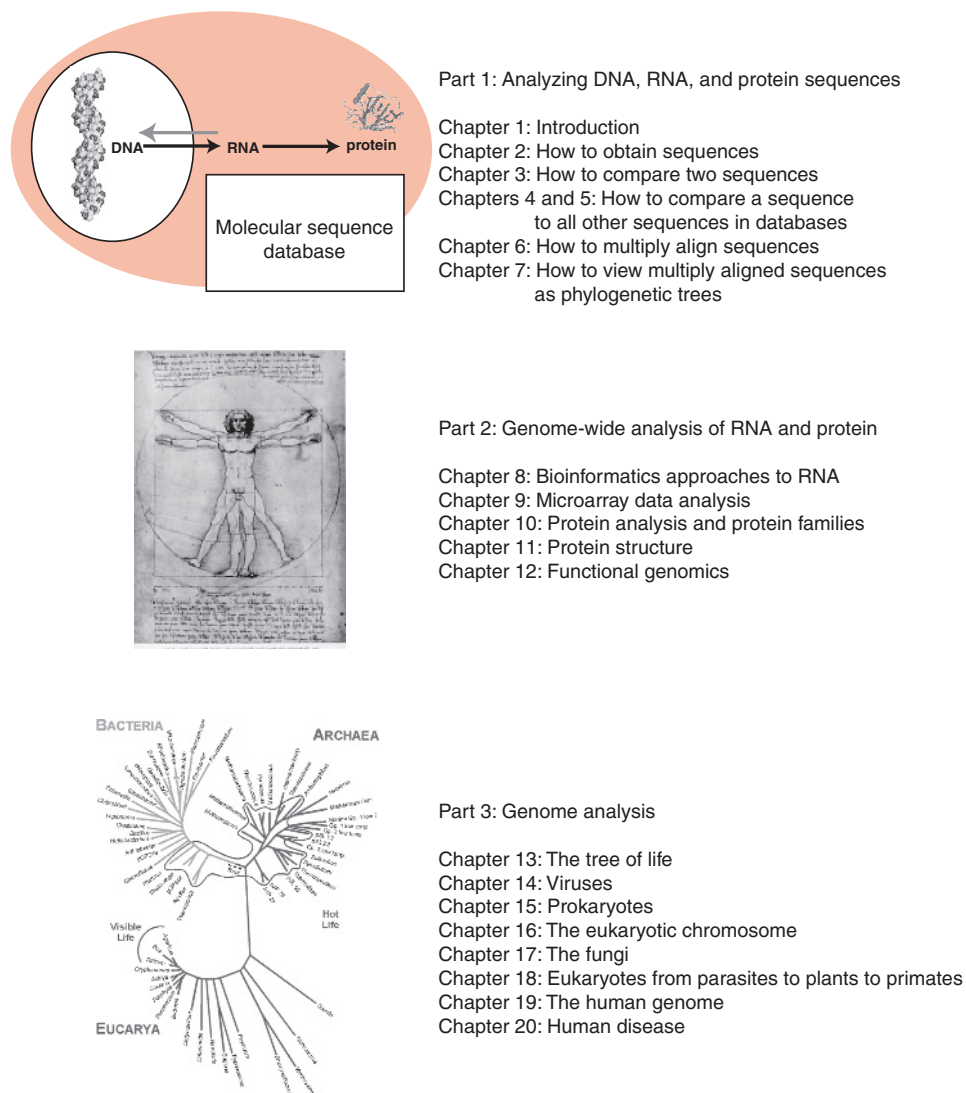


FIGURE 1.4. Overview of the chapters in this book.

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 α globin) and 11 (for β 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 apolipoprotein 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 parameters 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 features 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. AMLA 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).