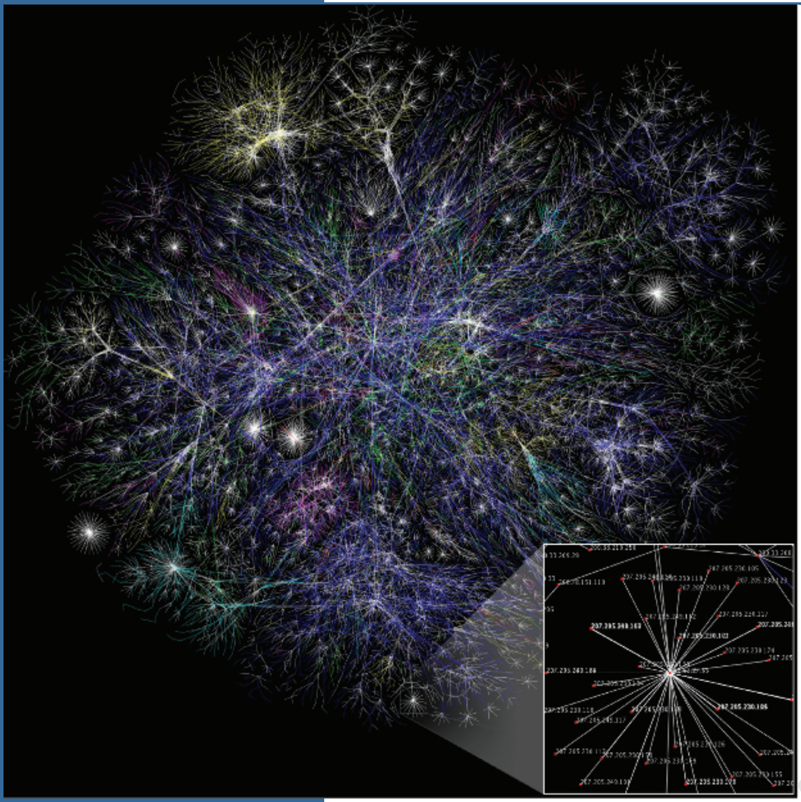
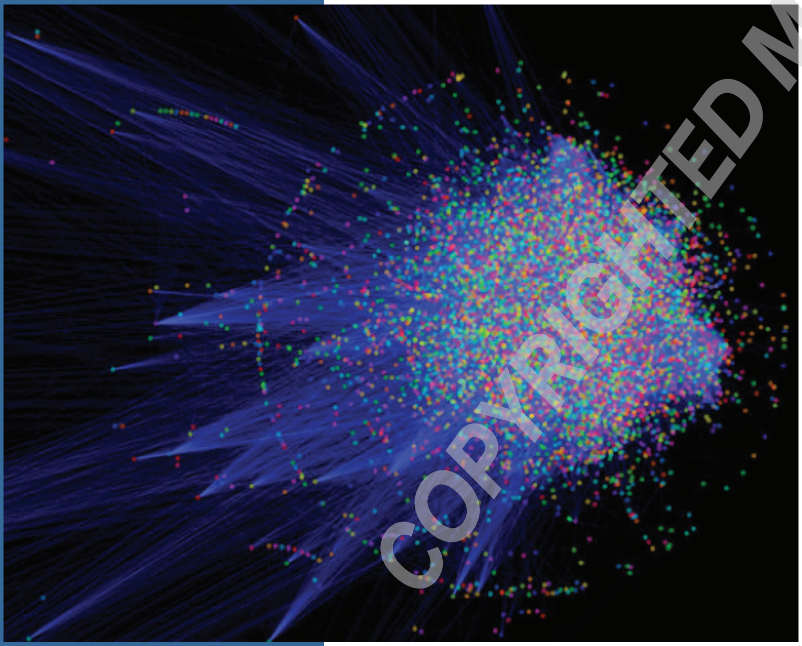


The first third of this book covers essential topics in bioinformatics. Chapter 1 provides an overview of the approaches we take, including the use of web-based and command-line software. We describe how to access sequences (Chapter 2). We then align them in a pairwise fashion (Chapter 3) or compare them to members of a database using BLAST (Chapter 4), including specialized searches of protein or DNA databases (Chapter 5). We next perform multiple sequence alignment (Chapter 6) and visualize these alignments as phylogenetic trees with an evolutionary perspective (Chapter 7).



The upper image shows the connectivity of the internet (from the Wikipedia entry for “internet”), while the lower image shows a map of human protein interactions (from the Wikipedia entry for “Protein–protein interaction”). We seek to understand biological principles on a genome-wide scale using the tools of bioinformatics.

Sources: Upper: Dcrjsr, 2002. Licensed under the Creative Commons Attribution 3.0 Unported license. Lower: The Opte Project, 2006. Licensed under the Creative Commons Attribution 2.5 Generic license.

Introduction

CHAPTER 1

Penetrating so many secrets, we cease to believe in the unknowable. But there it sits nevertheless, calmly licking its chops.

—*H.L. Mencken*

LEARNING OBJECTIVES

After reading this chapter you should be able to:

- define the terms bioinformatics;
- explain the scope of bioinformatics;
- explain why globins are a useful example to illustrate this discipline; and
- describe web-based versus command-line approaches to bioinformatics.

Bioinformatics represents a new field at the interface of the ongoing revolutions in molecular biology and computers. I define bioinformatics as the use of computer databases and computer algorithms to analyze proteins, genes, and the complete collection 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." Another definition from the National Human Genome Research Institute (NHGRI) is that "Bioinformatics is the branch of biology that is concerned with the acquisition, storage, display, and analysis of the information found in nucleic acid and protein sequence data."

The NIH Bioinformatics Definition Committee findings are reported at 🌐 http://www.bisti.nih.gov/docs/CompuBioDef.pdf (WebLink 1.1 at http://bioinfbook.org). The NHGRI definition is available at 🌐 http://www.genome.gov/19519278 (WebLink 1.2).

Russ Altman (1998) and Altman and Dugan (2003) offer two definitions of bioinformatics. The first involves information flow following the central dogma of molecular biology (**Fig. 1.1**). The second definition involves information flow that is transferred based

Bioinformatics and Functional Genomics, Third Edition, Jonathan Pevsner.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.
Companion Website: www.wiley.com/go/pevsnerbioinformatics

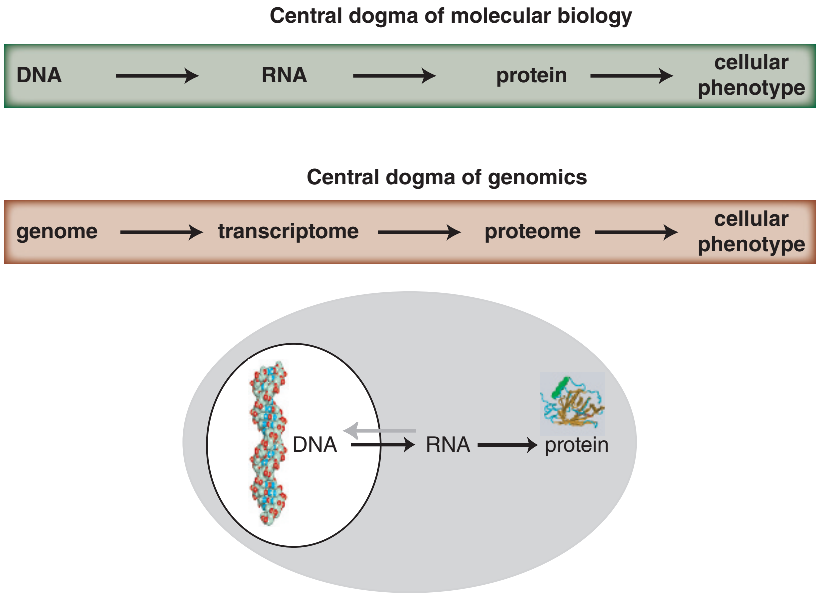


FIGURE 1.1 A first perspective of the field of bioinformatics is the cell. Bioinformatics has emerged as a discipline as biology has become transformed by the emergence of molecular sequence data. Databases such as the European Molecular Biology Laboratory (EMBL), GenBank, the Sequence Read Archive, and the DNA Database of Japan (DDBJ) serve as repositories for quadrillions (10^{15}) 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.

on scientific methods. This second definition includes problems such as designing, validating, and sharing software; storing and sharing data; performing reproducible research workflows; and interpreting experiments.

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 can be expressed as ribonucleic acid (RNA) transcripts and then, in many cases, further translated into protein. Functional genomics describes the use of genome-wide assays to study gene and protein function. For humans and other species, it is now possible to characterize an individual’s genome, collection of RNA (transcriptome), proteome and even the collections of metabolites and epigenetic changes, and the catalog of organisms inhabiting the body (the microbiome) (Topol, 2014).

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. Part I (Chapters 2–7) 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 or BLAST; Chapters 4 and 5). We introduce multiple sequence alignment (Chapter 6) and show how multiply aligned proteins or nucleotides can be visualized in phylogenetic trees (Chapter 7). Chapter 7 therefore introduces the subject of molecular evolution.

Part II describes functional genomics approaches to DNA, RNA, and protein and the determination of gene function (Chapters 8–14). The central dogma of biology states that DNA is transcribed into RNA then translated into protein. Chapter 8 introduces chromosomes and DNA, while Chapter 9 describes next-generation sequencing technology (emphasizing practical data analysis). We next examine bioinformatic approaches to RNA (Chapter 10), including both noncoding and coding RNAs. We then describe the measurement of mRNA (i.e., gene expression profiling) using microarrays and RNA-seq. Again we focus on practical data analysis (Chapter 11). From RNA we turn to consider proteins from the perspective of protein families, and the analysis of individual proteins (Chapter 12) and protein structure (Chapter 13). We conclude the second part of the book with an overview of the rapidly developing field of functional genomics (Chapter 14),which integrates contemporary approaches to characterizing the genome, transcriptome, and proteome.

Part III covers genome analysis across the tree of life (Chapters 15–21). Since 1995, the genomes have been sequenced for several thousand viruses, bacteria, and archaea as well as eukaryotes such as fungi, animals, and plants. Chapter 15 provides an overview of the study of completed genomes. We describe bioinformatics resources for the study of viruses (Chapter 16) and bacteria and archaea (Chapter 17; these are two of the three main branches of life). Next we explore the genomes of a variety of eukaryotes including fungi (Chapter 18), organisms from parasites to primates (Chapter 19) and then the human genome (Chapter 20). Finally, we explore bioinformatic approaches to human disease (Chapter 21).

The third part of the book, spanning the tree of life from the perspective of genomics, depends strongly on the tools of bioinformatics from the first two parts of the book. I felt that this book would be incomplete if it introduced bioinformatics without also applying its tools and principles to the genomes of all life.

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**). Here we follow the central dogma. A focus of the field of bioinformatics is the collection of DNA (the genome), RNA (the transcriptome), and protein sequences (the proteome) that have been amassed. These millions–quadrillions 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 single genes and proteins to cellular pathways and networks or even whole-genomic responses. 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 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

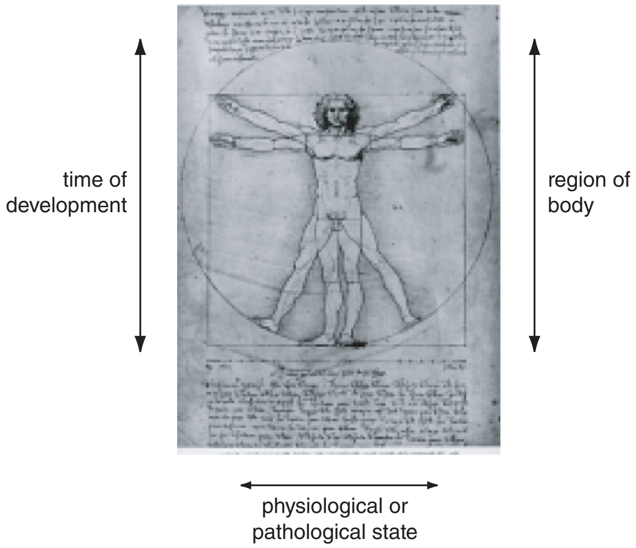


FIGURE 1.2 A second perspective of bioinformatics is the organism. Broadening our view from the level of the cell to the organism, we can consider the individual's genome (collection of genes), including the genes that are expressed as RNA transcripts and the protein products. For an individual organism,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

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 or RNA-seq to measure the expression of thousands of genes in biological samples.

At the largest scale is the tree of life (**Fig. 1.3**; see also Chapter 15). There are many millions of species alive today, and they can be grouped into the three major branches of bacteria, archaea, and eukaryotes. Molecular sequence databases currently hold DNA sequence from ~300,000 different species. The complete genome sequences of thousands of organisms are now 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. Through DNA sequence analysis we are learning how chromosomes evolve and are sculpted through processes such as chromosomal duplications, deletions, and rearrangements, and through whole-genome duplications (Chapters 8 and 18–19).

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

A Consistent Example: Globins

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 is one of the first proteins to be studied, having been described in the 1830s and 1840s by Gerardus Johannes Mulder, Justus Liebig, and others.
- Myoglobin, a globin that binds oxygen in the muscle tissue, was the first protein to have its structure resolved by X-ray crystallography (Chapter 13).

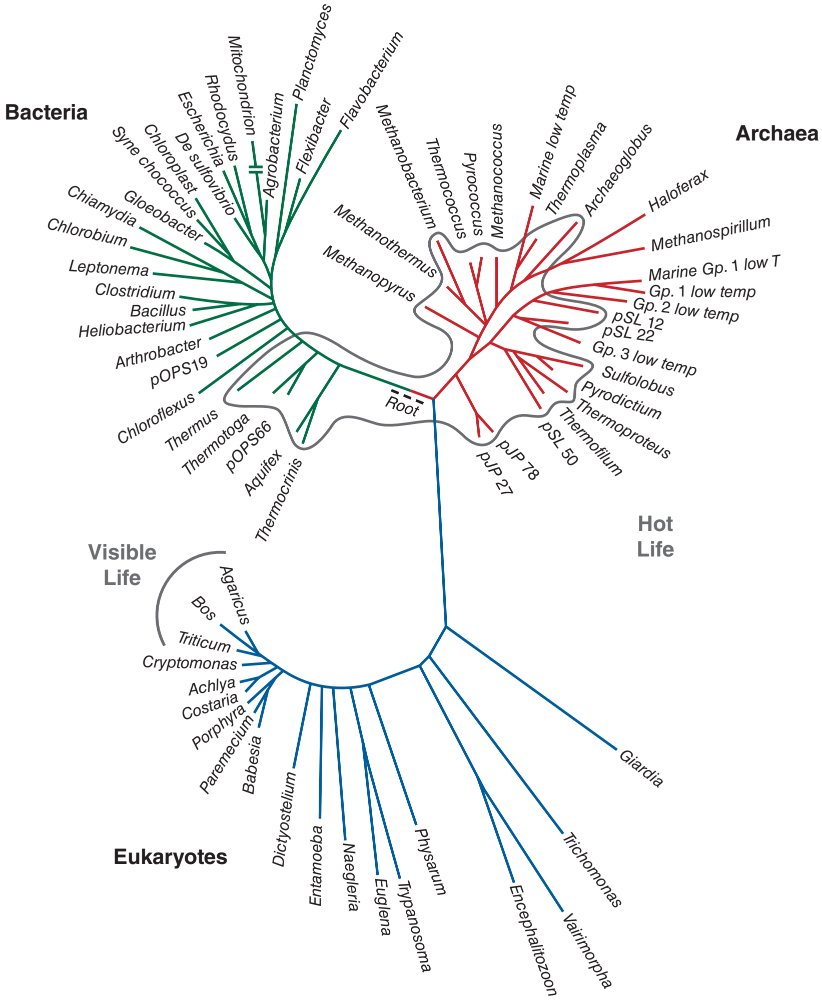
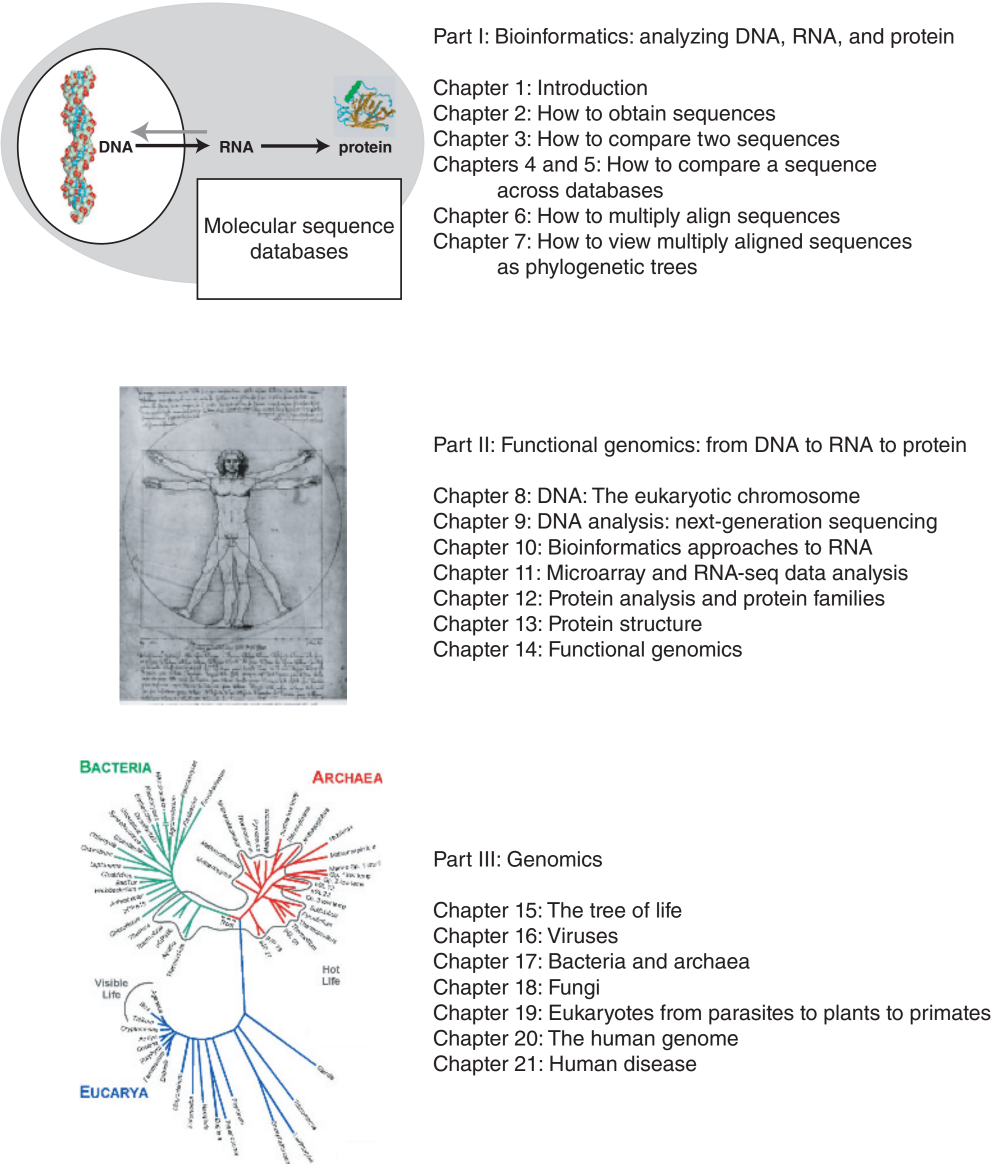


FIGURE 1.3 A third perspective of the field of bioinformatics is represented by the tree of life. The scope of bioinformatics includes all of life on Earth, including the three major branches of bacteria, archaea, and eukaryotes. Viruses, which exist on the borderline of the definition of life, are not 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. Adapted from Barns *et al*. (1996), Hugenholtz and Pace (1996), and Pace (1997).

- Hemoglobin, a tetramer of four globin subunits (principally $\alpha_2\beta_2$ in adults), is the main oxygen carrier in the 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).
- As DNA sequencing technology emerged in the 1980s, 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 (Chapters 10 and 14).
- While hemoglobin and myoglobin remain the best-characterized globins, the family of homologous proteins extends to separate classes of plant globins, invertebrate



Part I: Bioinformatics: analyzing DNA, RNA, and protein

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 across databases
Chapter 6: How to multiply align sequences
Chapter 7: How to view multiply aligned sequences as phylogenetic trees

Part II: Functional genomics: from DNA to RNA to protein

Chapter 8: DNA: The eukaryotic chromosome
Chapter 9: DNA analysis: next-generation sequencing
Chapter 10: Bioinformatics approaches to RNA
Chapter 11: Microarray and RNA-seq data analysis
Chapter 12: Protein analysis and protein families
Chapter 13: Protein structure
Chapter 14: Functional genomics

Part III: Genomics

Chapter 15: The tree of life
Chapter 16: Viruses
Chapter 17: Bacteria and archaea
Chapter 18: Fungi
Chapter 19: Eukaryotes from parasites to plants to primates
Chapter 20: The human genome
Chapter 21: Human disease

FIGURE 1.4 Overview of the chapters in this book.

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. The globin family is therefore useful as we survey the tree of life (Chapters 15–21).

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, Pitfalls, and Advice for Students. 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 repositories 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. Even for fundamental operations such as multiple sequence alignment (Chapter 6) and molecular phylogeny (Chapter 7), dozens of novel, ever-improving approaches are being 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 includes multiple-choice quizzes to 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 reference list at the end of each chapter is preceded by a discussion 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: EXERCISES, FIND-A-GENE, AND CHARACTERIZE-A-GENOME

This is a textbook for two separate courses: the first course is an introduction to bioinformatics (Parts I and II, i.e., Chapters 1–14), and the second is an introduction to genomics (Part III, i.e., Chapters 15–21). In a sense, the discipline of bioinformatics serves biology, facilitating ways of posing and then answering questions about proteins, genes, and genomes. Part III of this book surveys the tree of life from the perspective of genes and genomes, and could not progress without the bioinformatics tools described in Parts I and II of the book.

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

The websites described in this book are posted on the home page for this book (🌐 http://www.bioinfbook.org) as “WebLinks.” That site contains 900 URLs, organized by chapter. Each chapter also refers 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 a 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 4 (and summarized in Web Document 4.5 at 🌐 http://www.bioinfbook.org/chapter4). The student therefore 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. The student then 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; 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. Many students choose a gene (or protein) relevant to their own research area.

For a genomics course, students select a genome of interest and describe five aspects in depth (described at the start of Chapter 15):

1. The basic characteristics of the genome, such as its size, number of chromosomes, and other features, are described.
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 these new disciplines. Each chapter provides background on the subject matter. For more advanced students, key research papers are listed 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.

BIOINFORMATICS SOFTWARE: TWO CULTURES

There are two dramatically different approaches to bioinformatics: using web-based and command-line tools (**Fig. 1.5**). Web-based tools, sometimes called "point-and-click," do not require knowledge of programming and are immediately accessible.

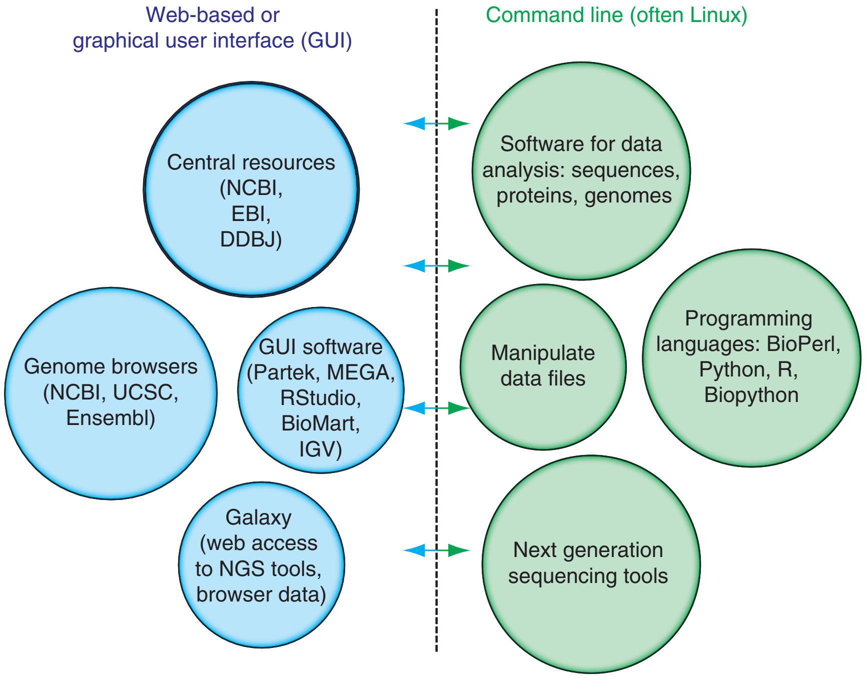

Web-based or
graphical user interface (GUI)
Command line (often Linux)
Central resources (NCBI, EBI, DDBJ)
Software for data analysis: sequences, proteins, genomes
Genome browsers (NCBI, UCSC, Ensembl)
GUI software (Partek, MEGA, RStudio, BioMart, IGV)
Manipulate data files
Programming languages: BioPerl, Python, R, Biopython
Galaxy (web access to NGS tools, browser data)
Next generation sequencing tools

FIGURE 1.5 Bioinformatics resources. Web-based or "point-and-click" resources are shown to the left, including the major portals (National Center for Biotechnology Information, European Bioinformatics Institute), major genome browsers (Ensembl, UCSC), databases, and specialized websites. Command-line resources are shown to the right. These include programming languages (such as Biopython, BioPerl, and the R language) and command-line software (typically accessed using the Linux operating system).

Command-line tools may have a steeper learning curve, but almost always offer more options for executing programs. They are more appropriate for analyzing large-scale datasets that are now routinely encountered in bioinformatics. Even for smaller datasets, command-line approaches can offer more flexibility and precision in accomplishing your tasks and more reproducible research since you can document your analysis steps.

Web-Based Software

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 main publicly accessible databases that serve as repositories for DNA and protein data. These include:

1. the National Center for Biotechnology Information (NCBI), which hosts GenBank and other resources;
2. the European Bioinformatics Institute (EBI);
3. Ensembl, which includes a genome browser and resources to study dozens of genomes; and
4. the University of California at Santa Cruz (UCSC) Genome bioinformatics site, including a web browser and table browser for a variety of species.

Throughout the chapters of this book we introduce almost 1000 additional websites that are relevant to bioinformatics. The main advantages offered by websites are easy access, rapid updates, good visibility to the community, and ease of use (since in general programming skills, command line skills, and the use of Linux-type operating systems are not required).

The URLs for these sites are NCBI, 🌐 http://ncbi.nlm.nih.gov (WebLink 1.3); EBI, 🌐 http://www.ebi.ac.uk/ (WebLink 1.4); Ensembl, 🌐 http://www.ensembl.org/ (WebLink 1.5); and UCSC, 🌐 http://genome.ucsc.edu/ (WebLink 1.6). For information on vast numbers of available databases, see the annual January issue of the journal *Nucleic Acids Research*, 🌐 http://nar.oxfordjournals.org/ (WebLink 1.7).

Command-Line Software

Command-line tools offer distinct, critical advantages. High-throughput approaches to biology result in the creation of both large and small datasets which require sophisticated analyses. We can think about command-line software in several ways.

1. The operating system is often Linux (a Unix-like environment). The Mac O/S is compatible with Linux as well (and is POSIX-compliant). However, while Windows-type operating systems are popular, they are not appropriate for the majority of command-line programs. In this book I assume the reader has no background in Unix. Beginning in Chapter 2, I provide basic instructions for becoming acquainted with Linux by providing examples of commands for a variety of software.
2. Programming languages are commonly used in bioinformatics. Examples are Perl (or its relative BioPerl; Stajich, 2007), Python (as well as Biopython), and R to manipulate data. Learning such languages is important as it is extremely useful to be able to write scripts and thus accomplish a broad range of tasks. Modules are available for hundreds of bioinformatics applications. For example, the BioConductor project currently includes > 1,000 packages that are useful for solving many tasks. Acquiring knowledge of R is a steep learning curve, and I provide suggestions of books, articles, and websites you can use to achieve this aim. It is also possible to use an `R` package without being an `R` "power user," however. For example, in Chapter 8 we use the `R` package `Biostrings` to extract information about the features on chromosomes, and in Chapter 11 we use R packages to analyze gene expression datasets from microarrays and next-generation sequencing. Once you learn to use a few packages, you will be in a postion to learn many more.

POSIX is an acronym for Portable Operating System Interface. It offers standards for maintaining compatibility between operating systems.

See http://bioinfbook.org/chapter1 for links to resources for learning Unix.

3. The command line of Unix systems offers Bash, a default shell for Linux and Mac OS X operating systems. We introduce a variety of Bash scripts in this book. Bash includes a series of utilities that can accomplish tasks such as sorting a table of data, transposing it, counting the numbers of rows and columns, merging data, or working with regular expressions. We'll see examples of Bash commands in Box 2.3 and in Chapter 9 on next-generation sequencing, for example.

Which operating system should you use? Linux is essential for many bioinformatics experts, often because it is used to access very large datasets (e.g., terabytes of data) with large amounts of RAM. For example, I recommend installing Bio-Linux on a laptop or a virtual machine. For many students approaching bioinformatics for the first time, the Macintosh O/S works well because it offers a Unix-like terminal. For Windows users, Cygwin provides a Unix-like environment. If you have access to a Linux server you can access it from a Windows or Mac environment using software such as PuTTY.

Bio-Linux 8 (released July 2014) is available at 🌐 http://environmentalomics.org/bio-linux/ (WebLink 1.8). Cygwin is available at 🌐 http://www.cygwin.com (WebLink 1.9). PuTTY is at 🌐 http://www.putty.org (WebLink 1.10).

We may further distinguish between using command-line software and using a programming language. Learning Perl, Python, or other languages offers tremendous benefits (Dudley and Butte, 2009). However, even if you do not program, you still should learn basic information about how to acquire, store, manipulate, and explore large files. Many files used in bioinformatics and genomics are simply too large to be handled efficiently (if at all) by web-based or GUI-based software. Many files that are generated by software tools require some level of restructuring to be further studied (e.g., to be analyzed by additional software tools). For many students, it has become essential to learn techniques to manipulate files on the command line.

Bridging the Two Cultures

Many bioinformatics resources are available to bridge the cultures of web-based and command-line software. This book introduces you to both (**Table 1.1**). For example, NCBI offers the web-based Entrez database that lets you type a query and obtain information. NCBI also provides EDirect, a set of command-line programs to access databases (see Chapter 2). Similarly, Ensembl offers programmatic access using Perl application programming interfaces (APIs). As another example, Galaxy hosts a broad range of web-based tools that are otherwise available as command-line software run on the Linux environment.

What is your best approach? Each person engaged in bioinformatics work should decide what problem he or she wants to solve, then choose the appropriate tool(s). If you are working with next-generation sequence data, it will be essential to learn how to use software tools in the Linux operating system. If that is new to you, you could use the more accessible Galaxy tools to start becoming familiar with the types of data and algorithms you will encounter as you transition to Linux-based tools. If you are doing phylogeny you can also start with MEGA software to learn a variety of approaches before complementing your analyses with command-line software to perform Bayesian analyses (see Chapter 7).

In this book we will use examples to try to help bridge these cultures. In Chapter 8 we will encounter both BioMart (an Ensembl web-based resource that interconnects hundreds of databases) and `biomaRt` (an R package that performs BioMart queries).

We will also see that the bioinformatics community is continuously improving existing software and developing new methods. There are often "competitions" in which organizers of an event obtain evidence of the gold standard "truth" for some problem, such as solving a protein structure or assembling a genome. Members of the community are then invited to compete to solve the answer within some time frame. By comparing the various results it is possible to assess the performance of each software

TABLE 1.1 Overview of some web-based (or graphical user interface (GUI)) and command-line software used in various chapters of this book.

Part: Chapter	Topic	Web-based or GUI software	Command-line software
I: 2	Access to information	BioMart Genome Workbench	EDirect
I: 3	Pairwise alignment	BLAST	BLAST+ Biopython needle (EMBOSS) water (EMBOSS)
I: 4	BLAST	BLAST	BLAST+
I: 5	Database searching	DELTA-BLAST Megablast	HMMER
I: 6	Multiple alignment	Pfam, MUSCLE	MAFFT
I: 7	Phylogeny	MEGA	MrBayes
II: 8	Chromosomes	Galaxy	geecee (EMBOSS) isochore (EMBOSS)
II: 9	Next-generation sequencing	Galaxy, SIFT, PolyPhen2	SAMTools, tabix, VCFtools
II: 10	RNA	RNAfam, tRNAscan	
II: 11	RNAseq	Galaxy	affy (R package), RSEM
II: 12	Proteomics	ExPASy	pepstats (EMBOSS)
II: 13	Protein structure	Cn3D, Pymol	psiphi (EMBOSS)
II: 14	Functional genomics	FLink, Cytoscape	
III: 15	Tree of life		Velvet (assembly)
III: 16	Viruses		MUMmer (alignment)
III: 17	Bacteria and archaea	MUMmer	GLIMMER (gene-finding)
III: 18	Fungi	YGOB	Ensembl (variants)
III: 19	Eukaryotic genomes		
III: 20	Human genome		PLINK
III: 21	Human disease	OMIM, BioMart	EDirect, MitoSeek

(i.e., true and false positives, true and false negatives); by defining the sensitivity and specificity of software we learn which tools to use. Examples of critical assessments are given in **Table 1.2**.

New Paradigms for Learning Programming for Bioinformatics

It is an excellent idea to learn a programming language to facilitate your bioinformatics work. You may want to run programs that are written in a language such as R or Python (as we do in this book), or you may want to write your own code and manipulate data to solve some task. In addition to available books and courses, many websites offer online training in the forms of tutorials or courses. David Searls (2012a, 2014) has reviewed many such online resources. These include Massive Open Online Courses (MOOCs) that tens of thousand of students may register for. Searls (2012b) also suggests ten rules for online learning. Briefly, these include: make a plan; be selective; organize your learning environment; do the readings; do the exercises; do the assessments; exploit the advantages (e.g., convenience); reach out to others; document your achievements; and be realistic

Excellent websites that guide you to learn a language include Code School (🌐 https://www.codeschool.com, WebLink 1.11), Code Academy (🌐 http://www.codecademy.com, WebLink 1.12), Data Camp (🌐 https://www.datacamp.com, WebLink 1.13), and Software Carpentry (🌐 http://software-carpentry.org, WebLink 1.14). Rosalind offers bioinformatics instruction through problem solving (🌐 http://rosalind.info/problems/locations/, WebLink 1.15).

TABLE 1.2 Critical assessment competitions in bioinformatics.

Name/Acronym	Competition	Chapter
Alignathon	Compare whole-genome alignment methods	6
EGASP	ENCODE Genome Annotation Assessment Project	8
Assemblathon	Compare the performance of genome assemblers	9
GAGE	Genome Assembly Gold-standard Evaluations	9
ABRF	Association of Biomolecular Resource Facilities (ABRF) assessment of phosphorylation	12
CASP	Critical Assessment of Structure Prediction	13
CAFA	Critical Assessment of Protein Function	14
CAGI	Critical Assessment of Genome Interpretation	14

about your expectations for what you can learn. These rules also apply to reading a textbook such as this one.

Reproducible Research in Bioinformatics

Science by its nature is cumulative and progressive. Whether you use web-based or command-line tools, research should be conducted in a way that is reproducible by the investigator and by others. This facilitates the cumulative, progressive nature of your work. In the realm of bioinformatics this means the following.

- A workflow should be well documented. This may include keeping text documents on your computer in which you can copy and paste complex commands, URLs, or other forms of data. Many people choose to maintain a traditional lab notebook, written by hand, but increasingly this must be accompanied by some form of electronic notebook.
- To facilitate your work, information stored on a computer should be well organized. In Box 2.3 we introduce a paper by Noble (2009), offering guidance on how to organize your files.
- Data should be made available to others. Repositories are available to store high-throughput data in particular. Examples are Gene Expression Omnibus (GEO) and Sequence Read Archive (SRA) at NCBI and ArrayExpress and European Nucleotide Archive (ENA) at EBI.
- Metadata can be equally as crucial as data. Metadata refers to information about datasets. For a bacterial genome that has been sequenced, the metadata may include the location from which the bacterium was isolated, the culture conditions, and whether it is pathogenic. For a study of gene expression in human brain, the metadata may include the post-mortem interval, the gender, the disease phenotype, and the method of RNA isolation. Metadata provide key information for statistical analyses, allowing the investigator to explore the effects of various parameters on the outcome measure.
- Databases that are used should be documented. Since the contents of databases change over time, it is important to document the version number and the date(s) of access.
- Software should be documented. For established packages, the version number should be provided. Further documenting the specific steps you use allows others to independently repeat your analyses. In an effort to share software, many researchers use repositories such as GitHub.

Git is the most popular distributed version control system for software development. It allows scientists to access software having specific versions. Github hosts both open and private projects. It is available online at 🌐 https://github.com (WebLink 1.16). As of early 2015, it has almost 20 million repositories and 8 million users.

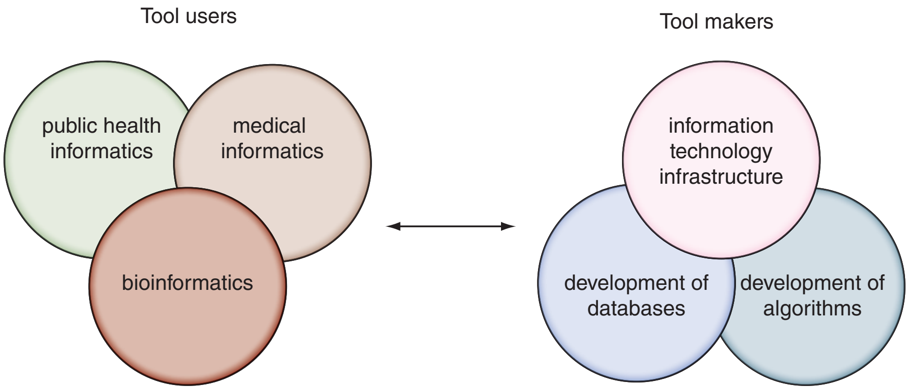

Tool users
public health informatics
medical informatics
bioinformatics
Tool makers
information technology infrastructure
development of databases
development of algorithms

FIGURE 1.6 Tool users and tool makers. The term "informatics" has been applied to an increasing number of disciplines in recent years including bioinformatics, public health informatics, medical informatics, and library informatics. Each of these disciplines is concerned with systematizing and analyzing increasingly large datasets. The focus of bioinformatics and genomics is on proteins, genes, and genomes in particular.

BIOINFORMATICS AND OTHER INFORMATICS DISCIPLINES

In recent years there has been a proliferation of other informatics fields including medical informatics, health care informatics, nursing informatics, and library informatics (**Fig. 1.6**). Bioinformatics has some overlap with these disciplines but is distinguished by its emphasis on DNA and other biomolecules. We may also distinguish tool users (e.g., biologists using bioinformatics software to study gene function, or medical informaticists using electronic health records) from tool makers (e.g., those who build databases, create information technology infrastructure, or write computer software). In bioinformatics, more than in other informatics disciplines, the tool users are also increasingly adept at being tool makers.

ADVICE FOR STUDENTS

The fields of bioinformatics and genomics are extremely broad. You should decide what range of problems you want to study, and what techniques are best suited to tackling those problems. Looking at **Figure 1.5**, you can see a broad range of available tools and approaches. As we move through the chapters it will likely become clear which is right for you. I encourage you to approach this textbook as actively as possible. When we discuss a website or a software package, take it as an opportunity to explore it in depth.

There are many ways to get help. Try using Biostars, an online forum in which you can post questions, get answers from the community, explore tutorials, and more (Parnell *et al*., 2011). By the year 2015, over 16,000 registered users have created >125,000 posts. Try joining Biostars or other bioinformatics forums to find others who have questions similar to yours.

Biostars was started in 2009 by Istvan Albert of Penn State University. Visit Biostars at 🌐 http://www.biostars.org (WebLink 1.17).

SUGGESTED READING

Dudley and Butte (2009) provide an excellent guide to developing effective bioinformatics programming skills (including the use of open source software and Unix). There have been relatively few general overviews of the field of bioinformatics in the past five years, perhaps because of its broadening scope. Thousands of reviews cover specialized topics. For all of Chapters 2–21 I provide sets of recent review articles.

In 2011 Eric Green, Mark Guyer and colleagues at the National Human Genome Research Institute published the highly recommended article: "Charting a course for genomic medicine from base pairs to bedside" (Green *et al*., 2011). This paper describes achievements in genomics and prospects for the coming decade.

Each January the journal Nucleic Acids Research offers a Database Issue that describes many central bioinformatics resources (Fernández-Suárez *et al*., 2014). That journal provides access to a vast number of papers via its website.

Visit the NAR database issue at http://nar.oxfordjournals.org/ (WebLink 1.18).

REFERENCES

Altman, R.B. 1998. Bioinformatics in support of molecular medicine. *Proceedings of AMIA Symposium* **1998**, 53–61. PMID: 9929182.

Altman, R.B., Dugan, J.M. 2003. Defining bioinformatics and structural bioinformatics. *Methods of Biochemical Analysis* **44**, 3–14. PMID: 12647379.

Barns, S.M., Delwiche, C.F., Palmer, J.D., Pace, N.R. 1996. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proceedings of the National Academy of Sciences, USA* **93**(17), 9188–9193. PMID: 8799176.

Dudley, J.T., Butte, A.J. 2009. A quick guide for developing effective bioinformatics programming skills. *PLoS Computational Biology* **5**(12), e1000589. PMID: 20041221.

Fernández-Suárez, X.M., Rigden, D.J., Galperin, M.Y. 2014. The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection. *Nucleic Acids Research* **42**(1), D1–6. PMID: 24316579.

Green, E.D., Guyer, M.S. 2011. National Human Genome Research Institute. Charting a course for genomic medicine from base pairs to bedside. *Nature* **470**(7333), 204–213. PMID: 21307933.

Hugenholtz, P., Pace, N.R. 1996. Identifying microbial diversity in the natural environment: a molecular phylogenetic approach. *Trends in Biotechnology* **14**, 190–197. PMID: 8663938.

Noble, W.S. 2009. A quick guide to organizing computational biology projects. *PLoS Computational Biology* 5(7), e1000424. PMID: 19649301.

Pace, N.R. 1997. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–740. PMID: 9115194.

Parnell, L.D., Lindenbaum, P., Shameer, K. *et al*. 2011. BioStar: an online question & answer resource for the bioinformatics community. *PLoS Computational Biology* **7**(10), e1002216. PMID: 22046109.

Searls, D.B. 2012a. An online bioinformatics curriculum. *PLoS Computational Biology* **8**(9), e1002632. PMID: 23028269.

Searls, D.B. 2012b. Ten simple rules for online learning. *PLoS Computational Biology* **8**(9), e1002631. PMID: 23028268.

Searls, D.B. 2014. A new online computational biology curriculum. *PLoS Computational Biology* **10**(6), e1003662. PMID: 24921255.

Stajich, J.E. 2007. An Introduction to BioPerl. *Methods in Molecular Biology* **406**, 535–548. PMID: 18287711.

Topol, E.J. 2014. Individualized medicine from prewomb to tomb. *Cell* **157**(1), 241–253. PMID: 24679539.