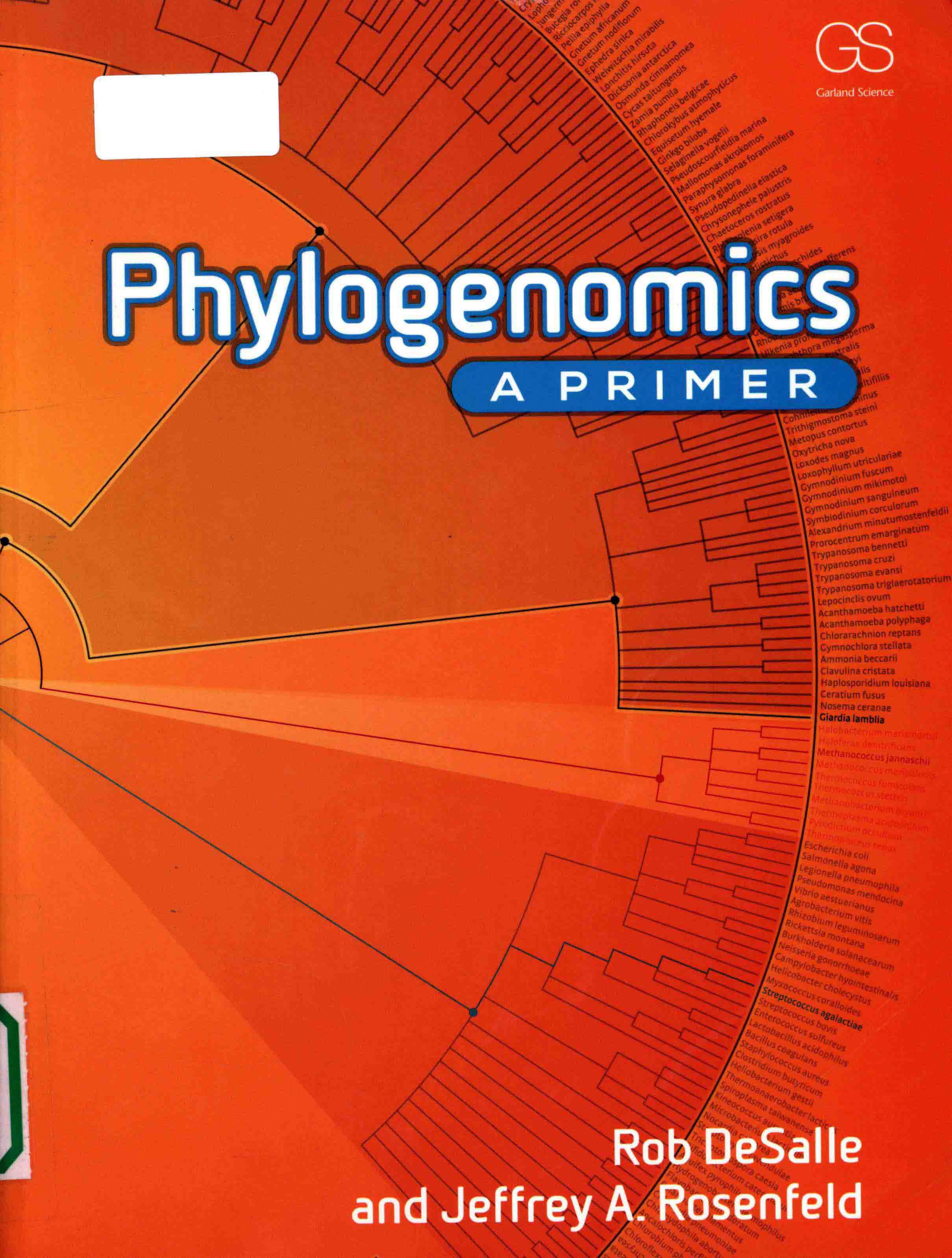
Garland Science

Phylogenomics

A PRIMER



Rob DeSalle
and Jeffrey A. Rosenfeld

Phylogenomics

A Primer

Rob DeSalle • Jeffrey A. Rosenfeld

Garland Science
Taylor & Francis Group
NEW YORK AND LONDON

Vice President: Denise Schanck
Senior Editor: Janet Foltin
Senior Editorial Assistant: Allie Bochicchio
Production Editor: Natasha Wolfe
Typesetter and
Senior Production Editor: Georgina Lucas
Copy Editor: Heather Whirlow Cammarn
Proofreader: Chris Purdon
Illustrations: Mill Race Studio
Indexer: Indexing Specialists(UK) Ltd

ISBN 978-0-8153-4211-3

Library of Congress Cataloging-in-Publication Data

DeSalle, Rob.
Phylogenomics : a primer / Rob DeSalle, Jeffrey A. Rosenfeld.
pages cm
Includes bibliographical references and index.
ISBN 978-0-8153-4211-3
1. Phylogeny--Molecular aspects. 2. Genomics. 3. Evolutionary genetics.
I. Rosenfeld, Jeffrey. II. Title.
QH367.5.D47 2013
572.8'38--dc23
2012036646

Published by Garland Science, Taylor & Francis Group, LLC, an informa business,
711 Third Avenue, 8th floor, New York, NY 10017, USA,
and 3 Park Square, Milton Park, Abingdon, OX14 4RN, UK.

Printed in the United States of America

15 14 13 12 11 10 9 8 7 6 5 4 3 2 1

Garland Science
Taylor & Francis Group

Visit our Website at http://www.garlandscience.com

Phylogenomics

Preface

This book is intended to serve as an introduction to a new area in comparative biology known as phylogenomics. Approximately 15 years ago, concurrent with the rapid and efficient sequencing of full genomes from living organisms, Clare Fraser and Jonathon Eisen coined the term "phylogenomics," a combination of phylogeny, which refers to the process whereby evolutionary trees are generated, and genomics, which represents the endeavor of obtaining genome-level data from organisms. Phylogenomics has developed into an important and compelling discipline in its own right. We developed this book in response to the students we have encountered over the last several years who are interested in applying genomics to comparative biology, specifically to phylogenetic, evolutionary, and population genetics problems.

Phylogenomics: A Primer is for advanced undergraduate students and graduate students in molecular biology, comparative biology, evolution, genomics, biodiversity, and informatics. Depending on their educational training, students can focus on the topics in the book that are of the most interest to them. Students who do not have strong backgrounds in evolution or phylogenomics will find the chapters that discuss evolutionary principles and the manipulation of phylogenomic-level data particularly useful. Conversely, students who are adept in ecology, taxonomy, and biodiversity will have the opportunity to learn about the evolution of genes and populations at the phylogenomic level and become familiar with applying phylogenomics to their genomics research.

We believe that there is no better way to understand the information that has been obtained about genes and genomes from life on this planet than to place it into context with the grand evolutionary experiment that has unfolded over the past 3.5 billion years. To this end, we have designed this book as a journey from the basic principles on which organic life has evolved, to the role of burgeoning databases in elucidating the function of proteins and organisms, and concluding with an interpretation of linear sequence information in the framework of organismal change.

Molecules are the currency of modern genomics and have an underlying linear arrangement of their component parts; that is, proteins and DNA can be broken down into linear sequences of amino acids and nucleotides, respectively. Chapters 1 and 2 present the essential principles underlying molecular biology and describe classical techniques used to analyze molecular sequences, including several high-throughput techniques that are known as "next generation" approaches. Chapter 3 explores evolution at the population level and introduces phylogenetic tree building. As a convenience, we make a simple demarcation between the evolutionary studies that focus on populations (microevolution) and those that focus on species relationships at higher-level systematic relationships (macroevolution).

Chapters 4 through 7 discuss the storage and manipulation of genomics-level data to enable the generation of the data sets that are used in phylogenomics. These processes include accessing databases and web programs such as PubMed, GenBank, and BLAST for downloading DNA and protein sequences; aligning linear sequences and producing matrices for evolutionary analysis; and assembling and annotating genomes.

Chapters 8 through 11 focus on the construction of evolutionary trees. Various approaches to phylogenetic analysis are presented, including distance, likelihood, parsimony, resampling, and Bayesian inference. In addition, the phenomenon of incongruence in relation to tree building is described as are the methods by which this problem is addressed.

Chapters 12 through 15 focus on the application of modern phylogenomics at the gene and population level. The transformation of population genetics by the use of DNA sequence information, the detection of natural selection on genes derived from genomic data, and the application of genome-level approaches to population genetics is essential to the understanding of natural populations in an evolutionary context.

The book concludes with a discussion of the basic applications of phylogenomics in the context of modern genome research. Chapter 16 examines the use of genome content to understand evolution. The role of phylogenomics in biodiversity studies, specifically the construction of the tree of life, DNA barcoding, and metagenomics, is explored in Chapter 17. The final chapter describes how functional genomics can be applied in a phylogenomic context, specifically transcription-based approaches and protein–protein interactions.

Working through the applications described in this book does not require an extensive computer science background beyond basic skills such as using a terminal or web browser. We have developed a set of Web Features that are linked to specific methods discussed in the book and are designed to introduce students to the websites used to obtain and analyze data. These features are designed to be accessed via a laptop or desktop computer and most are Web-based. A few stand-alone programs are referenced as well, all of which can be downloaded and installed on either a Mac or PC.

Rob DeSalle
New York, New York

Jeffrey A. Rosenfeld
Newark, New Jersey

Student and Instructor Resources Websites

Accessible from www.garlandscience.com, the Student and Instructor Resources websites provide learning and teaching tools created for *Phylogenomics: A Primer*. The Student Resources Site is open to everyone and users have the option to register in order to use book-marking and note-taking tools. The Instructor Resource Site requires registration and access is available only to qualified instructors. To access the Instructor Resource Site, please contact your local sales representative or email science@garland.com.

For Students
Web Features
Web-based exercises designed to assist students in working with the programs and databases used to analyze phylogenomic data.

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

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

Contents

Detailed Contents

CHAPTER 1

Why Phylogenomics Matters

Phylogenomics is a new way of looking at biological information. It refers to the intersection of several important aspects of modern biology such as molecular biology, systematics, population biology, evolutionary biology, computation, and informatics, with genome-level information as the source for testing hypotheses and for interpretation of data. Because the amount of information from genomes is orders of magnitude greater than previously available, novel approaches and new skills are needed by biologists to make sense of these data. In order to understand the biological information in a phylogenomic context, we first need to understand the nature of biological information and why and how we organize it. Understanding the nuances of computing therefore becomes an integral part of understanding phylogenomics. But we also need to have a good handle on the important molecular and evolutionary questions facing modern biology in order to formulate the right questions.

Phylogenomics and Bioinformatics

In 1976, the genome of the RNA virus MS2 (3569 nucleotides long) was sequenced by RNA sequencing. The next year, the first complete genome sequence of a DNA-based organism, φX174, was decoded. At 5386 nucleotides long, this genome opened the door for sequencing other DNA-based genomes. It took two decades to advance the technology enough that the whole genome of a living organism could be sequenced. The first living organism to be sequenced was *Haemophilus influenzae* (the bacterium that causes influenza) in 1996. In rapid succession, several bacterial genomes and eukaryotic model organism genomes were sequenced, including yeast (*Saccharomyces*), fruit fly (*Drosophila*), plant (*Arabidopsis*), mouse (*Mus musculus*), and worm (*Caenorhabditis elegans*).

As DNA sequencing technology has improved, the number of DNA fragments sequenced has risen. Recent advances in technology have resulted in an explosion of information. The trend for DNA sequencing for the three decades after genomes were first sequenced, compiled by the National Center for Biotechnology Information, is shown in **Figure 1.1**. In the years 2005–2011, advances in sequencing technology have reached what is called the "next generation" (see Chapter 2). From 2005 onward, the upswing in the amount of sequence generated by laboratories across the globe via next-generation sequencing approaches appears linear even on a logarithmic scale. With novel organisms being sequenced at extremely rapid rates, the onslaught of new gene sequences and the need to annotate, systematize, and archive them are now seen as a problem that is not solvable by simple comparative methods or simple computational approaches. The realization that billions of base pairs of sequence would soon be available to researchers studying cell biology, genetics, developmental biology, biochemistry, and evolution pushed researchers to think of the best ways to organize and interpret the data for making inferences about the functional aspects of newly sequenced genes. The first steps to achieving these goals were to use newly developed bioinformatics approaches.

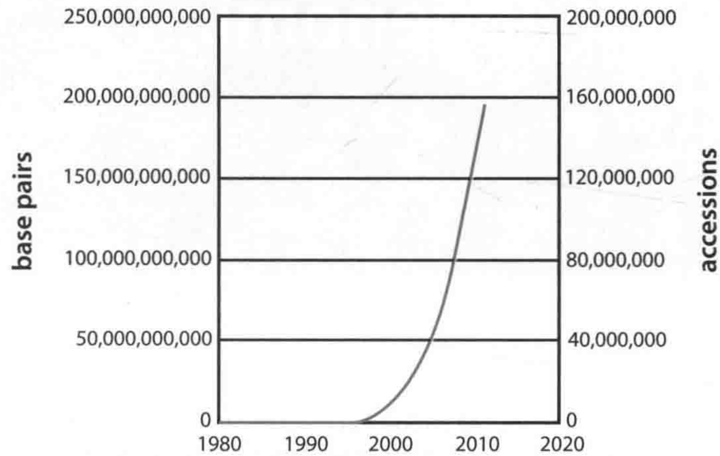


Figure 1.1 The increase in information being amassed by DNA sequencing projects. The curve represents the number of base pairs (on the left) and the number of "accessions" (on the right) for sequences in the National Center for Biotechnology Information database up to 2008. The *x*-axis shows the calendar year that the counts were made. The increase has reached a point where it is somewhat linear and nearly vertical. (Courtesy of the National Institutes of Health.)

Bioinformatics (Sidebar 1.1) is fundamentally the use of computational tools to answer biological questions and manage biological data. This term is generally synonymous with the term "computational biology," and the terms are used interchangeably. In this book, we will utilize the term bioinformatics. Examples of bioinformatics tasks include constructing phylogenetic matrices and building evolutionary trees, using microarray data to summarize the genes that are expressed in specific tissues, assembling the human genome, and predicting the three-dimensional fold of a protein. Bioinformatics is a wide field and its applications and needs are growing rapidly. While it was once considered a niche area that was separate from "wet-lab" biology, bioinformatics is now central to almost all biological investigations. Any time a biologist uses a computer to tabulate or analyze data, he or she is essentially doing bioinformatics.

The importance of bioinformatics has increased greatly with the introduction of technology that produces large amounts of data and the undertaking of large-scale projects. Below we briefly discuss two specific examples of high-throughput biology that have required a shift in the way we think about biological information. These two areas of modern biology—microarrays and the Human Genome Project—have given scientists the impetus to deal with large data sets in a bioinformatics context.

Sidebar 1.1. Origin of bioinformatics.

The origin of bioinformatics can be roughly traced to the publication of *What is Life?*, a monograph by Erwin Schrödinger, co-winner of the Nobel Prize in Physics in 1933. Schrödinger was one of the great physicists of the early twentieth century, and in this monograph, he discusses biology from a physicist's point of view. When the essay was published in 1944, physics was regarded as a mathematical science involved with quantum theory, while biology was considered an observational science with little need for mathematics. Schrödinger explained that there were many quantitative aspects to biological entities and that the proper understanding of these attributes would require the use of quantitative tools similar to those used by physicists. The ideas expressed spread throughout the world of physics, and James Watson and Francis Crick, who discovered the structure of DNA, trace their interest in biology to this work. Physicists were also motivated to investigate biological questions and applied their quantitative perspective to the biological field. Because their approaches were mathematical and computational, these aspects of physics were transferred to the study of biology to cope with the large amount of data flooding the field.

A microarray is a simple concept with powerful applications

The basic science behind a microarray is very simple, but the applications are very powerful. Its purpose is to determine the kinds and abundance of messenger RNA (mRNA) in a cell. Prior to the development of microarrays, measurement of mRNA levels was usually limited to a few genes at a time. Microarrays are discussed in more detail in Chapter 2, but a brief overview of the method is provided here. A microarray analyzes cellular RNA to determine the expression level (that is, how much mRNA is produced) for thousands of genes simultaneously. Single-stranded DNA sequences (probes) are affixed to a slide in specific positions, and the total RNA from a cell is extracted, labeled with a fluorescent dye, and hybridized with the DNA probes on the slide.

The objectives of microarray analysis include

- Determining where on the microarray the probes for various genes are located
- Determining the expression level of each gene from the fluorescence intensity of the probed DNA
- Determining which genes are significantly expressed
- Determining whether genes belonging to a particular functional category are overrepresented in the set of significantly expressed genes.

Microarrays are a fixture within biological laboratories devoted to diverse specialties, from bacterial genetics to cancer diagnostics. In order to effectively use this technology, bioinformatics skills are needed. Even the most basic microarray experiment involves a tremendous amount of bioinformatics. A large fraction of the bioinformatics may not be visible to the end-user, but it is impossible to ignore the impact of bioinformatics entirely.

The Human Genome Project was a watershed event in DNA sequencing

The Human Genome Project was the largest and most expensive biological project in history. It involved collaboration among genome centers around the world that were involved in sequencing the 3×10^9 bases in the human genome. It required a significant number and range of bioinformatics tools, many of which were specifically designed for this project. The bioinformatics tasks included

- Monitoring and organizing the data generated
- Efficient sharing of the data between genome centers
- Assembly of the sequence reads to compile the genome
- Annotation of the genome to determine the locations and functions of genes

Bioinformatics tools enable data analysis and identification of patterns in biological experiments

An important task in bioinformatics is processing the large amounts of data generated in high-throughput biological experiments. This information needs to be managed by computers so that it is accessible and understandable. As mentioned above, the human genome translates into a pattern of 3 billion letters of A, T, G, and C. If this information was written out on paper in 12-point Times font, it would be approximately the length of the complete Encyclopedia Britannica. Trying to find specific genes or DNA sequences in this much information without bioinformatics is like having to find a single specific word in all the volumes of the Encyclopedia Britannica *without* keywords or alphabetized arrangement of entries. Bioinformatics has produced tools that are used to sift through large amounts of information to discover patterns and processes. Informatics tools such as BLAST enable searches through the large number of DNA sequences that currently exist in the public databases (see Chapter 4). Besides the human genome, additional genetic sequence information has been collected from other organisms. The set of all publicly available DNA sequences is stored in GenBank and, as of April 2012, the size of the complete database was 471 gigabytes (471 GB = 471,000,000,000 bytes). The University of California at Santa Clara (UCSC) genome browser is a very highly utilized database of annotations for the human genome. It consists of 1.5 terabytes (1.5 TB = 1,500,000,000,000 bytes) of data.

The scope of the problems addressed by bioinformatics will continue to increase in the next few years (Figure 1.1). Several large high-throughput projects (the 1000 Genomes Project and the 10K Animal Genomes project are two examples) will increase the amount of sequence in the database by several orders of magnitude. The goal of the 1000 Genomes Project is to determine the complete genome sequences of 2500 individuals from diverse ethnic groups across the world. At 3 GB per genome, it is expected that this project will produce many terabytes of data. The 10K Animal Genomes project plans to produce the whole genome sequences of over 10,000 animals. This project will generate over 60 TB of data.

The Rise of Phylogenomics

The term phylogenomics (Sidebar 1.2) was first coined by Jonathan Eisen and Claire Fraser at The Institute for Genome Research (TIGR) at the turn of the century. Phylogenomics is an updating of the term phylogenetics and refers to focus on genome-level analysis. Whereas conventional phylogenetics is based upon the analysis of a few genes, phylogenomics would investigate complete genomes of data. At first, phylogenomics was applied to the functional annotation of newly sequenced genomes. **Table 1.1** (taken from Eisen) shows the comparative approaches that can be used to assign function to a newly sequenced gene. At the genome level for higher eukaryotes, this needs to be done tens of thousands

Sidebar 1.2. Where does the term phylogenomics come from?

To properly understand what phylogenomics is, we need to understand the two major roots of the word: phylo and genomic. The term is really a hybrid with Greek origins and more modern twists. The first part of the term, phylo, comes from the Greek root "phylon," which means group or tribe, which has been expanded into the modern word "phylogeny," or a diagram that represents grouping. Modern-day phylogenies are, at their simplest level, branching diagrams that represent the relatedness of organisms. But, as we will see, phylogenies can also carry information about the sequence of events that have occurred over evolutionary time. The second part of the term, genomics, comes from two subroots. The root word "gene" was first coined in 1903 by Wilhelm Johannsen, a Danish botanist, to refer to a unit of heredity. The suffix "omics" has a more modern origin: it has been applied to a number of root terms to signify an entirely new way of doing biology. When this suffix is applied to a root word, it usually means the exhaustive collection of information for a particular biological level. For instance, transcriptomics is the study of the entire array of transcripts made by a cell. Proteomics is the study of the entire array of proteins made by a cell. Similarly, genomics, a term first used in 1987 when scientists began to discuss the possibility of obtaining the DNA sequence of each base in the human genome, is the study of the entire array of DNA sequences contained in a cell. "Genome" studies proper began in 1996, when the first whole genome of a living organism (the bacterium *Haemophilus influenzae*) was produced by J. Craig Venter and his colleagues. Genomics includes the following steps:

- Obtaining sequences from the genome of an organism
- Assembly of those sequences into a single contiguous genome sequence (if the organism has a single chromosome) or sets of contiguous sequences (if the organism has more than one chromosome)
- Identification of the regions of the raw sequence that correspond to genes
- Annotation of the genes

Table 1.1. Comparative approaches to assigning gene function.

Name	Description of the approach	Example
Highest hit	The uncharacterized gene is assigned the function (or frequently, the annotated function) of the gene that is identified as the highest hit by a similarity search program.	Tomb et al., 1997
Top hits	Identify top 10+ hits for the uncharacterized gene. Depending on the degree of consensus of the functions of the top hits, the query sequence is assigned a specific function, a general activity with unknown specificity, or no function.	Blattner et al., 1997
Clusters of orthologous groups	Genes are divided into groups of orthologs based on a cluster analysis of pairwise similarity scores between genes from different species. Uncharacterized genes are assigned the function of characterized orthologs.	Tatusov et al., 1997

of times because the genomes of eukaryotes contain 10,000 to 30,000 genes. To date over a dozen species of *Drosophila* have had their genomes sequenced. The main reason for all of this fly sequencing was not because scientists were specifically interested in these other species, but rather because the sequences of these species gave scientists better tools to understand the function of the genome of *Drosophila melanogaster*, the model organism. In other words, these other species were sequenced simply because they would help with the annotation of a model organism genome. These kinds of approaches are called functional phylogenomics, because they attempt to get at the processes involved in the function of gene products. As time progressed, scientists realized the power of reconstructing phylogenetic relationships by use of genome-level information. So after about 5 years of usage of the term with its original meaning, other aspects of the use of genome-level sequences were assigned to the umbrella of phylogenomics. These include using an evolutionary approach to understand the function of genes and using whole genome sequences to interpret the relationships of organisms.

To give the student a sense of the power of a phylogenetic evolutionary approach to genomics, we present two examples. The first example concerns understanding the functional nature of protein products from genes (known as functional phylogenomics) and the second concerns the use of whole genome sequences to infer the pattern of relationships of organisms (known as pattern phylogenomics).

Functional phylogenomics employs common ancestry to infer protein function

Phylogenomic analysis allows for a way to use common ancestry to infer the function of an unknown protein. Brown and Sjölander have used the example of G protein coupled receptors to demonstrate how a phylogenetic approach can lead to annotation of function in a large group of proteins that might seem unrelated in the beginning. A branching diagram of protein sequences, derived from the opioid/galanin/somatostatin gene family that allows for two important inferences about assigning function to unknown proteins, is shown in **Figure 1.2**. Diamonds represent fully annotated and well-understood proteins, and ovals represent unannotated proteins. The structure of the tree allows researchers to focus on three subtrees—the opioid, galanin, and somatostatin receptors—and to assign a function for the unannotated proteins in the study. Thus unknown proteins can