

Phylogenomics

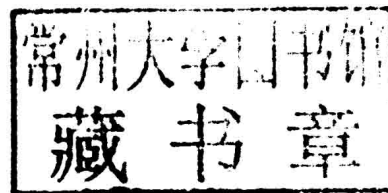
A PRIMER

Rob DeSalle
and Jeffrey A. Rosenfeld

Phylogenomics

A Primer

Rob DeSalle • Jeffrey A. Rosenfeld



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

© 2013 by Garland Science, Taylor & Francis Group, LLC

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means - graphic, electronic, or mechanical, including photocopying, recording, taping, or information storage and retrieval systems - without permission of the copyright holder.

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.

PowerPoint is a registered trademark of Microsoft Corporation in the United States and/or other countries.

Contents

Chapter 1	Why Phylogenomics Matters	1
Chapter 2	The Biology of Linear Molecules: DNA and Proteins	15
Chapter 3	Evolutionary Principles: Populations and Trees	35
Chapter 4	Databases	57
Chapter 5	Homology and Pairwise Alignment	71
Chapter 6	Multiple Alignments and Construction of Phylogenomic Matrices	93
Chapter 7	Genome Sequencing and Annotation	103
Chapter 8	Tree Building	117
Chapter 9	Robustness and Rate Heterogeneity in Phylogenomics	147
Chapter 10	A Beginner's Guide to Bayesian Approaches in Evolution	163
Chapter 11	Incongruence	185
Chapter 12	Adapting Population Genetics to Genomics	201
Chapter 13	Detecting Natural Selection: The Basics	219
Chapter 14	Refining the Approach to Natural Selection at the Molecular Level	233
Chapter 15	Genome-Level Approaches in Population Genetics	247
Chapter 16	Genome Content Analysis	263
Chapter 17	Phylogenomic Perspective of Biological Diversity: Tree of Life, DNA Barcoding, and Metagenomics	277
Chapter 18	Microarrays in Evolutionary Studies and Functional Phylogenomics	303
Index		325

Detailed Contents

Chapter 1 Why Phylogenomics Matters

Phylogenomics and Bioinformatics

- A microarray is a simple concept with powerful applications
- The Human Genome Project was a watershed event in DNA sequencing
- Bioinformatics tools enable data analysis and identification of patterns in biological experiments

The Rise of Phylogenomics

- Functional phylogenomics employs common ancestry to infer protein function
- Pattern phylogenomics gathers information from branching patterns of a group of organisms

The Phylogenomic Toolbox

- Sequence alignment programs and databases are essential computational tools in phylogenomics
- Statistical analysis enables comparison of biological sequences and generation of phylogenetic trees
- Parametric statistics are derived from the parameters of a given statistical distribution
- Nonparametric statistical analysis relies on resampling techniques to determine significance
- Maximum likelihood and Bayesian analysis are additional statistical methods used in phylogenomics

Key Attributes of Phylogenomicists

Summary

Discussion Questions

Further Reading

Chapter 2 The Biology of Linear Molecules: DNA and Proteins

Nucleic Acids

- DNA is a perfect molecule for transmitting information
- DNA is synthesized by specific pairing
- DNA can mutate and impart heritable information important in understanding descent with modification

1	Polymerase chain reaction is a milestone development	17
1	Proteins	17
3	Proteins are linear polymers of amino acids	17
3	Proteins have multiple levels of structure	18
3	Translation of the information in DNA is accomplished by the genetic code	20
3	A single nucleic acid sequence has multiple reading frames	22
4	The DNA Data Explosion	23
5	Sequencing methods for linear molecules have become more powerful	23
6	Next-generation sequencing allows small genomes to be sequenced in a day	24
7	Next-generation sequencing leads to practical applications	24
7	Alternatives to Whole Genome Sequencing: Detecting Variation in Populations	26
8	Single-nucleotide polymorphisms differ at one position	27
8	Methods that take advantage of DNA hybridization can be used to examine single-nucleotide polymorphisms	27
10	Analyzing Gene Expression	30
10	Northern blots assess RNA production on a small scale	30
11	Microarrays allow an entire transcriptome to be examined	31
12	Expressed sequence tags use reverse transcription to study RNA production	32
13	Summary	33
	Discussion Questions	33
	Further Reading	34
15	Chapter 3 Evolutionary Principles: Populations and Trees	35
15	Darwin and Evolutionary Theory	35
15	Four major contributions can be credited to Darwin	35
17	Darwin's research lacked a valid genetic or hereditary component	36

Evolution is now divided into microevolution and macroevolution	36	Discussion Questions	69
Microevolution	37	Further Reading	69
Microevolution is studied by population genetics	37	Chapter 5 Homology and Pairwise Alignment	71
Advances in molecular techniques led to new thinking in microevolution	37	Homology of Genes, Genomic Regions, and Proteins	71
Knowledge of codon changes and usage can provide evolutionary insights	38	Genomes can diverge by speciation and by duplication	71
Modern microevolutionary study relies on computational approaches and models	39	Alignment is a proxy for homology at the sequence level	72
Macroevolution	40	Nucleic acid sequence alignments can be evaluated manually	73
Macroevolution is studied by systematics	40	A paired protein alignment can be evaluated manually	75
The key to systematics is hierarchical arrangements	41	Dynamic Programming and Sequence Alignment	76
Competing philosophies and phylogenies form the basis of phylogenetics	41	Initialization sets up the matrix for alignment accounting	76
Modern phylogenetics is dominated by trees and tree thinking	42	Filling of the matrix is guided by preset rules	77
Modern phylogenetics aims to establish homology	43	Traceback uses the filled matrix to obtain the alignment	80
Species	46	Database Searching via Pairwise Alignments: The Basic Local Alignment Search Tool	82
The definition of species is a subject of debate	46	BLAST can identify inexact matches of high similarity	83
A simple solution is to define species phylogenetically	48	BLAST is optimized for searching large databases	83
Modern Challenges to Darwinian Evolution	50	There are several variations of BLAST for nucleotide and amino acid sequences	86
Punctuated equilibrium suggests that not all evolution is gradual	50	Performing a BLAST search is straightforward	86
Epigenetic changes caused by outside influences can be inherited	51	Whole genome alignments can also be performed	90
Markov chain models	51	Summary	91
Summary	54	Discussion Questions	91
Discussion Questions	55	Further Reading	92
Further Reading	56	Chapter 6 Multiple Alignments and Construction of Phylogenomic Matrices	93
Chapter 4 Databases	57	Multiple Sequence Alignment	93
Databases and Phylogenomics	57	Changing Alignment Parameters	97
DNA sequences are stored in large international databases	58	Multiple optimal alignments may exist	97
Specific data sets may be held in special repositories	58	Culling and elision are ways to explore the alignment space	98
Databases offer free access and availability for scientific inquiry	59	Choosing an Alignment Program	99
Information Retrieval from the NCBI Database	59	Automated alignment results are frequently adjusted "by eye"	99
Publications are archived in the PubMed database	60	Alignment programs can be compared by use of benchmark data sets	100
Sequences are collected in the GenBank database	60	Dynamic Versus Static Alignment	101
Whole genomes are collected on the Genome Page	66		
Other databases such as Online Mendelian Inheritance in Man (OMIM) can be used to obtain information	68		
Summary	69		

Summary	101	Rescoring is a simple way to weight characters in parsimony	129
Discussion Questions	102		
Further Reading	102		
 Chapter 7 Genome Sequencing and Annotation	 103		
Genome Sequencing	103		
Small viral genomes were the first to be sequenced	103		
Bacterial artificial chromosome-based sequencing employs a "divide and conquer" strategy for larger genomes	103		
The Human Genome Project	104		
Statistical Aspects of DNA Sequencing	105		
Lander–Waterman statistics tell how many sequencing reads are needed to cover a genome	105		
Sequence Assembly	106		
Next-Generation Sequencing	107		
Gene Finding and Annotation	109		
Gene finding can be accomplished via extrinsic, <i>ab initio</i> , and comparative approaches	109		
Gene functional annotation helps to determine gene function	110		
Gene ontology facilitates the comparison of genes	111		
A Phylo-View of Genomes Sequenced to Date	113		
Summary	114		
Discussion Questions	115		
Further Reading	116		
 Chapter 8 Tree Building	 117		
Distances, Characters, Algorithms, and Optimization	117		
The number of trees grows with each additional taxon	118		
Trees can be rooted by several methods	119		
Characters and Weighting	120		
Character states in molecular data may include the presence of genes and the sequence of nucleotides or amino acids	120		
Some discrete and numerical character states are ordered	121		
Characters can be weighted relative to one another	121		
Which characters should be used?	122		
A matrix of DNA sequences is used to illustrate different tree-building approaches	123		
Basics of Maximum Parsimony Analysis	124		
Fitch's algorithm uses set theory	125		
		Distance Methods	130
		Corrections for multiple hits may be introduced	131
		Corrections can be made by using evolutionary models	132
		Neighbor joining is a stepwise-based approach to tree building	134
		Minimum evolution uses minimal distance as a criterion to choose the best solution among multiple trees	134
		Basics of the Maximum Likelihood Approach	138
		Transformation and Probability Matrices	140
		Transformation cost matrices address the probability of character state changes in parsimony analysis	140
		Generalized probability matrices incorporate probabilities rather than integral weighting values in likelihood analysis	143
		Summary	144
		Discussion Questions	145
		Further Reading	145
		 Chapter 9 Robustness and Rate Heterogeneity in Phylogenomics	 147
		So Many Trees, So Little Time	147
		Tree space allows trees to be grouped by optimality	147
		Selection of a starting tree is the first challenge	149
		Peaks in tree space can be reached by branch swapping	149
		Moving from local optimality peaks to the true optimal tree is the most critical step in exploring tree space	150
		Rate Heterogeneity	151
		Rate heterogeneity can be included in likelihood models by use of a γ distribution	152
		The likelihood ratio test can be used to determine which models are more likely or probable than others	154
		Several programs can rapidly compare models	154
		Simple Metrics as Measures of Consistency in a Parsimony Analysis	155
		Tree length evaluates the degree of parsimony within a given solution	156
		Consistency index estimates convergence in a data set	156
		Retention index assesses degree of agreement with the maximum parsimony tree	157
		Rescaled consistency index addresses some shortcomings of retention and consistency indices	158

Decay index or Bremer index estimates robustness of a node	158	Character congruence uses total evidence or concatenation to create supermatrices	188
Determining the Robustness of Nodes on a Phylogenetic Tree by Resampling Techniques	158	Statistical methods assess incongruence to decide whether information should be concatenated	188
Bootstrapping analyzes a phylogenetic matrix by resampling with replacement	159	The incongruence length difference test uses parsimony to determine whether two gene partitions are incongruent	189
Jackknifing analyzes a phylogenetic matrix by resampling without replacement	160	Likelihood ratio tests compare likelihoods to determine whether two gene partitions are incongruent	191
Parametric bootstrapping applies a distribution model to the data	161	Fork indices provide measures of tree similarity	191
Summary	162	The Gene Tree/Species Tree Problem	191
Discussion Questions	162	Phylogeography adds a new dimension to population-level analysis	193
Further Reading	162	Lineage sorting was first observed in mtDNA	194
 Chapter 10 A Beginner's Guide to Bayesian Approaches in Evolution	 163	Three examples illustrate lineage sorting in studies of closely related taxa	194
Bayesian Inference	163	Genome-level examples of lineage sorting have also been documented	195
Generating a distribution of trees is an important application of the Bayesian approach	164	Coalescence offers a partial solution to the gene tree/species tree problem	196
MCMC is critical to the success of Bayesian analysis	167	Horizontal Transfer	197
Bayesian Analysis in a Phylogenetic Context	169	Programs That Consider Nonvertical Evolution and Lineage Sorting to Infer Phylogeny	197
Model selection can be utilized on any biologically meaningful partition	169	Coalescence programs use both gene trees and species trees as input	197
Selection of priors may involve default values, but priors can be adjusted manually	170	Programs that consider horizontal gene transfer generate nets and webs	198
Setting the MCMC parameters has greater impact for larger data sets	170	Summary	198
Increasing ngen in MCMC leads to better distribution of trees but at increased computational cost	172	Discussion Questions	199
An Example Using MrBayes	173	Further Reading	199
MrBayes can be run with default settings	174	 Chapter 12 Adapting Population Genetics to Genomics	 201
The MrBayes run can be interpreted by use of several parameters	176	Modernizing Population Genetics by Use of High-Throughput Methods	201
Models can be changed in MrBayes	177	Kimura and Lewontin contributed important new ways to think about genes in nature	202
Priors can be changed in MrBayes	179	The Hardy-Weinberg theorem has been extended in modern population genetics	202
Ending a Bayesian analysis involves an assessment of the run's efficiency	180	DNA Variation among Individuals	202
Summary	183	Single-nucleotide polymorphisms can be used to differentiate members of the same species	203
Discussion Questions	183	Microsatellites provide another analytical tool for species where SNPs are less abundant	204
Further Reading	183	Extending Fundamental Population Genetics	206
 Chapter 11 Incongruence	 185	Tajima's <i>D</i> distinguishes between sequences evolving neutrally and those evolving non-neutrally using allele frequencies	207
Incongruence of Trees	185		
Taxonomic congruence achieves consensus by construction of supertrees	187		

Corrections are needed in cross-species microarray studies	311
Protein-Protein Interactions	311
Various approaches are used to examine protein-protein interactions	311
Mendelian phenotypes in humans and model organisms are studied by Web-based approaches	312
Phylogenetic Approaches to Understanding Function	312
Phylogenomic gene partitioning is used to explore function	313
A gene presence/absence matrix was employed to compare 12 <i>Drosophila</i> genomes	316
Expressed sequence tags and phylogeny can be used to study plant function	320
Gene clustering in <i>Caenorhabditis elegans</i> was determined from RNA interference phenotype	321
Summary	322
Discussion Questions	323
Further Reading	323
Index	325

Why Phylogenomics Matters

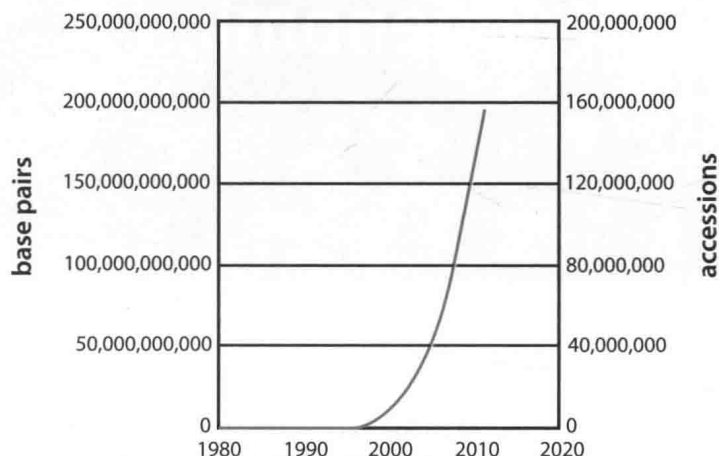
Phylogenomics is a new way of looking at biological information. It refers to the intersection of several important aspects of modern biology such as molecular biology, systematics, population biology, evolutionary biology, computation, and informatics, with genome-level information as the source for 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, ϕ X174, 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