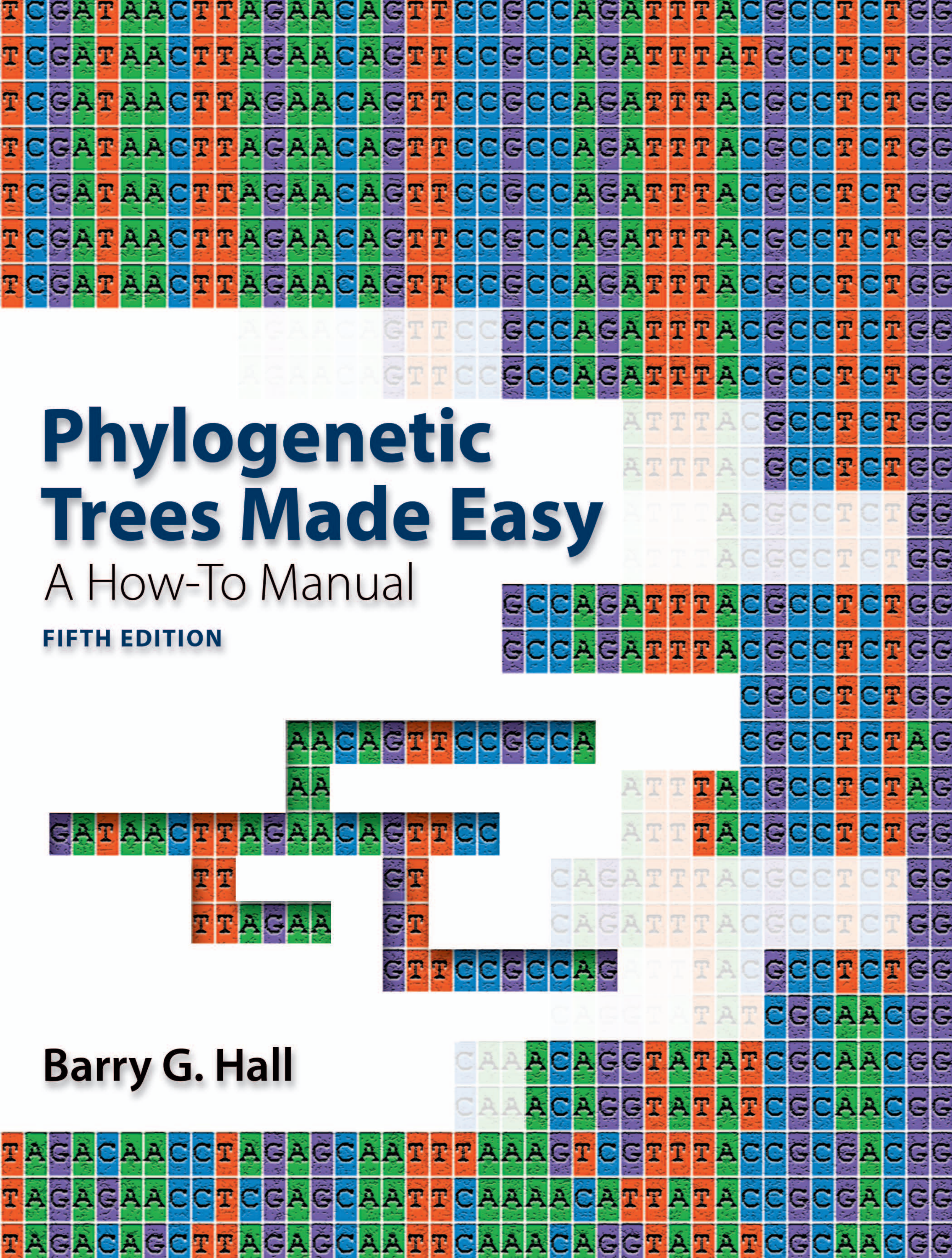


Phylogenetic Trees Made Easy

A How-To Manual

FIFTH EDITION

Barry G. Hall

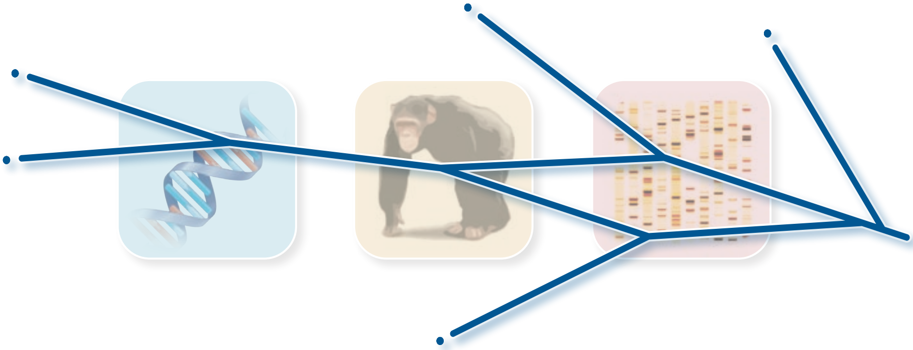
Phylogenetic Trees Made Easy

FIFTH EDITION

Phylogenetic Trees Made Easy

A HOW-TO MANUAL

FIFTH EDITION



Barry G. Hall

University of Rochester, Emeritus
and
Bellingham Research Institute

Sinauer Associates

NEW YORK OXFORD
OXFORD UNIVERSITY PRESS

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship, and education by publishing worldwide. Oxford is a registered trade mark of Oxford University Press in the UK and certain other countries.

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

Sinauer Associates is an imprint of Oxford University Press.

For titles covered by Section 112 of the US Higher Education Opportunity Act, please visit www.oup.com/us/he for the latest information about pricing and alternate formats.

Address editorial correspondence to:
Sinauer Associates
23 Plumtree Road
Sunderland, MA 01375 U.S.A.
publish@sinauer.com

Address orders, sales, license, permissions, and translation inquiries to:
Oxford University Press U.S.A.
2001 Evans Road
Cary, NC 27513 U.S.A.
Orders: 1-800-445-9714

Library of Congress Cataloging-in-Publication Data

Names: Hall, Barry G., 1942- author.
Title: Phylogenetic trees made easy : a how-to manual / Barry G. Hall, University of Rochester, emeritus, and Bellingham Research Institute.
Description: Fifth edition. | New York : Oxford University Press, [2018]
Identifiers: LCCN 2017026596 | ISBN 9781605357102
Subjects: LCSH: Phylogeny--Data processing.
Classification: LCC QH367.5 .H27 2018 | DDC 576.8/8--dc23
LC record available at https://lccn.loc.gov/2017026596

9 8 7 6 5 4 3 2 1

Printed in the United States of America

Downloadable files to be used with this text are available at
oup-arc.com/access/hall-5e

Notice of Trademarks
Throughout this book trademark names have been used and depicted, including but not limited to Macintosh, Mac, Windows, and Adobe. In lieu of appending the trademark symbol to each occurrence, the author and publisher state here that these trademarked product names are used in an editorial fashion, to the benefit of the trademark owners, and with no intent to infringe upon the trademarks.

To Miriam Barlow and John Logsdon

Acknowledgments

The dedication of this book requires some explanation:

Professor Miriam Barlow is in large part responsible for the existence of the First Edition of *Phylogenetic Trees Made Easy* (*PTME*). In 2000, Miriam was a Ph.D. student in my lab and had just successfully completed a course in Phylogenetic Analysis with Dr. John Huelsenbeck, but she was struggling with the mechanics of actually using software to make a phylogenetic tree. I gave her a fistful of notes I had written to myself when trying to learn phylogenetics a year earlier. She shared them with some fellow students who were similarly struggling, and soon came back to me virtually demanding that I complete them as a book. I did, and *PTME1* was the result.

Professor John Logsdon is similarly responsible for this, the Fifth Edition of *PTME*. John had been an enthusiastic supporter of earlier editions of *PTME*, and he contacted me in the Fall of 2016 asking when a new edition would be available, or more specifically, whether it would be available in time for a course he would teach in 2018. I hadn't thought much about another edition of *PTME*, but after some discussion I concluded that I could add sufficient new material to justify a new edition. This manual is the outcome of those discussions.

So, thanks to Miriam and John for being the bookends that initiated and concluded the *PTME* series. I would not have done it without you.

As always, writing *PTME5* depended on the active cooperation of the authors of some of the software described herein. My thanks to Sudhir Kumar for explaining the implementation of Timetrees in MEGA7, and to Glen Stecher of the MEGA team for responding so quickly to my queries and bug reports. My deep thanks to Andrew Rambaut for his help while learning BEAST and its associated programs. My thanks to Steven Salipante for his collaboration on MSTgold. Steve began as an undergraduate in my lab and is now a Professor at the University of Washington. Working with him has always been a joy.

Author John Scalzi once pointed out that a book is much more than the words written by the author; it is also the result of the efforts by the entire publication team. That is certainly the case for this edition of *PTME*.

Kathaleen Emerson, my Editor, has greatly improved my words and not just by catching typos and spelling errors. She has spotted inconsistencies between text and figures and rearranged the order of paragraphs to improve clarity. Most of all, she has played the role of "generic target reader" and made sure that the text always keeps that reader in mind. She is greatly responsible for making this book the best it can be.

Chris Small and Donna DiCarlo are responsible for the visual appearance of the book, its "look and feel." I am hopeless at anything visual, so I am very grateful to Chris and Donna for making *PTME5* not only informative but aesthetically pleasing.

Marie Scavotto is responsible for marketing and it is thanks to her efforts on this and earlier editions that *PTME* can be sold at a reasonable price. If she doesn't bring *PTME* to the attention of enough buyers it can't be sold at an affordable price, so I am very grateful to her for her efforts.

Last, as always, Andy Sinauer has been tremendously supportive of every edition of *PTME*. When I first thought of writing *PTME* I asked colleagues whom I should contact about publishing it. Sinauer Associates was the universal suggestion, and the universal comment about Andy was, "You can trust him absolutely." I have found that to be some of the best advice I ever received. Andy is both trustworthy and highly effective. He plans to retire in the near future, and he will be tremendously missed.

Finally, my deepest thanks to my wife of 53 years, Sue Hall, for her patience and encouragement while writing *PTME5*.

Table of Contents

Chapter 1 ▪ Read Me First! 1

New and Improved Software 2

Just What Is a Phylogenetic Tree? 2

Estimating Phylogenetic Trees: The Basics 4

Beyond the Basics 5

Learn More about the Principles 7

About Appendix VI: F.A.Q. 7

Computer Programs and Where to Obtain Them 8

MEGA 7 8

BEAST 8

FigTree 8

codeml 9

SplitsTree and Dendroscope 9

Graphviz 9

Utility Programs 9

Text Editors 9

Acknowledging Computer Programs 9

The *Phylogenetic Trees Made Easy* Website 10

Chapter 2 ▪ Tutorial: Estimate a Tree 11

Why Create Phylogenetic Trees? 11

About this Tutorial 12

Macintosh and Linux users 12

A word about screen shots 12

Search for Sequences Related to Your Sequence 13
Decide Which Related Sequences to Include on Your Tree 18
Establishing homology 19
To include or not to include, that is the question 20
Download the Sequences 22
Align the Sequences 27
Make a Neighbor Joining Tree 28
Summary 32

Chapter 3 ▪ Acquiring the Sequences 33
Background 33
Problems arising from the vast size of the sequence databases 33
The query sequences 34
Hunting Homologs: What Sequences Can Be Included on a Single Tree? 34
Becoming More Familiar with BLAST 35
BLAST help 36
Using the Nucleotide BLAST Page 36
Using BLAST to Search for Related Protein Sequences 39
Finalizing Selected Sequences for a Tree 43
Problems adding coding sequences of protein homologs to the Alignment Explorer 43
Adding Sequences to and Removing Sequences from the Alignment Explorer 47
Add a sequence 47
Import a file of sequences 48
Delete a sequence 48
Other Ways to Find Sequences of Interest (Beware! The Risks Are High) 48

Chapter 4 ▪ Aligning the Sequences 53
Aligning Sequences with MUSCLE 53
Examine and Possibly Manually Adjust the Alignment 57
Trim excess sequence 57
Eliminate duplicate sequences 58
Check Average Identity to Estimate Reliability of the Alignment 60
Codons: Pairwise amino acid identity 60
Non-coding DNA sequences 61
Increasing Alignment Speed by Adjusting MUSCLE's Parameter Settings 62
How MUSCLE works 62
Adjusting parameters to increase alignment speed 63

Aligning Sequences with ClustalW 64
Aligning Sequences with GUIDANCE2 65
Viewing the results 67
Eliminate unreliable parts of the alignment 68
Saving the GUIDANCE2 alignment 70

Chapter 5 ▪ Major Methods for Estimating Phylogenetic Trees 71
LEARN MORE ABOUT TREE-SEARCHING METHODS 72
Distance versus Character-Based Methods 74
LEARN MORE ABOUT DISTANCE METHODS 74

Chapter 6 ▪ Neighbor Joining Trees 77
Using MEGA7 to Estimate a Neighbor Joining Tree 77
LEARN MORE ABOUT PHYLOGENETIC TREES 78
Determine the suitability of the data for a Neighbor Joining tree 80
Estimate the tree 81
LEARN MORE ABOUT EVOLUTIONARY MODELS 83
Unrooted and Rooted Trees 88
Estimating the Reliability of a Tree 91
LEARN MORE ABOUT ESTIMATING THE RELIABILITY OF PHYLOGENETIC TREES 92
What about Protein Sequences? 99

Chapter 7 ▪ Drawing Phylogenetic Trees 101
Changing the Appearance of a Tree 101
The Options dialog 104
Branch styles 108
Fine-tuning the appearance of a tree 109
Rooting a Tree 112
Finding an outgroup 112
Subtrees 114
Saving Trees 118
Saving a tree description 118
Saving a tree image 119
Captions 119

Chapter 8 ▪ Parsimony 121
LEARN MORE ABOUT PARSIMONY 121

MP Search Methods 123
Using SeaView for Parsimony 125
Estimating a bootstrap tree in SeaView 133
Using MEGA to draw the tree estimated by SeaView 136

Chapter 9 ■ Maximum Likelihood 141
LEARN MORE ABOUT MAXIMUM LIKELIHOOD 141
ML Analysis Using MEGA 143
Test alternative models 144
Estimating the Reliability of an ML Tree by Bootstrapping 148
What about Protein Sequences? 149

Chapter 10 ■ Bayesian Inference of Trees Using BEAST 151
BEAST: An Overview 151
Installing BEAST 152
Prepare the Input Alignment File 152
Run BEAUti 153
LEARN MORE ABOUT BAYESIAN INFERENCE 158
Running BEAST 162
Run Tracer 165
Burnin 166
Run TreeAnnotator 168
What about Protein Sequences? 169
Visualizing the BEAST Tree 170
The icons above the tree 172

Chapter 11 ■ Which Method Should You Use? 175
Criteria to Consider 175
Accuracy 176
Ease of interpretation 176
Time and convenience 176
Results of the Major Methods 178

Chapter 12 ■ Working with Various Computer Platforms 183
Command-line Programs 183
MEGA on the Macintosh Platform 184
Navigating among folders on the Mac 184
Printing trees and text from MEGA 188

The Line Endings Issue 188
Running the Utility Programs 189

Chapter 13 ▪ Phylogenetic Networks 191
Why Trees Are Not Always Sufficient 191
Unrooted and Rooted Phylogenetic Networks 192
LEARN MORE ABOUT PHYLOGENETIC NETWORKS 193
Using SplitsTree to Estimate Unrooted Phylogenetic Networks 198
Estimating networks from alignments 198
Rooting an unrooted network 204
Estimating networks from trees 205
Consensus networks 206
Supernetworks 210
Using Dendroscope to Estimate Rooted Networks from Rooted Trees 211

Chapter 14 ▪ Minimum Spanning Trees 215
Minimum Spanning Trees Are *Not* Phylogenetic Trees! 215
Why Use Minimum Spanning Trees? 215
Origin of MSTs and the Issue of Reliability 216
What is a minimum spanning tree? 216
Using MSTgold to Estimate MSTs 217
The MSTgold input files 218
Two ways for MSTgold to calculate the initial distance matrix 219
Running MSTgold with the ebgC data 219
The MSTgold output 220
Bootstrapping MSTgold 223
Exporting MSTs from Graphviz 225
An Alternative Data Set to Illustrate Some Additional Features of MSTgold 226
An Alternative to Graphviz: Hypercube 229

Chapter 15 ▪ Time Trees 231
Preparations to Estimate a Time Tree 231
Estimating a Time Tree 234
Viewing the Relative Time Tree 238
An Absolute Time Tree 240
Effect of more calibration points on absolute time trees 244
Postscript 244

Chapter 16 ▪ Reconstructing Ancestral Sequences 245
Using MEGA to Estimate Ancestral Sequences by Maximum Likelihood 246
Create the alignment 246
Construct the phylogeny 246
Examine the ancestral states at each site in the alignment 247
Estimate the ancestral sequence 250
Calculating the ancestral protein sequence and amino acid probabilities 256
How Accurate Are the Estimated Ancestral Sequences? 257

Chapter 17 ▪ Detecting Adaptive Evolution 259
Effect of Alignment Accuracy on Detecting Adaptive Evolution 261
Using MEGA to Detect Adaptive Evolution 261
Why ebgC *is an interesting example for considering adaptive evolution 261*
Detecting overall selection 261
Detecting selection between pairs 263
Finding the region of the gene that has been subject to positive selection 264
Using codeml to Detect Adaptive Evolution 266
Installation 267
Run codeml 269
Questions that underlie the models 273
A closer look at ebgC_1.out 275
Summary 278
Postscript 278
Compiling PAML yourself and running codeml without PAMLX 278

Chapter 18 ▪ Estimating Phylogenetic Trees from Whole Genome Sequences 281
The Pan-Genome Problem 282
kSNP: An Alternative to Genome Alignment 282
kSNP 2.0 284
kChooser 284
FCK 285
Tree accuracy 286
Always run kChooser before running kSNP 286
Using kSNP3 286
The steps in estimating phylogenetic trees from WGS using kSNP3 287

Chapter 19 ▪ Some Final Advice: Learn to Program 297

Appendix I ▪ File Formats and Their Interconversion 299

Format Descriptions 299

The MEGA format 299

The FASTA format 300

The Nexus format 301

The PHYLIP format 304

Interconverting Formats 305

FastaConvert, MEGA, and SeaView 305

Appendix II ▪ Text Editors 307

Mac OS X: TextWrangler 307

Windows: Notepad++ 308

Linux: Gedit 308

Appendix III ▪ The Command-line Environment 311

Introduction and History 311

Terminal and Command Prompt: The Apps for Accessing the Command-line Environment 312

The current directory 313

Entering Commands 313

Navigating in Terminal 314

The Magic Trick 315

What Is in the CWD? The **ls** Command (§ **dir**) 316

The -l option 317

The -a option 317

The permissions column 318

The **chmod** *command 318*

§The **dir** Command in Command Prompt 319

Some Other Important Commands 320

Copy a file 320

Move a file 321

Rename a file 321

Make a directory 321

Remove a directory 321

Remove a file 321

Print the contents of a file to the screen 321

Clear the screen 321

Appendix IV ▪ Installing and Running Command-line Programs 323

Installing Command-line Programs 323

Mac OS X / Linux 323

Windows 324

Running Command-line Programs 329

An example 330

The line endings issue 330

The example command line 330

The output file(s) 330

Error checking 330

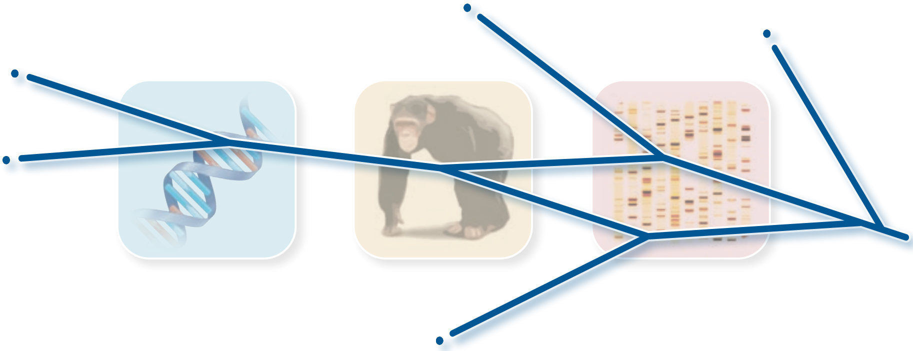
Appendix V ▪ Additional Programs 331

Appendix VI ▪ Frequently Asked Questions 335

Literature Cited 339

Index to Major Program Discussions 341

Subject Index 345



CHAPTER 1

Read Me First!

Phylogenetic analysis was once a tool used only by taxonomists, or as they are now called, systematists. Their interests were first in classification, then in elucidation of the historical relationships among organisms. Phylogenetics was considered an arcane and difficult aspect of biology that, fortunately, wasn't widely used by molecular and biochemically oriented biologists. With the advent of DNA sequencing that began to change, and the power of phylogenetics as a tool for understanding biology at all levels became apparent. Today there are few biological journals in which at least some of the published papers do not include phylogenetic trees.

Phylogenetic analysis lies at the core of the fields of genomics and bioinformatics. The emergence of bioinformatics as an important field, however, has had little effect on the perception of phylogenetic analysis as difficult and intimidating. Those who specialize in systematics, phylogenetics, genomics, and bioinformatics have done little to set aside the sense that phylogenetic analysis is a tool best used by specialists, with the result that molecular and cell biologists often seek collaboration with systematists when one of their papers requires a tree. But in most cases, preparing a robust, valid phylogenetic tree for a paper is no more challenging than using word processing software to write the paper or graphics software to prepare the figures. The purpose of this book is to make phylogenetic analysis accessible to all biologists.

Phylogenetic Trees Made Easy is a "cookbook" intended to aid beginners in creating phylogenetic trees from protein or nucleic acid sequence data. It assumes basic familiarity with personal computers and with accessing the web using web browsers. I have not attempted to explore all the alternative approaches that might be used, intending only to give the beginner an approach that will work well most of the time and is easy to carry out. I hope the book will also serve the investigator who has a modest familiarity with phylogenetic tree construction but needs to address some aspects and problem areas in more depth.

This book is not intended to be used as the primary text in systematics or phylogenetics courses. It can, however, be used to supplement the primary text and can serve as a tool for making the transition between a theoretical understanding of phylogenetics and a practical application of the methodology.

New and Improved Software

The six years that have elapsed since I wrote the Fourth Edition have seen significant changes in MEGA, the software for phylogenetic analysis and tree construction that is at the core of this book. MEGA has changed significantly both in terms of increased capabilities and in terms of its interface. Details of using some other programs have also changed significantly. In general, the changes have been positive and especially beneficial to those who are new to phylogenetics.

It is neither possible nor desirable to cover all of the programs available for phylogenetic analysis; that would require a tome too heavy to lift and too boring to read. The objective here is to describe programs that are sufficient to the task and relatively easy to use. Because I focus more on phylogenetic analysis and less on learning the vagaries of different programs, I try to minimize the number of programs described, and, as time goes on, my perception of the most appropriate programs changes.

The Fifth Edition continues the practice of updating software to reflect the interface of the current version and of adding value in the form of additional advanced chapters. MEGA7 adds two important new functions: Time tree estimation and Batch processing.

The Third and Fourth Editions each introduced “advanced” topics that went beyond just estimating phylogenetic trees per se. In keeping with that practice, the Fifth Edition adds a chapter on estimating phylogenetic trees from whole genome sequences without aligning those sequences. That chapter will be particularly important to those who consider phylogenies of bacteria and viruses based on whole genome sequences. It also adds a chapter on Minimum Spanning Trees (Chapter 14), an alternative to phylogenetic trees when too much homoplasy has reduced the phylogenetic signal below an acceptable level. Finally, because so many valuable and important (and mostly free) programs are command-line programs that do not use the familiar GUI interface, I have added an appendix on the command-line interface and an appendix on installing and running command-line programs.

To keep things simple, and to make my life easier, I will from now on refer to the Fifth Edition of *Phylogenetic Trees Made Easy* as “*PTME5*.”

Just What Is a Phylogenetic Tree?

To a mathematician, a phylogenetic tree is an abstract construct embodied in a special type of directed or undirected graph. To systematists and most evolutionary biologists, a tree is a representation of the relationships of different species to their ancestors. To a molecular biologist, a tree is a representation of the relationships of gene or protein sequences to their ancestral sequences. All agree, however, that a tree consists of nodes connected by branches (except that mathematicians tend to refer to branches as edges).

The tips of a tree, sometimes referred to as **leaves**, are the **external nodes** and biologically they represent existing taxa. Systematists and evolutionary biologists generally think of taxa as being synonymous with species, while molecular biologists think of taxa as being synonymous with sequences. In either case, these leaves/taxa/sequences are the only entities in a tree that we can be sure about. They represent the real data—factual information—from which everything else in the tree is inferred. That factual information can be the states of morphological characters or it can be the states of nucleotide or amino acid characters in macromolecular sequences. The discussion in this book will be limited to molecular sequence characters.

Even with sequence data, however, much is assumed. We may use sequences of the cytochrome *c* gene to make a tree in which each sequence is from a different species, so we may refer to the tips as the "cow sequence," "dog sequence," etc. In fact, the cow sequence is not the cytochrome *c* sequence for all cows; it is the sequence of that gene from one particular cow. We can use that sequence to represent all cows because we assume (1) that any variation in cow sequences occurred after cows diverged from dogs, and (2) that any one randomly selected cow sequence is as representative of cows as the next. Both these assumptions arise from a fundamental assumption of phylogenetics, that of genetic isolation—that there is no genetic exchange between taxa. When this fundamental assumption is violated, different parts of the information used to make the tree (gene sequence, set of sequences, full genome) can have different histories and valid trees cannot be constructed. This issue becomes particularly important when we try to estimate a phylogeny based on whole genome sequences. For microorganisms it is almost always the case that different parts of a genome have different evolutionary histories. Two alternatives to phylogenetic trees, minimum spanning trees and phylogenetic networks, are less disrupted by incomplete genetic isolation.

The interior nodes of the tree represent *hypothetical* ancestors. We don't know the character states of those ancestors and in most cases cannot know those characteristics because the ancestors no longer exist. We can only infer the characteristics of ancestral sequences from information about their descendants (the leaves of the tree). We assume a process of modification with descent in which mutations accumulate and are inherited by their descendants. **Branches** connect nodes and represent that descent, and **branch lengths** represent the amount of change between an ancestor and its descendant.

The most important fact to keep in mind about any phylogenetic tree is that it is not fact, it is an estimate—an inference. Many evolutionary biologists would say that a tree is a hypothesis, but I disagree. A hypothesis is useful only if it can be tested, and we cannot test the validity of evolutionary trees. We can estimate trees, we can estimate how accurate they might be, and we can change our estimates (and the trees) when new data are available, but that's it. An estimate is a fine and useful thing; just keep in mind that it is not absolute truth.

Estimating Phylogenetic Trees: The Basics

Chapters 2–10 of *PTME5* cover the basics of four distinct and equally important steps that are involved in making a phylogenetic tree based on molecular sequence data:

1. Identify and acquire the sequences that are to be included on the tree.
2. Align the sequences.
3. Estimate the tree by one of several methods.
4. Draw the tree and present it to an intended audience.

Chapters 2–10 describe implementing these four steps using the MEGA7 software package. The data acquisition, alignment, and tree-drawing functions of MEGA7 are so elegantly implemented and easy to use that this is the program of choice for most phylogenetic methods. Understanding these chapters will permit confident estimation of valid phylogenetics trees by a well-accepted, reliable method.

Chapter 2 is a tutorial that takes the reader through each step, using MEGA7 to construct a simple phylogenetic tree. The primary purpose of Chapter 2 is to familiarize you with both the mechanics of implementing the steps necessary to make a tree and the basics of the MEGA7 software. Chapter 2 will *not* lead to the optimal tree based on the data provided, so details of each step are covered in the subsequent chapters.

Chapter 3 deals in more detail with Step 1: identifying the sequences that might be included on a tree, deciding which to include and which to exclude, and downloading those sequences from international databases operated by government agencies.

Chapter 4 deals with the critical problem of aligning sequences (Step 2). Sequence alignments, whether of nucleotides or proteins, are the data upon which phylogenetics programs operate to estimate a tree. If you don't get this part right, nothing you do in later steps will matter; your tree will be worthless.

Chapter 5 very briefly discusses and compares the major methods for estimating phylogenetic trees. Chapter 6 then deals with Neighbor Joining (NJ), the most widely used tree estimation method, and discusses the often intimidating issue of bootstrapping to assess tree reliability.

In many ways, Chapter 6 is the meat of this book. The major phylogenetic concepts are covered in this chapter. Subsequent chapters dealing with other phylogenetic methods are mainly technical and simply illustrate how to accomplish the same goals with other methods and software.

Chapter 7 describes how to draw trees in various ways and how to choose which style will make it easiest for your audience to correctly interpret the tree and understand the biological point you are making.

Chapter 8 discusses using MEGA7 and SeaView to estimate trees by Maximum Parsimony (MP). MP is one of the first methods to have been applied to molecular sequence data.

Chapter 9 describes how to use MEGA to estimate trees by Maximum Likelihood (ML). ML is a statistical method that has long been appreciated, but was not widely used outside of the fields of systematics and phylogenetics because it had been perceived as being intimidating, slow, and not applicable to large data sets. Recent advances in software have solved the speed problem and have made ML applicable to large data sets. I hope that after reading Chapter 9 you will find ML no more intimidating than NJ or MP.

Chapter 10 discusses constructing trees by Bayesian Inference (BI) using BEAST. Like ML, BI is a powerful statistical method and it has proven to be slightly more accurate than the other methods (Hall 2005).

Chapter 11 considers ways to decide which of the major methods to use.

Chapter 12 discusses some issues that arise when working with MEGA7 on its non-native platforms, Mac OS X and Linux.

TABLE 1.1

Some Conventions Used in this Book

Convention	Description
Click	Use the mouse to position the cursor over the indicated button on the screen (as in "click the OK button"), then depress and quickly release the mouse button.
Double-click	Click twice rapidly without moving the mouse.
Drag	Position the mouse and, while holding down the mouse button, move the mouse to another position
Select	Highlight a menu item, section of text, or an object on the screen by dragging the mouse over or clicking (or double-clicking) on the desired operation. In the text these operations are indicated in **Blue, Bold Sans Serif** type.
Screen display	Text in **Black, Bold Sans Serif** type indicates information you will see on your screen, but no action is required.
Enter	For command line programs such as MSTgold and kSNP3 and when using this book's utility programs, text shown in the `Courier typeface` indicates commands that you will type into an input file (or see on the screen if you have downloaded a utility file).
Download	Indicates a file in the package that you can download from the PTME5 website

Beyond the Basics

Those readers without previous experience with phylogenetics will probably find the information in Chapters 2–12 sufficient to meet their needs for some time to come. Increased experience with phylogenetics, however, often leads to recognition of the utility of more advanced aspects of phylogenetic analysis. Some of those advanced topics are discussed in Chapters 13–19.

Chapter 13 concerns phylogenetic *networks* as opposed to phylogenetic trees. In some circumstances phylogenetic trees are an inadequate means of describing the historical relationships of taxa. For instance, trees based on different genes from the same set of species are often significantly different because those genes have different evolutionary histories as the result of recombination, horizontal transfer, or other "reticulate" events. In such cases phylogenetic networks more realistically describe evolutionary history. An interior node of a phylogenetic tree may have multiple nodes descending from it, but it can have only one immediate ancestral node. In a phylogenetic network, a node can have both multiple descendants and multiple immediate ancestors.

Chapter 14 discusses another alternative to phylogenetic trees, minimum spanning trees (MSTs). MSTs do not attempt to estimate relationships of taxa to hypothetical common ancestors, only relationships of taxa to each other. They consider identity by state, not identity by descent, and thus really only reveal clustering. In effect, they only ask, "Which sequences are most alike?," ignoring whether they are alike because of descent from a common ancestor, because of convergent evolution, or because of recombination. When phylogenetic signal is weak, MST can still offer some insights into relative similarity of sequences. Chapter 14 discusses the program MSTgold for estimating MSTs.

Chapter 15 discusses time tree estimation. Time tree estimation means estimation of the divergence times of interior nodes based on outside information about the times at which a few of the interior nodes diverged. Time trees are becoming increasingly frequent and important in the literature.

Chapter 16 deals with the reconstruction of the amino acid sequences of ancient ancestral proteins and nucleotide sequences of ancestral genes. This subject, which some call *paleobiology*, permits the estimation of ancestral sequences as a necessary prelude to actual synthesis of ancestral proteins. Biochemists and others interested in understanding protein function often have an interest in synthesizing, then studying, a protein they believe to be the common ancestor of a group of proteins with distinct, but mechanistically related, functions. The first step in the process is to estimate a reliable phylogenetic tree, and the second step is to estimate the most likely sequence of that common ancestor.

Chapter 17 deals with another advanced topic, detecting *adaptive evolution* (i.e., purifying or diversifying selection) along phylogenetic trees.

Chapter 18 discusses the problems associated with estimating phylogenetic trees from whole genome sequences (WGS). Because microbial genomes of the same species are replete with rearrangements, major deletions, and major insertions it is impractical to align more than about 35–40 WGS of microorganisms. There are now tens of thousands of microbial WGS in the databases, and there is a need for practical ways to estimate phylogenies of hundreds of such genomes. This chapter discusses an alternative to genome alignment and its implementation in the program kSNP3 (Gardner and Hall 2013; Gardner et al. 2015).

Chapter 19 discusses why it is important for you to learn at least the rudiments of programming and offers some suggestions about how to approach that learning.

Appendix I discusses file formats and their interconversion.

Three new appendices discuss the command-line environment. The command-line is a very old interface that predated the current GUI with its windows, menus, mouse, etc. Most free academic programs, such as MSTgold, and kSNP3, use the command-line environment and it is important to become comfortable in that environment. Appendix II discusses text editors that are used to make input files for command-line programs, and to read their output files.

Appendix III discusses navigating the command-line environment. Appendix IV describes how to install and run command-line programs.

Appendix V briefly discusses additional programs that are often useful.

Appendix VI answers some frequently asked questions—or at least some questions that should be frequently asked.

Learn More about the Principles

Just as it is possible to implement molecular methods without understanding them by following the protocols in commercial "kits," it is also possible to implement phylogenetic methods without understanding them by following the protocols in this book. Most of us insist that our students understand the principles underlying the methods implemented by these kits because we know that without such an understanding it is impossible to spot and troubleshoot many problems.

It is in this spirit that the reader will find *Learn More* boxes scattered throughout the text. These boxes present somewhat more detailed background on the biological and/or mathematical principles underlying the various methods and suggest further reading. It is not necessary to read the boxes to be able to estimate a valid phylogenetic tree, but understanding these principles can help troubleshoot the phylogenetic problems that arise when estimating trees from molecular alignment data.

Readers who want to go beyond the *Learn More* boxes will find Dan Graur and Wen-Hsiung Li's *Fundamentals of Molecular Evolution* (2000) and Masatoshi Nei and Sudhir Kumar's *Molecular Evolution and Phylogenetics* (2000) very helpful and enjoyable. Li's *Molecular Evolution* (1997) and Chapters 11 (Swofford et al. 1996) and 12 (Hillis et al. 1996) of *Molecular Systematics* (David Hill, Craig Moritz, and Barbara Mable, eds.) provide more detailed insights into these topics. Joe Felsenstein's outstanding book, *Inferring Phylogenies* (2004), is very technical, very mathematical, and not for the faint of heart. At the same time, it is delightfully written and well worthwhile for anyone who really wants to understand phylogenetic theory. *Phylogenetic Networks*, by Daniel Huson, Regula Rupp, and Celine Scornavacca (2011) provides an extremely detailed explanation of the concepts, algorithms, and applications of this relatively new approach (described briefly in Chapter 13 of this book). Like Felsenstein's book, *Phylogenetic Networks* is highly technical and mathematical, but it is also a very worthwhile presentation of this recent alternative to phylogenetic trees for understanding relationships among taxa.

About Appendix VI: F.A.Q.

Well, not really "Frequently Asked Questions," but "Sometimes Asked Questions," or more likely, "Questions That Should Be Asked but Probably Aren't."

This book is organized and presented as though you will use MEGA whenever possible for every step in the process of estimating and presenting a tree. What if you want to use MEGA to download sequences, but you want to use another program for alignment? What if you have estimated a tree using other programs, but you want to use MEGA to draw and present that tree? The more you use phylogenetics the more likely you are to want to explore and use alternative programs. Such issues can interrupt the flow of an individual chapter and be distracting, so I have collected them in Appendix VI. Look there first for tips and hints. I hope that these "What if?" questions will both help and encourage you to expand your horizons beyond the methods that are presented in this book.

Computer Programs and Where to Obtain Them

Programs are continuously being updated. Readers should always download the latest versions from the sources listed below. Readers should also be aware that new versions might have screens that differ from the examples in this book, or they may accept commands in slightly different ways. Usually the modifications are minor and will present few problems to the careful user.

MEGA7

MEGA7 (Kumar et al. 2016) is a modern, integrated phylogenetics package that elegantly handles the following: downloading sequences through its own specialized web browser; aligning sequences through its implementations of ClustalW and MUSCLE; estimating phylogenetic trees by a variety of methods including Neighbor Joining (NJ), Maximum Parsimony (MP), and Maximum Likelihood (ML); and drawing those trees in a variety of ways. MEGA7 is free and can be downloaded from www.megasoftware.net.

BEAST

BEAST(Bayesian Evolutionary Analysis by Sampling Trees) is used for Bayesian Inference (BI) of phylogenetic trees (Drummond et al. 2012). The BEAST home page beast.bio.ed.ac.uk/ provides helpful information and links to the download site github.com/beast-dev/beast-mcmc/releases/tag/v1.8.4. BEAST is free and is available for Macintosh, Linux, and Windows computers.

The current version is 1.84. The BEAST package includes several programs and some documentation. The BEAST Manual and Practical Beast in the Docs folder are helpful, but they have not been revised in some time so many of the screen shots are not consistent with the current version.

FigTree

FigTree is an excellent tree-drawing program. It reads tree files in both Nexus and Newick formats. It also has an extended Nexus format that includes such features as fonts, branch colors, etc. FigTree is the default program for drawing the consensus tree written by BEAST. Like MEGA's tree-drawing program, FigTree permits re-rooting, rotating clades around a branch, showing or hiding branch lengths, and other features. The current version is 1.4.3. Download from tree.bio.ed.ac.uk/software/figtree/.

codeml

Codeml is part of the free PAML (Phylogenetic Analysis by Maximum Likelihood) package. You can download PAML for Macintosh, Windows, and Unix/Linux from abacus.gene.ucl.ac.uk/software/paml.html. The current versions of PAML are 4.8a for Mac OS X and Linux, and 4.9.d for Windows. This book discusses the use of PAMLX, a graphical user interface for PAML (Xu and Yang 2013).

SplitsTree and Dendroscope

SplitsTree and Dendroscope are free programs for estimating and drawing phylogenetic networks. They are available from www-ab.informatik.uni-tuebingen.de/software/splitstree4/welcome.html and www-ab.informatik.uni-tuebingen.de/software/dendroscope/welcome.html, respectively.

Graphviz

Graphviz is a program that draws minimum spanning tree graphs from .dot files. It is required to display and print the minimum spanning trees that are discussed in Chapter 14. Graphviz is freely available from www.Graphviz.org/Download.php.

Utility Programs

I have provided a couple of utility programs that are very helpful for changing file formats and for reconstructing the sequences of ancestral proteins and nucleic acids. Those programs, along with various examples, templates, and so forth can be downloaded from the *Phylogenetic Trees Made Easy* website (oup-arc.com/access/hall5e). Appendix IV gives details on installing and using these programs.

Text Editors

Users who intend to use codeml, kSNP3, MSTgold, or the utility programs will need a text editor program to prepare the necessary input files. Word processor files such as those written by Microsoft Word, WordPerfect, etc. will not work. See Appendix II for recommended text editors and instructions on how to use them.

Acknowledging Computer Programs

Phylogenetic analysis is completely dependent upon computer programs; without those programs, molecular phylogenetics simply would not be possible. The programs described in this book are free—at least they are free to us, the users. The authors of these programs have paid dearly for them in terms of time, effort, and creativity. The only reward the authors get for their effort (and it is an enormous effort) is having that software cited in publications. Please be sure to cite the software you use just as rigorously as you cite published papers. You can find the appropriate citation in the program itself, in the documentation, or at the program's website. If you can't find a paper, simply cite