

Read Me First

This is a "cookbook" intended as a tool 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 World Wide Web using browsers such as Netscape or Microsoft Explorer. 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.

Whatever the user's level of experience, this book devotes a significant amount of attention to the problem of aligning proteins and nucleic acids. Although one of the major purposes of alignments is to create phylogenetic trees, it is far from the only important application of alignments. Studying alignments of a series of related proteins can provide insight into possible active sites by identifying those regions that are most strongly conserved and that contain probable catalytic residues. Thus those investigators whose interests are more in protein function than in phylogeny *per se* are also the intended audience for this book.

This book is not intended to be used as a primary text in a systematics or phylogenetics course, and it is not appropriate for that purpose. It can, however, be used as a supplement to 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.

A Brief Overview of the Second Edition

As pointed out in the first edition, software changes quickly, and some screens may well look different from those depicted in this book. Indeed, within a few weeks of publication of the first edition, the appearance of the Web-based BLAST screens were substantially different. It is not only cosmetic appearances that change. Programs change the names of input parameters, what is contained within various output files, etc. Such changes are intended to simplify things for users and are often part and parcel of improving a program's functionality, but they do make it awkward for a beginner trying to follow a book in order to learn both a method and the software to implement that method. One of the major motives for this edition was to make the book current with respect to the software as of the end of 2003.

When a simple repair of a shower in my home revealed serious and extensive dry rot my contractor said, "You know, you don't *have* to put this bathroom back exactly the way it was. If you are ever going to remodel it, this is the time." That turned out to be good advice. In the same sense, I decided that if I was ever going to reorganize this book to make it more useful, this is the time.

One of the critics who reviewed the first edition was kind enough to take the first sentence of the book seriously and compared it to a well-written cookbook. To take that analogy a bit farther, this edition is organized as one might a cooking class for someone who had never prepared food—indeed had never even seen a kitchen—and assumed that food simply appeared by magic.

One might begin such a class with an entire evening devoted to boiling an egg, beginning with purchasing the egg, boiling water, putting the egg into the water for a specific duration, cracking the egg, and finally, presenting the egg at a table. It's not a lot, but it allows the student to experience the basic elements of cooking and results in something edible at the end of the class. That is exactly the purpose of Chapter 1, "Tutorial: Create a Tree!". The reader learns one way to create a tree that is a valid representation of the historical relationships among a set of related sequences. It is not enough for a meal, but it is a start.

The next chapter, "Basic Elements in Creating and Presenting Trees," might be compared with demonstrating sautéing, baking, broiling, and roasting. Each is an important method and it is necessary to know something about all of them if one is to prepare any but the most limited meals. Chapter 2 explains how to fine-tune alignments—the raw data from which phylogenies are constructed—and then presents each of the major methods for constructing phylogenies in some detail, with examples. The chapter finally turns to methods of presenting trees in publications and how to troubleshoot some common problems. Chapter 2 is the meat of this book. Understanding Chapter 2 will permit confident construction of phylogenetic trees under circumstances that anyone but an expert systematist is likely to encounter.

Throughout Chapter 2 you will find notes such as

Paup for Windows/Unix: pages xxx-yyy

and

PHYLIP: pages xxx-yyy.

These notes guide you to the pages in Chapter 4 where the same method is discussed.

Chapter 3 presents some advanced topics for those who may want to go beyond the basics. It is not necessary to understand, or even to read, Chapter 3 in order to construct sound, valid phylogenies. Chapter 3 is like learning to make soufflés or spectacular flaming desserts. It is not necessary, but it can sure add a nice touch. The advanced topics in Chapter 3 include reconstructing ancestral states (i.e., estimating the sequences of ancestral genes); analyzing trees for evidence of adaptive evolution; and constructing very deep phylogenies from protein crystal structure data.

Chapter 4 recognizes that not everyone will want to invest in the purchase of PAUP*, the primary software package discussed here for tree construction. It is also the case that PAUP* for Windows does not use the same interface as PAUP* for Macintosh computers, from which the examples in book are drawn. Chapter 4 therefore covers the details of tree construction using PAUP* for Windows and using the free and widely used PHYLIP program. The only cooking analogy for this chapter is that it is like low-calorie cooking. Not as much fun, but sometimes necessary.

Time-Limited Version of the PAUP*4.0 Program

The CD accompanying this book is compatible with both Windows and Macintosh computer platforms. It includes a fully functional *but time-limited* version of PAUP*4.0 beta10 for each platform. Users can install PAUP* onto *one* computer, and it will be functional for a period of six months from the date of installation.

The purpose of including this time-limited PAUP* is to permit students to become familiar with the program, especially in the context of a course for which this book may be required. Those students who wish to continue using PAUP* for phylogenetic analysis after that time are encouraged to order PAUP* from Sinauer Associates (<http://www.sinauer.com>). An advantage of purchasing PAUP* is that a registered purchaser is entitled to free updates of all beta releases, and to the final package release (which is expected to include a user's manual). Registered purchasers are also notified by email of any issues that may arise with respect to the use and/or performance of PAUP*.

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 various methods and suggest further reading. It is not necessary to read the boxes to be able to construct a reliable, valid, phylogenetic tree, but understanding the principles outlined will help troubleshoot the phylogenetic problems that arise when creating 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* (Graur and Li 2000) very helpful and enjoyable to read. Li's *Molecular Evolution* (Li 1997) and Chapters 11 (Swofford et al. 1996) and 12 (Hillis et al. 1996) of *Molecular Systematics* (David Hillis, Craig Moritz, and Barbara Mable, eds.) provide more detailed insights into these topics.

Computer Programs Discussed and Where to Obtain Them

All of the programs described in this book are available for both Macintosh and Windows platforms, and most are available for Unix as well. The examples in this book use the Macintosh versions of software that were current in November of 2003. The Windows and Unix versions of the programs may vary slightly in detail, but not so much as to preclude using this book as a guide for those platforms. The Windows versions of these programs have been tested by the author, and where they differ significantly from the described Macintosh programs the differences are noted in the text. Software authors are continuously upgrading (and sometimes even improving) their software, so some menus, windows, and dialogs may differ from the examples shown. If you don't see a feature that is mentioned here, look around a bit. It may be in a different menu or have a slightly different name. Users should ALWAYS register these programs when registration is available. Registration often allows software authors to alert users of updates and new features.

If you have access to a Macintosh computer I strongly recommend that you use that machine for phylogenetic analyses and that you obtain PAUP* for Macintosh. Modern Macs, especially the new G5 computers, are as fast or faster than the fastest Windows machines. The Macintosh implementation of PAUP* is superior to the Windows/Unix implementation because it includes a high-resolution tree-drawing program that allows you to present your trees easily and conveniently. If you can possibly do so, buy PAUP*. It is considerably more convenient to use than the free PHYLIP.

ClustalX

ClustalX (Thompson et al. 1997) is the primary multiple alignment program. Files created by ClustalX can be used by other programs to display and print trees, as well as to display alignments in ways that facilitate recognizing regions of high similarity. ClustalX for Macintosh, Windows, Linux, and some Unix machines is available free over the Web at

<ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/>

Help for ClustalX is available at

www-igbmc.u-strasbg.fr/BioInfo/ClustalX/Top.html

Examples in this book were created using ClustalX version 1.81. If you have an earlier version of ClustalX, or if you are still using ClustalW, you will find it useful to download the most recent version of Clustal from one of the above sites. If your version is more recent than 1.81, the appearance of some screens may differ slightly from the illustrations.

TreeView

TreeView is a free program for drawing phylogenetic trees. It does not create those trees, it simply uses files created by phylogeny programs to display and print the trees. TreeView allows you to modify the appearance of a tree to suit your taste and needs. TreeView for Macintosh, Windows, Unix, and Linux is available at

<http://taxonomy.zoology.gla.ac.uk/rod/treeview.html>

TreeView will soon be replaced by TreeViewX, currently under development. The incomplete development version is currently available from

<http://darwin.zoology.gla.ac.uk/~rpage/treeviewx/>

TreeViewX promises some interesting new features, so it is worth checking on the progress of this major revision.

PAUP*

PAUP* 4.0 (Swofford 2000) is the primary tree-building program discussed in this book. PAUP* uses the files created by ClustalX to build trees by any one of several methods. PAUP*4.0 beta is available for Macintosh, Windows, and

Linux/Unix. PAUP* is an inexpensive commercial program available from Sinauer Associates, Sunderland, MA. Information about ordering PAUP* 4.0 is available at

<http://www.sinauer.com/detail.php?id=8060>

and at the PAUP* home page,

<http://paup.csit.fsu.edu/>

Wherever you order PAUP* be sure to go to the PAUP* home page and download both the Command Reference Document and the Quick Start Tutorial.

Although it is a superb program, PAUP*4.0 is not yet in its final form. The current release is PAUP*4.0b10, meaning beta version ten. It is not clear that PAUP*4.0 will ever be finalized or that a manual for PAUP* will ever be written, but each beta release represents a significant improvement over the previous beta. In fact, each beta release corresponds to a minor new version. If you cannot be comfortable without a manual for PAUP*, use PHYLIP to construct your trees and TreeView to draw them. Despite the absence of a manual, I recommend PAUP* and I strongly recommend using the Macintosh version if a Macintosh computer is available to you. The Macintosh version (but not the Windows or Unix versions) includes an excellent interface for drawing and printing trees.

PHYLIP (PHYLogeny Inference Package)

This is a package of programs for inferring phylogenies (evolutionary trees). It is available free at

<http://evolution.genetics.washington.edu/phylip.html>

PHYLIP is written to work on as many different kinds of computer systems as possible. The source code is distributed (in C), and for some operating systems executables are also distributed. In particular, already-compiled executables are available for Windows95/98/NT, Windows 3.x, DOS, and Macintosh systems. Complete documentation is available on documentation files that come with the package.

The PHYLIP source code is available to be compiled on any Unix machine.

Tree-Puzzle

Tree-Puzzle is a program for constructing Maximum Likelihood (ML) trees from DNA and protein sequences. It is available free of charge from

<http://www.tree-puzzle.de/> (Tree-Puzzle home page)

<http://iubio.bio.indiana.edu/soft/molbio/evolve> (IUBio archive www, USA)

<ftp://iubio.bio.indiana.edu/molbio/evolve> (IUBio archive ftp, USA)

<ftp://ftp.pasteur.fr/pub/GenSoft> (Institut Pasteur, France)

MrBayes

MrBayes (Huelsenbeck and Ronquist 2001) is a program for constructing phylogenetic trees by the Bayesian method. MrBayes is available from

www.mrbayes.net

The currently available version (November, 2003) is a beta version, MrBayes 3.0b5, but the final MrBayes 3.0 should be released by the time you read this. Please note that some commands have changed from MrBayes 2.0, so readers interested in using the examples in this book should update to MrBayes 3.0.

CodonAlign

CodonAlign is a simple program that creates a DNA alignment based on alignments of the corresponding proteins. It introduces into each DNA sequence a triplet gap at the position of each gap in the aligned protein sequence. CodonAlign is available from this book's companion website,

<http://sinauer.com/hall/>

The new version, CodonAlign 2.0, is somewhat easier to use than the original. Users of the earlier version should be sure to read the new documentation for input file format.

Other Programs

There are many phylogenetics programs that I have not mentioned here, some of them widely distributed. My failure to mention a program does not imply that the program is not useful or valuable. I have chosen to include a minimal set of programs that will allow you to implement the methods I discuss in this book, and I have tried to include only programs that are available for Macintosh, Windows, and Unix platforms. See

<http://evolution.genetics.washington.edu/phylip/software.html>

for a more extensive list of phylogenetics programs.

Files and Utilities on the "Phylogenetics Made Easy" Website and CD

The "Phylogenetics Made Easy" (PME) website was created especially for this book. Located at

<http://www.sinauer.com/hall/>

the site contains a variety of files that will make it easier for you to follow the tutorial without actually downloading the sequences to create an input file for ClustalX. It also includes copies of all of the relevant output files so you can compare your results with those I obtained in the event that you have difficulties.

In addition, the CD that comes with this book includes all of the files that are on the website at the time of publication. Instead of downloading, you can simply copy those files from the CD.

The site and CD also include templates, or boilerplate, for various input files. You can copy those templates directly into your own input files in lieu of typing everything manually. It is amazingly easy to make a tiny typing error, such as "1st_pos" instead of "1st_pos", and spend hours trying to figure out why your program won't run. Copying blocks of critical text, then modifying those blocks for your specific needs, is a good way to avoid such problems. Indeed, many of us use such templates routinely. Throughout the text these files and templates are indicated with an icon. For example,



Chapter 1 files:Clustal#1:Metallo.aln

means on the website in the Chapter 1 files folder, in the Clustal#1 folder, the Metallo.aln file.

The site and CD also include a copy of the program CodonAlign 2.0, mentioned earlier. They do not include copies of ClustalX, TreeView, or MrBayes. Although these programs are distributed at no charge, it is much better for you to download the most recent versions from the websites listed above than to work with the versions that were current at the time this book was assembled.

Some Conventions Used in This Book

- **Click**, as in "click the OK button," means to use the mouse to position the cursor over the indicated button on the screen and to depress and quickly release the mouse button. **Double Click** means to click twice rapidly, without moving the mouse.
- **Drag** means to position the mouse and, while holding down the mouse button, move the mouse to another position.
- **Select** means to highlight a section of text or an object on the screen by dragging across the indicated region or by double clicking on the object.
- In the text, the **Chicago font** indicates a menu item or a button that you will see on the screen.
- For command-line programs such as PHYLIP, MrBayes and CodonAlign, the **Courier font** indicates text that you will see on the screen or that you will type into an input file.