

Book Review

“MacTrees made easy”—Review of “Phylogenetic trees made easy: a how-to manual for molecular biologists” by Barry G. Hall

We are living in the golden age of molecular phylogenetics. Phylogenetic trees are appearing everywhere, from specialized journals in molecular evolution and evolutionary genetics to specialty journals such as the *Cell* and *Development*. The practice of molecular phylogenetics now permeates most disciplines of molecular biology and genetics, as the summarizing and predictive power of phylogenetic analysis has begun to be appreciated widely. With this expansion of interest in building phylogenetic trees beyond traditional boundaries, there is a clear need for books and how-to manuals for molecular phylogenetic analysis software aimed at beginning investigators and non-specialists. Entries in the former category have increased quickly over the last few years, but the latter is largely represented by manuals written by the authors of academic software (e.g., *PAUP*, *MacClade*, *MEGA2*, *PHYLIP*, and *PAML*). These software manuals generally describe program functionality along with some illustrated examples. They are not intended to be step-by-step cookbooks on how to accomplish specific tasks and “read” the results generated. The publication of “Phylogenetic trees made easy” is, therefore, a welcome addition to the how-to category of books for molecular phylogenetics.

The “Phylogenetic trees made easy” is intended to bring the world of phylogenetic analysis to the fingertips of molecular biologists, particularly the Macintosh users. It is a tutorial on various aspects of tree making using programs with extensive user interface for the Mac operating system. It begins with a long chapter (called section) that occupies almost one-third of the book and provides a hands-on tutorial on how to build a tree. It starts with assembling a set of homologous sequences using a BLAST search, follows it with sequence alignment using *ClustalX*, and then builds a phylogenetic tree neighbor-joining (NJ) in *PAUP**. The author advises against using *ClustalX* for building NJ trees: it is unclear why. For a molecular biologist, it would actually have been better to use (and thus have to install) only one program (*ClustalX*) to build the NJ tree to get the initial feel for the process. It would have made the discussion more general, also since BLAST and *ClustalX* are both available on non-Macintosh systems. Instead, the reader is now hampered by the steep learning curve involved in

learning any sophisticated phylogenetics program and has to deal with frustrations that come from program-specific data file format conversion (p. 39). In this respect, this book is more like a how-to manual for *PAUP**, and surely be a good supplement for the *PAUP** manual.

In this section, the author states his favorite method (Bayesian inference), which is mentioned to be gaining rapid acceptance (pp. 37 and 89). Recent enthusiasm for Bayesian methods for phylogenetic inference comes from their ability to alleviate time-requirement issues in maximum likelihood analysis, especially for phylogenetic hypothesis testing. However, molecular systematics field is known to experience waves of “beliefs” in certain methods every few years, and the final outcome about this relatively new method is yet to be known. It is certainly promising and we are learning more every day. For unsuspecting beginners, who are the target of this book, an unqualified endorsement may be harmful at this stage. Also, the book unequivocally designates the Bayesian posterior probabilities (BPP) as substitutes for ML bootstrap analysis (p. 61). It is well known that these two are distinct quantities and may have no clear-cut relationship, except when bootstrap values are very high. In fact many investigators are now realizing that BPPs are usually very high for most datasets, even when all other methods suggest lack of significant resolution of the phylogenetic relationships. We are yet to understand why this is so. In this regard, this book digresses from its tutorial mission to advocating role, which has led to presentation of an unbalanced view of the molecular phylogenetic field.

The author mentions that the NJ method is algorithmic because it uses a “specific series of calculations to estimate the tree.” This is a long-standing misconception. By this definition, all methods as implemented in any program are algorithmic because they undergo a specific (predefined) series of calculations. The difference is that the NJ method usually produces a single tree, whereas others may produce multiple trees. However, the NJ method (even UPGMA) indirectly examines a large number of phylogenies at each step of sequence clustering and chooses the best result locally. Computer simulations have shown that this method known to work as well as, or better than, more exhaustive searches in which a very large number of trees are explicitly examined under the minimum evolution criterion. Therefore, misstatements in this book may bias the beginning

investigators' viewpoint. In general, however, the book is to be commended for its extensive coverage of different types of methods in terms of the space allocated. I would have, however, liked to see a greater emphasis on precise writing for new comers, because simplistic, overloaded statements come with potential for misrepresentation and misunderstanding.

The second section of this book is relatively short and is called "Additional methods for creating trees." It goes beyond a simple tutorial and discusses methods other than NJ. In this section, I found the concepts of tree building and tree representation to be confounded in the discussion. For instance, it is stated that a consensus tree built from multiple trees in parsimony analysis is better than the NJ tree because the latter does not contain polytomies (p. 85) even if the true gene tree may contain multifurcations (polytomies). However, NJ trees and likelihood methods also produce trees with polytomies, because they may contain branches of effectively zero length in the tree produced. In fact, some branches can even have negative lengths in the NJ trees, which are direct indicators of polytomies. These misconceptions stem from the fact that unweighted parsimony works with exact integer counts, so the concept of 0 is clear-cut. In other statistical methods, the concept of 0 is not exact as most numbers tend to be positive, but rather close to 0, which essentially result in polytomies. Also, polytomies will manifest in bootstrap majority-rule (or strict) consensus trees for any tree-making method. Therefore, there is little or no difference among methods. Similarly, it is stated that the Bayesian method produces a large number of trees with similar [maximum] likelihoods, suggesting that the direct maximum likelihood methods do not. By default, *PAUP** saves and presents the ML tree(s). However, it can be told to hold multiple ML trees, as we do when we use the Bayesian programs. There are, of course, differences between the Bayesian and Likelihood methods for selection of trees, but again the default program implementation details differences between *MrBayes* and *PAUP** are confused with difference between the methods.

In this section, the book also contains brief tutorials for *MrBayes* and *Puzzle*. One of my students, who had never used *MrBayes* before, was successful in running the Bayesian analysis using this book as a guide. Therefore, the author succeeds in bringing difficult to use programs to novice users as they (e.g., *MrBayes*) require the use of the command line prompt that is a major impediment for most first time users, who are accustomed to using mice and menus on Macintosh.

The third section is on "Presenting and printing of your trees." Clarifications on various types of cladograms and phylograms as well as rooting strategies are presented in order to demystify terms that appear in menus and dialog boxes in *PAUP**. They are nicely

explained with clear-cut illustrations. However, the unreadability of a specific NJ tree (with branches drawn to scale) with two clusters of closely related sequences (almost a star phylogeny in each cluster) is mentioned to be a shortcoming as compared to the MP tree topology (cladogram) for the same dataset in which the actual branch lengths are written on each branch (p. 116; also note that in Fig. 3.1 legend it should be a reference to Fig. 3.2 rather than 3.7). In fact, one could draw NJ tree in exactly the same way as the MP tree (it is possible in *PAUP**), or the MP tree in the same way as the as the NJ tree. There is exactly the same type of information in both trees; only the visual representation is different. Also, it is mentioned that a rooted-style tree (Fig. 3.1) is to be preferred over other "unrooted" representations (Figs. 3.2 and 3.3). It can be argued that the unrooted representation in Fig. 3.2 is actually superior in that it does not "force" a rooted perspective on to the reader (as does the tree in Fig. 3.1). For multigene family evolutionary trees, the primary interest of most molecular biologists, the unrooted style (e.g., Fig. 3.3) style is often more desirable. In addition, an unresolved bush in Fig. 3.2 may be useful to population geneticists who may not be interested in within population relationships of closely related haplotypes. Therefore, the relative merit of different representations depends on the objective of the study. This is the reason why authors of various academic software packages spend considerable time in providing such flexibility.

Furthermore, the use of the actual number of substitutions (number of substitutions per *sequence*) in writing branch lengths on the tree is advocated (p. 132), as opposed to the number of substitutions per *site*. Multiplying the number of substitutions per site by the sequence length can accomplish this. While it might seem intuitively more appealing, this strategy may lead to incorrect outcomes. This is because when the sequences contain sites with alignment gaps and missing data (which is often the case in multigene family sequence alignments), the actual number of substitutions cannot be computed by simply multiplying the branch length (in terms of the number of substitutions per site) by a common sequence length, as the actual numbers of nucleotide bases (or amino-acid residues) differ among sequences.

The fourth section discusses the fine-tuning of alignments, which could have been easily merged into section one. The following section is on reconstruction of ancestral DNA sequences and contains an extensive tutorial on how to use *MrBayes*. It would have been useful to have an additional short tutorial on *MacClade*, which is a popular Macintosh program with extensive facilities for reconstructing and visualizing ancestral sequences using maximum parsimony analysis. In the last section, the book discusses another program called *CodonAlign*, which is intended to facilitate alignment of coding

sequences such that the corresponding amino-acid sequences are first aligned followed by adjustment of codons in accordance with the amino-acid alignment. In our tests, we found that the *CodonAlign* program was extremely picky regarding the input file formatting, but it does produce a useable Nexus file ready for *PAUP** analysis. It was almost impossible to get this program to work on PCs, even though it could be compiled into an executable. However, at least Macintosh users will find *CodonAlign* useful in their daily activities.

It would have been nice if the book were more inclusive in its treatment of other software. For instance, *PAML* (available on Macs also; free) contains an extensive set of functionalities for hypothesis testing and other likelihood inference. *MacClade* is widely distributed software with an excellent user interface for Macintosh platform. Other programs, such as *MEGA2* (free) are available for Windows operating systems only, but can be easily run on Macs using emulators (e.g., VirtualPC and SoftWindows, which cost <US \$100). Non-mentioning of these programs, especially given their extensive user base, even in the introduction (p. 4) is puzzling. This along with a number of long-standing

misconceptions in molecular systematics included in this book as statement of facts is a major weakness of this book.

In conclusion, the *Phylogenetic Trees Made Easy* is a hands-on tutorial for beginners who want to use *PAUP** (a well established software) or *MrBayes* (a relatively newer entry). The book is written in a lucid style with a language accessible to non-experts and non-systematics. Its pricing provides the flexibility to the instructors to designate it as a supplementary material for courses with significant phylogenetic components in senior undergraduate and graduate level. For readers wanting a start-up guide to the software packages used in this book, it will be a useful resource.

Sudhir Kumar
Center for Evolutionary Functional Genomics
Department of Biology
Arizona State University
Tempe, AZ 85287-1501
USA
E-mail address: s.kumar@asu.edu