ACADEMIC
PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Cladistics 19 (2003) 276–285

Cladistics

www.elsevier.com/locate/yclad

Book review

The perils of ‘point-and-click’ systematics

Phylogenetic Trees Made Easy, a How-To Manual for Molecular Biologists. By Barry G. Hall. Sinauer Associates, Sunderland, MA, USA, 2001. 179 pp. $29.95 US.

> The rising preeminence of Phylogenetic Systematics runs the risk of being self defeating, for it is becoming more and more common for practitioners of other approaches to pay lip-service to phylogenetic principles... This tendency seems to be most pronounced when the alternative approaches are of a mathematical nature or are implemented by computer programs, and the practice hinders continued development of truly phylogenetic methods.
>
> (Farris et al., 1982, p. 317)

Four factors seem responsible for the rising preeminence of phylogenetic systematics in recent years. First, it is now recognized widely that evolutionary history must be considered when addressing any biological problem and that phylogenetic systematics is the field of biology that aims specifically to elucidate that history. Second, explicit numerical techniques have eliminated much of the subjectivity that characterized systematics in the post-Darwinian era. Third, technological advances in computer hardware and software make it possible to collate data and perform complex phylogenetic analyses in reasonable amounts of time and user-friendly environments. Fourth, standardized molecular techniques have been incorporated into systematics research, bringing massive amounts of data to bear on phylogenetic problems. These advances have helped phylogenetic systematics prosper, but they have also given rise to the modern phenomenon of ‘point-and-click’ systematics. It has become possible to carry out extremely sophisticated-looking analyses without any consideration of the theory behind the commands, the underlying algorithms and assumptions, the pros and cons of the different approaches, or even the biological principles that provide the foundation of phylogenetic systematics. To a large degree user-friendliness has replaced scientific rigor in the design and implementation of systematics studies. This is the world into which Barry G. Hall’s *Phylogenetic Trees Made Easy, A How-To Manual for Molecular Biologists* was born.

The first page of the book begins with a disclaimer:

> This is a “cookbook” intended to aid beginners in creating [sic; see below] phylogenetic trees from protein or nucleic acid sequence data... 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.

The book claims to assume “basic familiarity with personal computers and with accessing the World Wide Web” but, importantly, no biological knowledge whatsoever. Hall openly admits in the Acknowledgments that he “can (almost, sometimes) understand” (p. vii) many of the subjects he covers in this book. With this in mind, we expected Hall’s book to be (as advertised) a simple how-to computer program manual, a sort of ‘ClustalX, PAUP*, Tree-Puzzle, and MrBayes for Dummies’ that described the basic user-specified commands and options and provided illustrative examples and exercises to accompany undergraduate computer labs. Although we were doubtful that such a book would inspire the next generation of biologists to undertake phylogenetic studies, a practical manual describing clearly and simply what the different programs and commands do would be a step toward the elimination of ‘point-and-click’ systematics.

Had Hall remained within the limits established in the Introduction, we may have had little positive to say about this book, but we also would have found little to criticize. However, *Phylogenetic Trees Made Easy* is not simply a practical manual. Theoretical discussions replete with Hall’s personal interpretations and justifications occur throughout the book—presumably for the benefit of “the investigator who has a modest familiarity with phylogenetic tree construction but needs to address some aspects and problems in more depth” (p. 1). Although incorporating theoretical background into a how-to manual would undoubtedly assist in eliminating ‘point-and-click’ systematics, Hall’s cavalier treatment of almost every subject creates the impression that phylogenetics is a weak, poorly argued collection of methods instead of a rigorous analytical system sufficient to provide the historical context for all other branches of biology. Far from a step toward the elimination of ‘point-and-click’ systematics, the many misconceptions, inaccuracies, misrepresentations, and inconsistencies perpetuated throughout this book serve to exemplify the perils of doing without knowing why. The fact that it was written to introduce beginners to the field is of special concern.

doi:10.1016/S0748-3007(03)00029-X

Our review of *Phylogenetic Trees Made Easy* is divided into three sections. First, we give an overview of the structure and presentation of the book. Next, we examine the content of the text, addressing subjects largely (but not exclusively) in the order in which they are presented in the book. Finally, we conclude with a brief summary of strengths and weaknesses and a series of recommendations for authors and publishers who may consider a similar undertaking.

Structure

Physically, the book consists of 179 pages, plus front- and back-matter, and is soft bound in heavy bond, laminated paper. The quality of the binding is not good and we doubt that it would withstand a semester of heavy class use; within a few weeks of occasionally reading our copy, the glue began to break and the pages to fall out. The text font and size are easy to read, and the many boxes and boldfaced and italicized asides are used effectively. The text of several of the screen-captured images is hard (or impossible) to read, but that is difficult to avoid in computer how-to manuals.

The book is divided into an Introduction (“Read Me First,” 6 pp.), six major Sections (“Tutorial: Create a Tree!,” 62 pp.; “Additional Methods for Creating Trees,” 45 pp.; “Presenting and Printing Your Trees,” 18 pp.; “Fine-Tuning Alignments,” 4 pp.; “Using MrBayes to Reconstruct Ancestral DNA Sequences,” 7 pp.; and “Dealing with Some Common Problems,” 7 pp.), each with numerous subsections, and two appendices (“File Formats and Their Interconversion Using PAUP*,” 10 pp.; and “Printing Alignments,” 2 pp.). The remainder of the 179 pages of the book are dedicated to Literature Cited (only 2 pp.), a useful “Index to Major Discussions” (3 pp.), and a more complete “Subject Index” (7 pp.). Its presentation is typical of user manuals, equipped with step-by-step procedures and screen-captured images to help orient the uninitiated. All images are gray scale; this is understandable, given economic constraints, but the 14 screen shots of ClustalX alignments would have benefited greatly from color reproduction. Sections 1 and 2 include nine “Learn More” boxes that “present somewhat more detailed background on the various methods and suggest further reading” (p. 2). Given the introductory level of this book, a glossary would have been a helpful addition, as would a flow chart illustrating the steps from obtaining data to printing trees.

Substance

Setting the tone for the remainder of the book, Hall starts off on the wrong foot, billing Section 1 as: “Tutorial: *Create* a Tree!” (p. 7, italics added) and setting off such subsections as “Why *create* phylogenetic trees?” (p. 7, italics added) and “Using PAUP* to *Create* a Tree” (p. 37, italics added). By presenting the problem in this way, Hall creates the impression that phylogenetic hypotheses are somehow generated or induced from the data, when in fact the hypotheses already exist and data are used to choose among them, either by evaluating evidential support comparatively (i.e., tree searching) or by computing a single solution (i.e., “algorithmic” methods). Not until p. 50 (“Learn More about Phylogenetic Trees”) is the reader informed that there is a finite number of possible binary topologies defined by the number of terminal taxa (curiously, this information is repeated in detail on p. 86 to explain maximum likelihood, but was not mentioned in the preceding section on parsimony), and the concept of tree searching is not brought up until p. 70 (“Learn More about Tree-Searching Methods”). Even after these subjects are addressed, they are not coupled with notions of hypothesis testing; in fact, hypothesis testing is never mentioned in this book.

The first substantive issue. Hall addresses in Section 1 is the question, “Why create phylogenetic trees?” Hall’s answer: to understand protein function. As Hall explains (p. 7),

> We are frequently forced to assign biological functions to proteins on the basis of sequence homology alone... Examination of a phylogeny can allow you to determine just how closely or distantly your sequence relates to a sequence whose function is actually known from biological or biochemical information.

Throughout the book, Hall focuses exclusively on protein-coding sequences, despite the fact that ribosomal, ‘nonfunctional’ (e.g., introns, pseudogenes), and other kinds of DNA sequences are also of great interest to molecular biologists and systematists alike. Furthermore, according to Hall (p. 7), the reason that we study protein function via sequence alignment and phylogenetic analysis instead of using “a table of pairwise homologies, expressed as percent identities or percent similarities” is simply a result of the size of sequence databases:

> As databases grew, it became impossible to present tables of all the homologs, so we started to create multiple alignments with programs such as Clustal and Pileup.

Hall neglects to mention that homology is a historical concept unrelated to “percent identities or percent similarities,” as was argued decisively by Hennig (1966) some 35 years ago and then in an explicitly molecular context by Reeck et al. (1987). Even more fundamentally, he neglects to mention that the concept of homology is unrelated to *function*—an argument that dates back much further (for historical review see Ghiselin, 1976)—the inference of which is (according to Hall) the very purpose for "creating phylogenetic trees"! In fact, although the word is used repeatedly throughout the book, "homology" is not defined until p. 149. But then "phylogeny" is *never* defined, which may explain why Hall believes that the reason we "create phylogenetic trees" is to infer sequence function. For Hall, it seems that understanding the evolutionary history of the DNA sequences (not to mention the terminal taxa) is not the objective of phylogenetic analysis. The fact that "phylogenetic trees have been used to represent the historical relationships of groups of organisms—often species" is not mentioned until p. 44, and even then it is brought up only because that is how phylogenetic trees were used "[t]raditionally." For Hall, a phylogenetic tree is merely "a simple object consisting of two elements: nodes and branches" (p. 44). That is correct as a technical definition useful in formal representation and computer programming, but in this introductory book it would seem much more useful to define these branching diagrams as evolutionary hypotheses.

Following the subsection "Why create phylogenetic trees?," Hall provides an overview of the steps involved in searching for and downloading sequences from GenBank using BLAST. It strikes us as unlikely that the trained molecular biologists for whom this book was written would require basic instructions on using GenBank. Instead, the detailed step-by-step instructions, screen-captured images, and definitions of bit score and E value would be useful to those who have never downloaded sequences from GenBank, such as beginning undergraduates and systematists trained in analysis of traditional data.

Apart from the confusion over the intended audience, the theoretical shortcomings of Hall's treatment have practical implications in this subsection, particularly when deciding which sequences to exclude from phylogenetic analysis. Hall's advice is (p. 16):

> Except that duplicates are to be avoided, there is no hard and fast rule about deciding which sequences to include and which to exclude. If you are interested in proteins that are probable biological homologs of your protein of interest you will most likely want to exclude very small proteins or protein fragments.

There may be something to this if by "biological homologs" Hall means functional equivalents (assuming that proteins of vastly different lengths have vastly different functions), but it is bad advice if one is interested in the evolutionary history of those sequences. For example, what if those small proteins are really homologous with the large ones? To assume otherwise would imply the impossibility of indels. Or what if they are partial sequences? Why not break up the larger sequences to permit alignment of the homologous fragments? This is one of the most common procedures in using downloaded sequences in phylogenetic analysis, because few studies use exactly the same primers or sequence entire genes or functional regions, yet it is never considered in Hall's treatment. The "hard and fast rule" in phylogenetic analysis is that the sequences must be homologous, but that is at best only indirectly related to sequence length. It is puzzling that Hall does not suggest consulting the author's description and primary literature associated with the sequence, stating only that a downloaded file "includes a lot of information about the sequence" (p. 17).

The next step in "creating a tree" is to "create" the multiple sequence alignment. Unfortunately, Hall addresses this subject in part of Section 1 ("Creating the Multiple Sequence Alignment"), Section 4 ("Fine-Tuning Alignments"), and Appendix II ("Printing Alignments"). It would have been more effective for these sections to be combined into a single chapter on alignment, as this would eliminate repetition (e.g., "Refining and Improving the Alignment" in Section 1 and "Fine-Tuning Alignments" are highly repetitive) and prevent readers from having to flip from section to section to deal with alignment issues. In total, Hall devotes 33 pages to alignment, a significant portion of his book.

As in previous parts of Section 1, the instructions on using ClustalX are clear and straightforward. Hall covers input and output file formatting, importing input files, setting alignment parameters, interpreting and modifying ("refining" or "fine-tuning") results, adding new sequences to an existing alignment, and aligning two sets of aligned sequences with each other. Although it may be important to know how to print alignments, the emphasis Hall places on it (e.g., Appendix II) is unwarranted. Journal editors are increasingly reluctant to fill precious pages with alignments that can be deposited online, and alignment visualization, comparison, and manipulation may be performed much more effectively (not to mention economically and environmentally responsibly) by opening multiple windows in programs such as the freeware BioEdit (T.A. Hall, 1999; available at http://www.mbio.ncsu.edu/BioEdit/bioedit.html).

Given the emphasis that Hall places on alignment and the level of detail included on other topics, we were disappointed not to find a "Learn More" box covering the Needleman–Wunsch dynamic programming algorithm used to perform alignments (Needleman and Wunsch, 1970). In fact, the details on what ClustalX actually does are remarkably scarce. Early in Section 1 we are informed that (p. 7)

> The process of creating a multiple alignment begins with computing all pairwise alignments, then making a rough "guide tree" from those pairwise comparisons.

Later Hall adds that "[ClustalX] uses that guide tree to help create the multiple alignment" (p. 23). However, although we are informed that guide trees are not "[v]alid phylogenetic trees" (p. 8), we are never told how they are used or why they are needed in multiple sequence analysis. A brief summary of the problem of implementing the Needleman–Wunsch algorithm with multiple sequences and the Feng–Doolittle heuristic solution (Feng and Doolittle, 1987) is essential in a book like this (see Wheeler (1994) for a concise summary; Phillips et al. (2000) is more up to date, but may not have been available when this book went to press).

Unlike many contemporary workers who seem to consider sequence alignment to be almost irrelevant, Hall rightly warns readers repeatedly that "the quality and value of [a phylogenetic tree] will be no better than the quality of the alignment" (p. 29; see also pp. 34, 37), and his efforts to draw attention to the importance of alignment in phylogenetic analysis are commendable. Unfortunately, the procedures Hall proposes in order to evaluate and improve alignment quality are not so laudable. The only piece of advice that has any objective basis is the deletion of nonhomologous regions (p. 30). However, in Hall's example involving short sequences that only partially overlap with long sequences, a better approach than discarding evidence (the segments of the long sequences that do not overlap with the short sequences) is to break up the longer sequences into homologous fragments, align each set of fragments independently, and then merge the alignments into a single matrix for phylogenetic analysis. Of course, in that case there would be some amount of missing data, which is another important subject that is never addressed in this book.

Following initial automated sequence alignment in ClustalX using default settings for DNA sequences or gap opening 15.00 and gap extension 0.30 for protein sequences (Hall does not explain this preference), Hall instructs that (p. 23, italics added)

> It is necessary for the user to carefully and thoughtfully examine each alignment to see if it *makes biological sense*.

Likewise (p. 30, italics added),

> At this stage you need to examine the alignment to see if most of the gaps *make sense*. If many of the gaps *seem to be arbitrary* (i.e., you think you could have done *better by eye*), then you will need to *improve* the alignment.

What does it mean for an alignment to "make biological sense" or to "seem to be arbitrary?" On what basis can one decide whether a "better" alignment could have been produced by eye? Without an explicit definition of these criteria, the objectivity obtained from algorithm-based sequence alignment is lost. Biologically, any nucleotide at any position could be homologous with any nucleotide of a homologous sequence, so how can one assess the quality of an alignment without invoking some optimality criterion? In partial answer to these questions, Hall offers a novel optimality criterion for selecting gap (indel) penalties, stating authoritatively (p. 31):

> We should attempt to minimize the number and size of gaps while maximizing the extent of conserved blocks.

He defines a conserved block as "a region in which similar or identical residues occur across all or most of the sequences" (p. 31). In light of his optimality criterion, Hall describes an iterative alignment procedure whereby gap costs are increased, and (p. 32)

> we simply check to see if we are reducing the number of gaps, which is good, or whether we are starting to break up homologous [= conserved?] blocks that were present at lower gap penalties, which is not good.

What is the basis of this procedure? Why is this preferable to minimizing the number of gaps in shorter sequences or minimizing the number of gaps in longer sequences or minimizing discontinuous gaps? Hall's examples involve amino acid sequences, but the same procedure is meant for DNA sequences also. In that case, why optimize this function instead of, for example, the number of gaps that occur in triplets or the number of conserved blocks divisible by three? Why are gap costs only increased from initial settings and not decreased? Why should Hall's optimality criterion be implemented by rerunning the alignment program with different gap costs instead of by modifying the alignment manually (by eye)? The fact that Hall does not discuss manual refinement suggests he may be opposed to it, but because that is unquestionably the most common refinement technique, the topic should have been addressed explicitly. Why is it necessarily "good" to reduce the number of gaps? That procedure may lead to explanations of sequence evolution that involve many more transformations to account for the observed variation. Is that "good"? Of course, Hall may have defensible reasons for preferring his method, but he never argues his case; he simply states his personal, unfounded opinion as fact and instructs the uninformed reader to proceed. As the past 35 years of debate demonstrate, systematists have little patience for such authoritarianism, and one can only hope that molecular biologists do not either.

The final 32 pages of Section 1 and all of Section 2 are dedicated to the phylogenetic analysis portion of "creating a tree." Given the attention paid to sequence alignment, we were surprised that Hall never discusses treatment of indels in phylogenetic analysis. Hall seems to accept PAUP*'s default settings that treat gaps as missing data, a position that is not widely accepted (and one that seems at odds with Hall's view that "gaps are assumed to represent insertions or deletions that occurred as the sequences diverged from a common ancestor'' [p. 20]). Regardless of his own preferences, what is most problematic is that Hall does not even mention that there is a controversy or that PAUP* allows the user to specify how to treat indels.

Also missing from the book is any discussion of model testing in maximum likelihood, even though Hall includes a Learn More box on ''Evolutionary Models'' (p. 91) and asserts repeatedly that ''[a] major advantage of the ML method is that it allows users to specify the model for evolution'' (p. 90; see also p. 73). This hardly seems an advantage if there is no procedure by which the adequacy of models can be assessed. The section on maximum likelihood analysis of proteins using Tree-Puzzle (Schmidt et al., 1999–2000) is limited to a description of input data format, some default options, and output tree format. Given that this section is so poorly developed and that a concise and clear manual for Tree-Puzzle (http://www.tree-puzzle.de/manual.html) already exists, interested readers would do better to go straight to the original manual.

Hall's description of PAUP* procedures (all in the Macintosh environment; no instructions are given for molecular biologists who use PCs), Tree-Puzzle, and MrBayes are accurate (if not fully developed), but his discussions of the theoretical basis for procedures would be amusing if they were not so misleading. The three substantive issues he focuses on are (1) which method to choose, (2) rooting, and (3) reliability, which we address in order.

Methods for phylogenetic analysis are introduced briefly in Section 1. After naming the four ''primary methods'' (he does not specify the basis for this distinction) of neighbor joining (NJ), parsimony, maximum likelihood (ML), and Bayesian analysis, Hall suggests that ''method choice depends on both what you want to learn and the size and complexity of the dataset'' (p. 37). At best this introduction confuses the issues, because all these methods are intended to achieve the same goal (i.e., to provide a hypothesis of phylogenetic relationships). Instead of underscoring the differences in the assumptions made by these competing methods, Hall depicts them as complementary approaches designed for different questions and dataset sizes.

Hall chose neighbor joining for his tutorial, but Section 2 opens with the question ''Which method should I use?'' (p. 69). According to Hall, the fact that ''the field of phylogenetics is quite contentious with respect to which method is best'' (p. 69) boils down to subjective opinion. As such, he reports (p. 69)

> Much of the opinion amounts to religious conviction, and you need not worry about it. You could just stick with Neighbor Joining, but the other methods offer some advantages and some disadvantages when compared with Neighbor Joining.

What are the advantages of neighbor joining? ''It is fast, and it yields only a single tree'' (p. 69). Hall does not consider arbitrary selection among ties and sensitivity to input order to be disadvantages (Farris et al., 1996), but he does point out that ''[t]oday's fast, powerful desktop computers have greatly reduced the speed problem'' (p. 72)—overlooking the much more significant advances in search algorithms (e.g., Farris et al., 1996; Goloboff, 1996, 1999, 2002; Goloboff and Farris, 2001; Nixon, 1999)—and that obtaining a single tree is not an advantage at all. Despite having dismissed the only putative ''advantages'' of neighbor joining, Hall still insists that ''I do not think you should automatically discount publishing an NJ tree'' (p. 75).

Hall further observes that (p. 73),

> It would be lovely if there were some objective way to select the ''best'' method for constructing evolutionary trees, but no such way exists.

What does Hall suggest in light of this epistemological dilemma (p. 76)?

> My own rule of thumb is that I am willing to use a method that will run overnight while I am home. Therefore, if it takes longer than about 14 hours, I will probably choose another method.

This emphasis on time is puzzling, given his dismissal of speed a few pages earlier. But it is even more problematic when one considers that the duration of an analysis is determined not only by the optimality criterion but also by the exhaustiveness of the implementation. Hall misleads newcomers to the field of phylogenetics by conflating (1) the epistemological basis for preferring one method over another and (2) the practical constraints that determine the exhaustiveness of an analysis. The difference in times required by different analytic strategies applying the same optimality criterion may be at least as important as the difference in times between different methods (consider that Goloboff (1999) analyzed ''Zilla'' at least 15,000 times faster than Rice et al. (1997), even though both studies employed parsimony). Of course, Hall is aware that duration may be controlled by the user, but he regards this to be a special virtue of Bayesian analysis (p. 108):

> MrBayes is unusual in that it is the user who determines how long the run will take by setting… the number of generations.

Perhaps the reason that Hall did not realize that analysis duration may be determined by the user in other methods too is that he did not address heuristic tree searching in any detail. In the Learn More box on ''Tree Searching Methods'' (pp. 70–72), Hall devotes almost two pages to exact and branch-and-bound solutions and a few lines to step-wise addition and star decomposition, but he says nothing about multiple random addition sequences and only offhandedly remarks that "[a]nother heuristic approach, **branch swapping**, involves making predefined rearrangements of trees by one of a variety of means" (p. 72, boldface in the original). He says nothing about TBR or SPR, the two branch swapping algorithms mentioned in almost all published analyses involving heuristic solutions.

Hall further confuses matters by misrepresenting the different methods. For example (p. 76),

> I do not trust the underlying assumption of Parsimony: That the most likely scenario involves the fewest number of changes. That assumption implies an efficiency of the evolutionary process in which I have little confidence.

That interpretation of parsimony has not been entertained seriously since it was dismissed decisively by Farris (1983). Besides, the assumption of evolutionary efficiency is not required to interpret the most parsimonious solution as "most likely" (Tuffley and Steel, 1997).

Hall also dislikes parsimony because (p. 76)

> Parsimony does not allow me to have branch lengths on consensus or bootstrap trees, whereas Bayesian analysis as implemented in MrBayes does that.

That seems a strange complaint for Hall to make, given that a few pages later (p. 83) he points out that

> In a consensus tree, branch lengths have no meaning. How could you average over situations in which some branches do not exist?

Precisely! Nevertheless, the fact that branch lengths cannot be depicted is not a property of parsimony, but of the program Hall uses. If Hall really wants to obtain the branch lengths for consensus trees, he could use NONA (Goloboff, 1993–1999), which allows the synapomorphies common to all most parsimonious trees to be shown on the consensus, or WinClada (Nixon, 1999–2002), which gives (meaningless) branch lengths by optimizing characters on the consensus topology.

It should be pointed out that Hall never actually explains what consensus cladograms represent or what the different consensus techniques he refers to actually do. For Hall, it is a matter of personal choice (p. 83):

> How can you choose between these two [most parsimonious] trees? In one sense it doesn't matter; each of the trees is equally parsimonious and therefore as good as the other tree, so you can pick a tree at random. Another possibility is to compare the Parsimony trees with the NJ tree and pick the Parsimony tree that most resembles the NJ tree... Another option is to present a consensus tree... I like to use the 50% majority rule to compute the consensus, but you can use either strict or semistrict rules if you prefer.

Hall continues (p. 84),

> If you wish to choose a single tree to present, in many cases you can choose the tree that most closely represents the consensus.

One is left to wonder why parsimony analysis should be undertaken at all if the goal is to choose the single tree that most resembles the neighbor joining tree—why not just use the neighbor joining tree? More importantly, Hall neglects to mention that the advantage of the strict consensus is that it includes only clades that are unambiguously supported by the available evidence (thereby summarizing objective knowledge of relationships) and that any of the other trees that he suggests will include groups that are contradicted by equally optimal solutions.

It actually appears that Hall does not consider summarizing the unambiguously supported groups to be a goal of consensus representation. Consider his statements on polytomies (p. 84, italics in original):

> Multiple trees are often the result of very real *polytomies* in the tree. Like most of us, phylogeneticists prefer to keep things simple. The simplest situation is a strictly bifurcating tree: from every internal node there are exactly two branches... Sadly, evolutionary history is not always so simple, and at times an ancestor may have given rise to multiple descendents within such a short time span that the order of descent cannot be resolved.

Sadly, Hall overlooks the fact that a method by which one could distinguish between "very real *polytomies* " and those due to inadequate or ambiguous data (which he admits may result in polytomies earlier, on p. 48) has yet to be proposed, and his claim that this is "often" the case is completely unfounded.

The difficulties posed by consensuses and polytomies prompt Hall to ask (p. 85),

> Is the inconvenience of dealing with consensus trees a reason to simply accept the Neighbor-Joining tree and get on with it? Not necessarily. Compare figure 2.5 with figure 1.46. Both are derived from the same data. Which is a more accurate representation of history? If the polytomy is real, there is a problem with the NJ tree in that in PAUP* distance trees are strictly bifurcating—no polytomies allowed... Sometimes the best thing is to turn to another method.

By this he means maximum likelihood. Yet, as far as we are aware, no one has ever claimed that maximum likelihood is able to determine whether a polytomy is real. In fact, turning to another method is likely to exacerbate Hall's dilemma, because the maximum likelihood tree may differ from both the neighbor joining and parsimony trees.

Hall also misrepresents statistical methods of phylogenetic analysis. For example, (p. 73; see also p. 101):

> Bayesian analysis is a recent variant of Maximum Likelihood. Instead of seeking the tree that maximizes the likelihood of observing the data, it seeks those trees with the greatest likelihoods given the data.

First, if by "recent variant" he means that both methods employ a likelihood term and evolutionary models he is correct, but that is where the similarities end. The theoretical underpinnings of Bayesian estimation are generally understood to be fundamentally different from those of maximum likelihood estimation, and the analytical procedures are completely different.

Second, maximum likelihood maximizes the *likelihood* of the hypothesized tree (given the observed data and the model), which is proportional to the *probability* of observing the data (given the model and the hypothesized tree). Bayesian analysis maximizes the *posterior probability* of the hypothesis (given the data and the model). This may seem trivial, but Hall's description misrepresents the difference between maximum likelihood and Bayesian analysis, and such errors are bound to confuse people just coming to systematics—especially if they have a statistical background, because the distinction between likelihoods and probabilities has far-reaching theoretical consequences (likelihoods do not obey the axioms of the probability calculus, for example).

Third, the Bayesian method Hall refers to does *not* seek "those trees with the greatest likelihoods" or "the *best set* of trees" (p. 101, italics in original) or even the set of trees with the greatest probabilities. The theoretical basis of Bayesian analysis using the Markov Chain Monte Carlo method is that it obtains a mixed sample of trees with low, moderate, and high likelihood scores, the frequency of a given tree being an estimate of its posterior probability. The acceptance probability of the general Metropolis–Hastings algorithm (Hastings, 1970) is designed to achieve precisely that frequency (Tierney, 1994). If Hall's description were correct, the Metropolis–Hastings algorithm would not be needed to generate the distribution, as one would only have to obtain the maximum likelihood tree and the desired number of next most-likely trees. The burn in (which Hall defines only as "the number of trees… that will be ignored when the consensus is created," p. 107) is needed not to maximize the likelihood (because the actual maximum likelihood solution is irrelevant) but to avoid starting-point sensitivity and prevent the frequency distribution from being distorted by trees not sampled according to their posterior probabilities. Nor is "the frequency of any particular tree in that set of saved trees… taken as the probability that it is the best of the equally likely trees" (p. 102); those trees are not equally likely, and the frequency is taken as the probability that the tree is true. Furthermore, the tree(s) with the greatest likelihood (or the greatest posterior probability) may be contradicted by the MrBayes tree (W. Wheeler, pers. comm.). Just as it is possible for groups absent from the most parsimonious tree to be present and apparently well supported in the parsimony jackknife tree (and vice versa), it is possible for groups absent from the maximum likelihood (or the most probable) tree to be present and apparently well supported in the MrBayes tree (and vice versa). It is unfortunate that Hall's close association with the development of MrBayes (see Acknowledgments and Huelsenbeck et al., 2001) did not enable more accurate representation.

Hall concludes his discussion of the "advantages and disadvantages" of different methods by observing that (p. 77),

> In the end, it probably matters little which method you use. NJ, Parsimony, ML, and Bayesian analyses are all perfectly respectable methods that will be accepted by most journals and most readers as valid. It is often reasonable to use all four methods.

Respectability aside, these four methods rely on fundamentally different assumptions, and what matters scientifically is the objective validity of those assumptions and not whether a journal will accept them. Hall continues,

> If your data are good, in the sense that the sequences you chose really are related by descent, and your alignment is robust, then the trees will be so similar that the differences won't matter.

Under this definition one can only conclude that all data must be good (assuming that life is monophyletic). In any case, Hall's pragmatic advice is epistemologically flawed (robustness to variation in assumptions has no bearing on either the objective validity of those assumptions or the strength of evidential support for competing hypotheses) and practically unhelpful, because empirical results often differ among methods.

The problem of rooting phylogenetic trees is of great concern in *Phylogenetic Trees Made Easy*. And so it should be, because Hall seems to believe that choice of root determines tree length, at least in parsimony (p. 81):

> The other possible rootings of the tree are considered in the same way, and if a different rooting of the tree produces fewer changes, that is the score for that site.

Of course, the placement of the root actually has no effect on the number of changes (at least not in standard character analysis), which may be why Hall also advises that (p. 52)

> The least arbitrary (and therefore always correct) means to present the tree is to use the unrooted phylogram method.

This "method" is "correct" only if the "phylogram" is not intended to depict phylogeny, because it says nothing about the evolutionary history of the lineages. Minimally, we suggest that it would have been useful for the uninformed molecular biologists to have been told that rooting is necessary to determine monophyly.

Hall goes on to describe in some detail midpoint rooting and outgroup rooting (he does not mention other methods of polarizing characters). He uses those terms in different ways in different parts of the book. In discussing outgroup rooting, he defines the outgroup as "a designated outsider to the rest of the sequence [sic]" (p. 53) and, further, that "[a]n outgroup is a taxon that is more distantly related to each of the ingroup taxa than any of the ingroup taxa are to each other" (p. 54). Accordingly, one can conclude only that outgroup rooting involves the designation of a single taxon as the root. Hall defines midpoint rooting as "designating a point in the middle of a branch as the common origin of these sequences" (p. 53), which would apply both to midpoint rooting as usually defined and to cases where the branch between the ingroup and the multiple outgroup taxa is treated as the root. But then on p. 55 Hall reveals that "[a]n outgroup need not consist of a single sequence" and in Section 3 he discusses how to do outgroup rooting by placing the root between the ingroup and an outgroup consisting of multiple terminals. It is simple enough for a systematist to see that Hall conflates the concepts of "outgroup" and "root" in his discussions, but we can imagine how confusing this must look to a novice.

To his credit, Hall does provide an accurate definition of midpoint rooting on p. 54, and he even points out that midpoint rooting relies on the assumption of constant evolutionary rates, which may be highly problematic. Unfortunately, the rest of the discussion of rooting is not as well informed. For instance, in figures 3.1–3.4 (pp. 116–119) several unrooted trees that each have an extremely long internal branche are shown. Hall observes that although all these trees are unrooted, "it seems likely that these two clades descended from a common ancestor so long ago that there have been many changes since diverging from that ancestor" and that "[t]o convey all that information, we need to root the tree" (p. 118). Moreover, Hall suggests that "when the two clades are separated by a very long branch, midpoint rooting makes perfectly good sense" (p. 119) and that outgroup rooting is relevant only when there is not "an especially long branch between two clades to help us decide where to place a root" (p. 120). Overlooking the inappropriate use of the term "clade" in reference to unrooted trees, this suggestion would certainly be misleading if a small clade nested within a monophyletic group is extremely divergent from the rest of the taxa.

Hall characterizes the problem (brought about, according to him, by unequal evolutionary rates; see p. 54) of rooting on a taxon that is too distantly related as (p. 55, boldface in original)

> A distantly related sequence may be so distantly related that it does not share a common ancestor with the ingroup sequences, i.e., it is not **homologous**.

This conflates two problems that should be addressed separately. First, the sequences being compared must be homologous (and orthologous, although Hall never mentions the problem of paralogy), regardless of their degree of similarity or divergence. Second, the taxon used to root the tree must not be so distantly related that its states are effectively randomized with respect to the ingroup. It is generally accepted that the most effective means of avoiding this is to include numerous intermediate taxa between the root and the ingroup to break up what would otherwise be an excessively long branch, but Hall does not suggest this strategy.

The third subject that Hall addresses as part of "creating" phylogenetic trees is reliability, which can be assessed by bootstrapping or Bayesian probabilities. Although we agree that a comprehensive discussion of the many methods of assessing support is not necessary in this beginner's manual, the other common ones (e.g., parsimony jackknifing, Bremer support) should have at least been mentioned to orient newcomers who will find references to them throughout the literature.

According to Hall (p. 60 italics in original),

> Reliability is measured as *the probability that the members of a given clade are always members of that clade*.

We are not sure exactly what that means, but it certainly is not a definition that we have read before. Given that Hall endorses a probabilistic interpretation of reliability, a more intuitive (and accurate) definition for beginners to grasp would be simply *the probability that the members are truly (really, actually) members of that clade.*

Hall goes on to explain that experimental scientists test the reliability of a conclusion by repeating the experiment with independent data, whereas in systematics (p. 60, boldface in original),

> Since the data in this case are the sequences themselves, and sequences are what they are, there seems to be little point to repeating the data unless we just want to test the reliability of the sequencing... Phylogeneticists use a sampling method called **bootstrapping** that pseudorepeats data collecting as a method to estimate the reliability of the tree.

Hall apparently does not recognize that the equivalent of repeating an experiment with independent data is gathering additional sequences from new specimens or loci, not repeating the sequencing. Repeating the sequencing is equivalent to the experimental scientist reviewing his notes! Furthermore, bootstrapping is at least as applicable in experimental sciences as it is in phylogenetics. Hall's description of pseudoreplicates is adequate, but there is no discussion about why this is believed to assess reliability or what assumptions must be met to justify a probabilistic interpretation (e.g., random sampling). Likewise, Hall's description (pp. 111–113) of the steps involved in importing trees into PAUP* and determining the Bayesian probability of each clade is straightforward, but there is no discussion of the basis for treating this as a posterior probability or the assumptions that must be met. This lack of discussion would not be a deficiency had Hall actually written this book as a practical manual; but, because he also claims that this is an important advantage of Bayesian analysis (e.g., pp. 73, 111), some discussion is needed.

Summary

In general, the step-by-step instructions in *Phylogenetic Trees Made Easy* are accurate but incomplete. Even practical discussions on using commercial graphical software packages (e.g., Corel, Canvas) to edit and print trees are at best incomplete, because there are numerous phylogenetic software packages designed specifically for that purpose (e.g., MacClade, TreeView, WinClada) that in some cases are freely available online. Likewise, Appendix I presents a concise description of several file formats for sequence data (FASTA, Clustal, Nexus, Phylip, GCG/MSF, NBR/PIR) and a brief introduction to the interconversion of these formats using PAUP*, but it omits common data file formats (e.g., HENNING86NONA) and software for managing sequence data (e.g., BioEdit, MacClade).

The seemingly arbitrary choice of topics to be covered in depth is a further general deficiency of this book. For example, exhaustive tree searching (which is relatively unimportant in modern research) is covered in extensive detail, but branch swapping is all but ignored. Similarly, a Learn More box reviews the details of several evolutionary models, but there is no mention of the problem of choosing which of those models to use. Perhaps the greatest disparity is in the coverage of alignment, which includes 33 pages dedicated to generating, interpreting, refining, and even printing alignments using ClustalX, but does not provide any details whatsoever on what ClustalX actually does or discuss alternative approaches to treating indels in phylogenetic analysis.

What is most astonishing is that Hall apparently does not see his admitted lack of understanding (see p. vii) as a limitation, but instead licentiously passes on his novel perspectives and suggestions to unsuspecting molecular biologists with little or no background in systematics. Almost without exception, Hall's theoretical discussions are false, incomplete, misleading, or contradictory. We can only assume that the manuscript of the book was never reviewed by a trained systematist; several are acknowledged as having taught Hall systematics, but none for having read or corrected the manuscript. Hall's warnings (e.g., p. 76, italics in original) that his procedures and justifications are merely "*my* reasons for a preference, not general reasons" and that "[t]hey are personal and should not be interpreted as recommendations" are disingenuous, as they are obviously meant to instruct beginners on designing and running phylogenetic analyses—that being the purpose of the book. To call *Phylogenetic Trees Made Easy* nothing more than a "cookbook" is misleading and serves only as a poor excuse for bad theory. By claiming to assist in the "transition between a theoretical understanding of phylogenetics and a practical application of the methodology" (p. 1), Hall is able to make far-reaching claims without ever having to defend them. And by specifically targeting an audience that lacks formal systematics training and is therefore unable to discern well-founded theory from unsubstantiated conjecture, the book is ideally positioned to have maximal influence while attracting minimal criticism.

What we found most frustrating about this book is that it is a missed opportunity to move away from 'point-and-click' systematics. To that end, future authors and publishers can learn much from the shortcomings of *Phylogenetic Trees Made Easy*. To begin with, rather than trivialize the differences between the methods, newcomers to the field should be exposed to the arguments directly. Only by understanding the assumptions that underlie the different methods can they make an informed choice. Characterizing the ongoing debates in systematics as "religious conviction" (p. 69) hardly encourages students to consider seriously the basis of the disagreements. Furthermore, we suggest that a choice be made early in the book-planning phase between writing (1) a software manual, (2) a theoretical text, or (3) a combination. Modern phylogenetic studies cannot be carried out without the assistance of computer programs, and computer manuals are necessary to use and understand them. If this book had remained within the limits of a practical software manual, it would have served a useful purpose. Likewise, theoretical texts are useful, especially those that present phylogenetic concepts in a way that facilitates their integration into other fields of biology. However, though undoubtedly more challenging, the combined approach is by far the most useful option, provided that an effort is made to explain and defend the theory and couple it directly with the commands; a prime example of such an effort is the manual for MacClade (Maddison and Maddison, 1992). A manual covering multiple software packages and equipped with a readable, accessible theoretical discussion of the basis of the commands would allow users to understand the reasoning behind the options that they click, and that would represent an important step away from 'point-and-click' systematics. The problem with *Phylogenetic Trees Made Easy* is that Hall advertised (1) and tried to write (3), but lacked the theoretical background to do the job properly.

To the pessimist, the fact that such a book could be published, distributed, sold, and recommended illustrates the extent to which 'point-and-click' systematics has achieved predominance and suggests that the rising preeminence of phylogenetic systematics is now, more than ever, on the verge of self defeat. To the optimist, it underscores the broader scientific community's desire to incorporate phylogenetic systematics into their research programs and reveals a prime opportunity to educate colleagues and newcomers on the perils of 'point-and-click' systematics. To the pragmatist, *Phylogenetic Trees Made Easy* represents $29.95 US misspent.

References

Farris, J.S., 1983. The logical basis of phylogenetic analysis. In: Platnick, N.I., Funk, V.A. (Eds.), Advances in Cladistics. Columbia University Press, New York, pp. 7–36.

Farris, J.S., Albert, V.A., Källersjö, M., Lipscomb, D., Kluge, A.G., 1996. Parsimony jackknifing outperforms neighbor-joining. Cladistics 12, 99–124.

Farris, J.S., Kluge, A.G., Mickevich, M.F., 1982. Phylogenetic analysis, the monothetic group method, and myobatrachid frogs. Syst. Zool. 31, 317–327.

Feng, D.F., Doolittle, R.F., 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J. Mol. Evol. 25, 351–360.

Ghiselin, M.T., 1976. The nomenclature of correspondence: A new look at "homology" and "analogy." In: Masterton, R.B., Hodos, W., Jerison, H. (Eds.), Evolution, Brain, and Behavior: Persistent Problems. Erlbaum, Hillsdale, NJ, pp. 129–142.

Goloboff, P.A., 1993–1999. NONA. Ver. 2.9. Published by the author, Tucumán, Argentina.

Goloboff, P.A., 1996. Methods for faster parsimony analysis. Cladistics 12, 199–220.

Goloboff, P.A., 1999. Analyzing large data sets in reasonable times: solutions for composite optima. Cladistics 15, 415–428.

Goloboff, P.A., 2002. Techniques for analyzing large data sets. In: DeSalle, R., Giribet, G., Wheeler, W.C. (Eds.), Techniques in Molecular Systematics and Evolution. Birkhäuser, Basel, pp. 70–79.

Goloboff, P.A., Farris, J.S., 2001. Methods of quick consensus estimation. Cladistics 17, S26–S34.

Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. 41, 95–98.

Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57, 97–109.

Hennig, W., 1966. Phylogenetic Systematics. University of Illinois Press, Chicago.

Huelsenbeck, J.P., Ronquist, F., Hall, B., 2001. MrBayes: A program for the Bayesian inference of phylogeny. Published by the authors, available at http://morphbank.ebc.uu.se/mrbayes/.

Maddison, W.P., Maddison, D.R., 1992. MacClade Version 3. Sinauer Associates, Sunderland, MA.

Needleman, S.B., Wunsch, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequences of two proteins. J. Mol. Biol. 48, 443–453.

Nixon, K.C., 1999. The parsimony ratchet, a new method for rapid parsimony analysis. Cladistics 15, 407–414.

Nixon, K.C., 1999–2002. WinClada. Ver. 1.0000. Published by the author, Ithaca, NY, USA.

Phillips, A., Janies, D., Wheeler, W.C., 2000. Multiple sequence alignment in phylogenetic analysis. Mol. Phylogenet. Evol. 16, 317–330.

Reeck, G.R., de Haen, C., Teller, D.C., Doolittle, R.F., Fitch, W.M., Dickerson, R.E., Chambon, P., McLachlan, A.D., Margoliash, E., Jukes, T.H., Zuckerkandl, E., 1987. "Homology" in proteins and nucleic acids: A terminology muddle and a way out of it. Cell 50, 667.

Rice, K.A., Donoghue, M.J., Olmstead, R.G., 1997. Analyzing large data sets: *rbc*L 500 revisited. Syst. Biol. 46, 554–563.

Schmidt, H.A., Strimmer, K., Vingron, M., Haeseler, A.V., 1999–2000. Tree-Puzzle 5.0. http://www.tree-puzzle.de/.

Tierney, L., 1994. Markov chains for exploring posterior distributions. Ann. Stat. 22, 1701–1762.

Tuffley, C., Steel, M., 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull. Math. Biol. 59, 581–607.

Wheeler, W.C., 1994. Sources of ambiguity in nucleic acid sequence alignment. In: Schierwater, B., Streit, B., Wagner, G.P., DeSalle, R. (Eds.), Molecular Ecology and Evolution: Approaches and Applications. Birkhäuser, Basel, pp. 323–352.

Taran Grant,[a,b,*] Julián Faivovich,[a,b] and Diego Pol[a]
[a] *American Museum of Natural History*
Central Park West at 79th Street
New York 10024, USA
[b] *CERC/Department of Ecology*
Evolution, and Environmental Biology
Columbia University
1200 Amsterdam Ave.
New York 10027, USA
E-mail address: grant@amnh.org

[*] Corresponding author