ACADEMIC
PRESS

Book Review

Easy trees?—Review of "Phylogenetic trees made easy: a how-to manual for molecular biologists" by Barry G. Hall

Within the biological sciences, phylogenetics has long been relegated to a selected sector of scientists interested in the evolutionary history of organisms. Recently, it has become widely accepted that molecular sequence data is best analyzed in a phylogenetic context if predictive statements are to be made concerning lineage-based relationships among genes in large gene families, origins of genes via events such as speciation and gene duplication, potential function of newly discovered gene sequences, and reconstruction of ancestral sequences. Many molecular biologists have been turning to the phylogenetic literature or consulting their local evolutionary biologist in order to interpret molecular sequence data in a phylogenetic context, often finding themselves overwhelmed by the number of available programs and methods for phylogenetic analysis of sequence data. With the introduction of Barry Hall's "Phylogenetic trees made easy" How-to manual, any researcher with data in the form of molecular (DNA or protein) sequences and a desktop computer (preferably Macintosh) can easily build a branching hierarchy from those data. Step-by-step instructions are given that take the reader through the process of building a tree, starting with the acquisition of homologous sequences, and finishing with the aesthetics of printing and publishing trees. As the first book to make tree-building a step-by-step process, Hall's *Phylogenetic trees made easy* has been heralded as a "tutorial manual that can initiate (phylogenetic novices) into this important field of biology" (Van de Peer, 2001). But what kind of initiation is the reader receiving? Is it merely a form of hazing? The question arises; does this manual truly enable a researcher to produce a valid phylogenetic representation of their hard-earned data? Or does it over-simplify a complex analytical process potentially resulting in weak analytical results or, worse yet, misleading interpretations of phylogenetic signal and misinterpretation of evolutionary scenarios.

PTME is a self-proclaimed cookbook, "intended as a tool to aid beginners in creating phylogenetic trees..." This is a formidable task. Hall has attempted to take the analytical tools of an entire field of biology and reduce them to a list of point and click options. He has done so in an easily accessible, user-friendly format, displaying pull-down menu options and summarizing (particularly

well for Bayesian methods, page 109) the steps required to make a tree. The book is easier to understand and contains more hands-on instruction than the manuals available for the programs utilized in the text. However, Hall fails to indicate choices being made along the way. In the process of detailing which buttons to click and when, Hall does not inform the reader about the decisions being made concerning phylogenetic inference, or the effects these decision may have on the way the data is treated in the analysis. While Hall claims that the book helps in "making the transition between a theoretical understanding of phylogenetics and a practical application of the methodology," there is no evidence of transition. Hall ultimately advocates using phylogenetic tools without fully understanding the theoretical underpinnings, and advises users to construct trees without considering important aspects of their data and the potential impacts on phylogenetic reconstruction.

Hall compares his book to a commercial kit (page 2), justifying its simplicity by stating that, molecular biologists can use the kits to produce a reliable product without understanding the underlying methods or chemistry involved. Equivalently, it would seem, with the help of his manual one can construct a "reliable, valid" phylogenetic tree without understanding the underlying assumptions. However, there is a major difference that becomes apparent when following the steps outlined in Hall's book. Following the protocol of a kit does not produce the results of your research; it merely enables you to complete a tedious task in significantly less time. The protocols outlined in Hall's book also may help to save time, or enable production of tress without fully understanding the underlying "chemistry" (to keep with the kit analogy). But what you get from following Hall's protocols is a final analytical product upon which conclusions concerning evolutionary history are based. Simplification of data analysis and interpretation of results should never be compared to the simplification of a molecular or chemical procedure as available in commercialized kits. The protocols in Hall's How-to manual impart a delusion of simplicity in phylogenetic reconstruction. Unlike with a kit, understanding the principles does more than "help troubleshoot the phylogenetic problems that arise..." (page 2). An understanding will actually help the researcher to determine reliability and validity of the phylogenetic reconstruction.

Along this vein, Hall does provide several opportunities to develop a greater understanding. Scattered throughout the book are nine "Learn More About" boxes that provide additional information about topics of concern to users of phylogenetic methods. Hall suggests that these boxes serve as optional reading to provide the user with a more detailed understanding of the protocols outlined in the text. While the boxes do indeed provide more details about the intricacies of phylogenetic reconstruction, they do not treat issues such as how these methods handle different types of data. Referring the reader to some key literature would be more appropriate. The boxes could have provided a venue to introduce further reading, but Hall did not take advantage of this opportunity. Only one reference is made in each box, and usually to a very general overview paper rather than primary literature involving real data. In general, Hall does not provide the reader with easy access to primary literature references of interest. His Literature Cited section alone is testament to the lack of scientific consideration. With only 17 references, 4 of which are software packages, the reader is not given any idea of the vast amount of publications and primary literature that address the very issues that the author argues are not necessary to understand in order to create reliable and valid phylogenetic reconstructions. The only additional reading providing a level detail beyond that found in the boxes are two general textbooks and two review chapters.

The first section, "Create a Tree," is by far the longest section of the book. This section takes you through the assessment of homology, alignment of sequences, and creation of a publication-quality tree using Neighbor Joining (NJ). There is strong focus on assessment of homology at the level of sequence selection and sequence alignment, which is indeed an important first step in the phylogenetic process and something often ignored in the rush to produce a tree. As the tree is being created, those who are truly interested in what they are doing can opt to read the "Learn more about Phylogenetic Trees" box. This box, actually explains what it is that you are creating. If the point of this tutorial is for a beginner to create a phylogenetically meaningful branching diagram for publication, it seems as though it would be hard to interpret the results without realizing that "a tree is composed of branches that intersect and terminate at nodes" or understanding the significance of a bootstrap value, or jackknife value, or for that matter Bremer Support (noticeably lacking in the book). The use of boxes to segregate background information from step-by-step instructions is helpful in organizing the text, but we strongly suggest that any reader consider the boxes as an integral part of the manual, and not be tempted by Hall's offer to point and click without understanding the basics.

After the reader has completed the How-to instructions of the first section and successfully created and printed a phylogenetic tree (which can be touched up for publication with the advice on "printing and presenting" trees given in Section 3), the second section introduces the reader to the additional methods for analyzing data in a phylogenetic context. This seems a bit out of order considering the desired tree is already built and saved and/or printed. These methods should be discussed first, as an introduction to the diversity of analytical tools available. As presented, the reader is unlikely to see any benefit in repeating the steps just to try out another method, especially after reading Hall's cursory discussion of the "additional" methods available and the supposed "religious convictions" associated with an evolutionary biologists' preference for using one versus another. Based on his discussion, it would seem as though there are no really good reasons for opting to use one method of data analysis over another. In fact, on page 83 Hall mentions that a choice between two most parsimonious trees can be made by comparison of the parsimony trees to the NJ tree and selection of the most parsimonious tree that best matches the topology NJ tree, implying that the best topology is the one most easily acquired. If this is the case, why would the reader bother with the parsimony tree at all? It makes no sense. Especially when considering Hall's blatant promotion of time-saving techniques.

While Hall does discuss a total of four desirable performance criteria in the methods used to construct phylogenetic trees, (efficiency, robustness, computational speed, and discriminating ability), his overwhelming favorite seems to be computational speed. As an example, on page 76, Hall states

> "...I am willing to use a method that will run overnight while I am home." (Those of you without tenure may well wonder about the full 14 h that Hall considers "overnight").

Following this admission, one notes that much of Hall's advice is based on the time that it takes to run an analysis or prepare the data for analysis rather than on the accuracy or robustness of the results. While it is true that phylogeneticists are concerned with speed, speed itself is a factor that encourages us to produce faster algorithms and even to construct faster computers. It is never considered as a factor that encourages sloppy or incomplete data analyses!

Throughout the book, many arbitrary decisions are made for the reader and are left unexplained, for example the use of 50% majority rule for consensus trees and presenting bootstrap and jackknife values, but neither discriminating between the computational differences of the latter two nor providing references to the germane literature. (Is the author aware of the fact that the 50% majority rule tree may not have a topology found in any of the original trees?) The type of

consensus tree presented makes a difference when interpreting evolutionary relationships as well as robustness of phylogenetic signal and reconstruction at ancestral nodes. It would seem that the more conservative approach would be to choose strict consensus, and a decision to do otherwise merits explanation. On page 77, Hall goes through the steps of creating a parsimony tree. Although there is a box that discusses tree searching methods, Hall gives the instruction to choose Heuristic search and leave everything in its default state. The discussion concerning searching algorithms provided in the "Learn More" box on page 70 is not even referred to in the text. Based on the overall tone of the book, we presume that Hall would respond that it does not really matter which choices are made at these junctures. Which is exactly our salient point. A reader could easily walk away from this book believing that certain decisions (which tree construction method to use, which most parsimonious tree to publish, how to create or interpret a consensus tree, how to determine reliability or robustness of the results) are arbitrary, or even worse readers may remain completely unaware of the types of decisions that are being made and the impact of those decisions on data analysis and interpretation.

Considering his ability to neglect the importance of these decisions in making "reliable and valid" phylogenetic trees, it is understandable why Hall neglects to include citations of the relevant literature. Oddly enough, the only primary literature cited is with reference to the Bayesian approach, an obvious favorite of Dr. Hall's. This is a method which Hall describes as being very new and remaining largely unfamiliar to most systematists. One should be curious as to why Hall believes that this is so. When a novel technique becomes available to molecular biologists, do they not read the relevant journals to learn about it, discuss its merits, and perhaps even attempt to use it if they find those merits worthwhile? Why would systematists be any different? Hall does not list any citations for NJ, ML, or Parsimony, yet he lists three primary literature citations for Bayesian (...to prove its infamiliarity??). Furthermore, on page 109, a box summarizing MrBayes outlines the exact steps needed to produce a tree using the Bayesian algorithm. As this book primarily functions as a how-to manual, this type of outline and this level of primary-literature citation would be helpful for all methods discussed, including the neighbor joining exercise detailed in Section 1.

Skipping ahead to Section 5, the reader is led to believe that only MrBayes can be used to reconstruct ancestral sequences, nevermind that the reconstruction of ancestral character states has been a major undertaking in evolutionary biology long predating the arrival of MrBayes (See various references in journals such as Cladistics, MPE, MBE, and Systematic Biology among others). No mention is made as to the effects of the use of prior probabilities in constructing ancestral states, and Hall merely states that "MrBayes provides a better and certainly much more convenient way to reconstruct sequences at interior nodes." As reconstruction of ancestral states is important in gene tree analysis, this topic should be dealt with in greater detail with access to alternative programs that do not employ Bayesian methods. Most phylogenetists and evolutionary biologists concerned with the reconstruction of ancestral character states employ a variety of easy to use and graphically-pleasing programs. One of these is the popular MacClade (Maddison and Maddison, 2000), which will reconstruct character states based on the parsimony in a menu-driven format. Several other programs, such as WinClada (Nixon, 2000), also include algorithms for reconstructing character states using parsimony although in a slightly less user-friendly manner.

In the final section, Hall presents us with a series of opinions concerning: (a) the use of DNA sequences vs Protein sequences in creating a phylogeny and (b) the effects of including non-homologous sequences on phylogenetic topology. This second topic was touched briefly in section one as well, couched in a discussion of the definition of homology and the selection of homologous sequences for phylogenetic reconstruction. As this is actually a very important matter, some references to the primary and review literature would be useful. It is clear that Hall is aware of the importance of these issues and their impact on data analysis, but his methods of dealing with these difficult issues are more based on avoidance than on understanding. While Hall earlier explains that similarity and homology are not equivalent, here he proceeds to describe how to use similarity analyses to determine potential homology. The discussion on the use of DNA vs Protein sequence data also falls short. There is a good discussion of the use of different types of data for different levels of phylogenetic reconstruction, but no mention of the fact that using protein sequences is functionally decreasing the amount of data used in the analysis and the effects this may have on the phylogenetic interpretations (Cladistics, MPE).

The real value of this book is its guide through the myriad of pull-down menus (and pull-downs of the pull-downs) in various programs. Most likely, many previous users of these programs will discover options they did not know existed. Unfortunately, Hall's manual gives no basis for choosing amongst various options found among the pull-downs, nor does he provide any access to the literature which might enable the reader to make informed choices. These, and his obsession with speed, leave too much left behind. His only true guide to making decisions is to make one choice; that which is computationally fastest.

Maybe the bottom line is that reliable and valid phylogenetic trees cannot be made easy.

References

Maddison, D.R., Maddison, W.R., 2000. MacClade 4: Analysis of Phylogeny and Character Evolution. Sinauer Associates, Sunderland, MA.

Nixon, K.C., 2000. WinClada. Published by the author, Ithaca, NY.

Van de Peer, Y., 2001. Phylogeny branches out. Book Review. Nature 414 (6863), 490.

Chelsea D. Specht, Dennis Wm. Stevenson
The New York Botanical Garden
Bronx, NY 10458
USA
E-mail addresses: chelsea@nybg.org (C.D. Specht),
dws@nybg.org (D.Wm. Stevenson)