[Review published in *SIAM Review*, Vol. 59, Issue 2, pp. 450–458.]

Phylogeny: Discrete and Random Processes in Evolution. *By Mike Steel.* SIAM, Philadelphia, 2016. $64.00. xii+297 pp., softcover. ISBN 978-1-611974-47-8.

Mike Steel has contributed enormously to the mathematical aspects of phylogenetics and does the field a major favor in summarizing its key results. Phylogeneticists would do themselves a major favor by becoming more familiar with these results.

This is a high-quality book by one of the leaders in the field. It covers combinatorial properties of trees, different representations, their enumeration, and metrics on tree space, as well as the possible composition and decomposition of trees into subtrees and the compatibility of trees. Tree inference from characters and distances, stochastic models on trees, and models for evolution or generation of trees are discussed. It is a mathematical book, but one with clear applications. The presentation throughout is generally succinct, and as such it covers a large amount in its 300 pages, including many results since *Phylogenetics* [9] was published with Charles Semple 12 years ago. As an introductory book it is not an easy read; there are often a number of new concepts presented on a single page. However, plentiful figures provide an additional, and often easier, way to understand the material. There are few numerical examples or examples taken from real applications, which could have provided stronger insight into the applications as well as the concepts themselves.

The book contains a lot of new material on stochastic models which underlie many current biological applications. There is a final chapter on networks, but these are not covered in depth. The ancestral recombination graph is only mentioned in passing, which is a significant omission since it describes the relationships among a set of sequences sampled from a population. Important algorithmic results are highlighted, although the issues are rarely discussed, which seems to be a sensible choice for maintaining the focus of the book. The discussion of tree inference methods presents a number of important results which are of significance to current disagreements in the literature over conflicting, seemingly well-supported phylogenies. Overall, it is both a very readable introduction to the subject as well as an overview of many important recent results and current research topics. Some parts of the book might be hard reading for the average biologist, but many in the field have a strong background in statistics or algorithms, and for them the book is well worth the effort.

The book is logically structured, and the preface contains advice on different paths that could be taken through the book. Each chapter has a few exercises embedded in the text. The presentation style is very different from the earlier *Phylogenetics* [9] in that it gives more space to intuitive explanations and less to theorems followed by rigorous proofs; however, the level of mathematics is about the same.

Chapter 1. Phylogeny. Chapter 1 covers the prerequisites: *graphs*, *unrooted*, and *rooted trees*, as well as some structures that will be put to use later on in the book—the *intersection graph* and *chordal graphs*. A *phylogenetic tree* is defined; this is the central object of interest for the book in contrast to the *X-tree*, which was used in the author's earlier book, *Phylogenetics* [9]. Rooted and unrooted trees are treated in parallel throughout the book as properties of the one frequently have direct equivalents in the other.

Chapter 2. Basic Combinatorics of Discrete Phylogenies. Despite being only 30 pages long, Chapter 2 covers a lot of ground—a feat repeated frequently throughout the book! It begins with the *enumeration of trees*. We see Cayley's formula for the number of trees with all nodes labeled. The formula, $n^{(n-2)}$, is simple, but deceptively so. Bifurcating trees and spanning trees are then considered in a similar manner. The results are obtained using formal power series in a maddeningly compact page that should be expanded into at least four pages.

Steel then presents two key equivalences

that are central to much of the book and the subject itself: a rooted tree can be represented as a *hierarchy* of sets on the leaf-set and an unrooted tree as a system of *splits* (bipartitions) on the leaf-set. He then explores a series of alternative characterizations of hierarchies and their equivalence to rooted trees. In a hierarchy of sets, any two sets must satisfy one of the following: one of the sets is a subset of the other, or they are disjoint. An unrooted tree is shown to be encoded by a set of *quartet* trees and a rooted tree by a set of *rooted triples*.

The set representation of phylogenetic trees leads naturally to the concept of the *refinement* of trees. Loosely speaking, a phylogenetic tree may not be fully resolved and a tree, T, is a refinement of a second tree, T', if it is equivalent to T' but resolves more of the edges. A set of trees is *compatible* if there is a tree that is a refinement of all the trees in the set.

The *Buneman graph* makes an appearance here and is revisited later in the book when multistate characters are discussed. The Buneman graph is a graph which may be constructed from an arbitrary set of splits, and is a phylogenetic tree precisely when the splits are pairwise compatible. It will reappear when Steel considers data that cannot be represented by a tree, but we are getting ahead of ourselves.

Three more ways to encode an unrooted tree are discussed (other than as an unrooted tree itself or as a system of splits), the most fun being *circular orderings*, which are natural when there is no root and the leaves can hang down and be placed on a line. Next is a fun section on the size of neighborhoods and the diameter of *tree-space* with respect to rearrangement operations (*NNI*, *SPR*, *TBR*) on trees. It would be fun to sit and puzzle with this section and discuss which other operations are imaginable. Then comes more on the metric space of trees with edges (edit operations). Finally, consensus functions are discussed. Important here are the *strict consensus*, the *majority consensus*, and the *Adams consensus*. The Adams consensus is only defined for rooted trees and satisfies some desirable properties not satisfied by the other consensus methods. Furthermore, an impossibility result is given showing that no method is possible that satisfies a more relaxed, seemingly desirable property for a consensus function!

Chapter 3. Tree Shape and Random Discrete Phylogenies. Chapter 3 discusses the shape of trees generated according to a number of models. First, we see how to enumerate the number of phylogenetic trees using the orbiter-stabilizer theorem to handle symmetries that do not change the tree, such as the symmetry of flipping the children of an inner node. The *Yule–Harding* model for evolving trees and *uniformly sampled trees* are discussed. The "big picture" is presented (with a wonderful figure) showing how you can link and count trees with different kinds of labeling. Gaining an understanding of these details was very rewarding. There are six classes of trees and seven arrows between them, each showing a simple combinatorial conversion.

Measures of *tree imbalance* are defined (Colless and Sackin). Yule–Harding trees are more balanced than uniformly sampled trees and observed trees tend to be somewhere between the two. The *Aldous beta-splitting model* for generating trees is presented. The beta parameter varies the amount of imbalance and the Yule–Harding and uniform model trees come out as special cases of this model. An interesting question is to ask what biologically realistic models could lead to Aldous beta-splitting trees, since naive assumptions of independence will lead to Yule–Harding. There is a very nice generalized urn model with cherry, noncherry, and pendant edges that can be used to calculate the distribution of cherries in a Yule–Harding tree, showing how using a Polya urn model formulation can make certain theorems easier. We all loved this section.

Chapter 4. Pulling Trees Apart and Putting Trees Together. Chapter 4 considers how trees can be analyzed and described in terms of trees restricted to subsets of labels (taxa). Conversely, how can trees be combined into a "supertree," for example, in the construction of a "tree of life." The chapter discusses *tree restriction* to a subset of labels and its converse, the *display* of a tree by a tree on a superset of labels. *Compatibility* and *consensus trees* are discussed, as is the question of when a

set of trees on subsets uniquely *defines* a binary tree. Also, when is the restriction of trees to a collection of subsets *decisive* in distinguishing between the two trees.

Results on when the members of a set of trees are compatible (can all be viewed as restrictions of some common larger tree) are explored. The problem is typically harder for unrooted trees, where you need to go via a *display graph*. After triangulation of this graph, simple criteria use the tree width to determine if the trees are compatible. The somewhat confusing concept of *tree width* was only scantily treated earlier in the book, so the implications of these results are hard to interpret without better explanation.

Quartets are a special class of unrooted trees that have received a lot of attention. Besides whether a collection of trees is compatible, it is of interest to know whether a set of compatible trees uniquely (in some sense) defines a single tree of which all trees in the collection are a restriction. A natural concept pertaining to collections of trees—excess—is defined as the difference between the number of internal edges in the full tree on all vertices and the number of edges of the subtrees in the collection. Ideally you would want each subtree to give you one internal edge in the tree on all leaves. It is interesting to identify the differences with and similarities to the previous Semple and Steel book [9]. Even those topics present in both books are treated from a different perspective, and it is definitely worth reading both books.

The chapter continues with two pages on (rooted and unrooted) maximal agreement subtrees (MASTs) and discusses some conjectures on the how small a MAST can be and what their expectation is for randomly picked binary trees. This is followed by analogous questions for the *quartet metric*. There is a practically very relevant section on trees on subsets of a full set. Then random coverage is discussed: when the subsets of species for which you have trees are chosen randomly, when can you infer the total tree? The chapter ends with another interesting situation where you have a set of subtrees or quartets and want to reconstruct more than one tree, which could well be the case with horizontal transfer or recombination. The chapter is very compact, often with several new definitions given within ten lines.

Chapter 5. Phylogenies Based on Discrete Characters. Given *characters* (i.e., traits) on selected nodes, typically on the leaves of a tree, the first section discusses when these characters can have a *perfect phylogeny*. A perfect phylogeny is a tree explaining the character states for the observed taxa, allowing each trait to have evolved only once and no reversals to have occurred, i.e., there are $r - 1$ mutations for an r-state character. This is easy for *full binary characters*, but much harder for *multistate characters* or partial characters. For multistate characters, three alternative avenues are exolored: (i) reduce the characters to quartets and find their span (trees that can display them); (ii) construct their intersection graph and see if it has a "restricted chordal completion"; or (iii) create all binary splits and check if they are compatible. Approach (ii) implies a restriction on the tree width of the intersection graph.

Four interesting extensions of the basic problem of having full binary characters are highlighted: (i) *persistent perfect phylogeny* (PPP), where a single reversal is allowed in a phylogeny; (ii) *partial binary characters*, where the character states for some leaves are unknown for some characters; (iii) *perfect haplotyping problem* (PPH), an extension of PP to diploid individuals; (iv) *incomplete directed perfect phylogeny* (IDPP), a rooted tree must be found so that each particular state can evolve multiple times but no reversals are allowed, and the state is not known for all leaves.

There is a nice result presented showing an extremely clever way to reconstruct any tree from only four characters (each of which may, however, have a large number of states)! So why on earth do we sequence all these genomes? The answer is simple the moment you see the trick behind the surprising result. We find this section very important, since many biologists have good intuition about binary characters but not about more general characters such as for four nucleotides! They should read this section.

Some simple combinatorial questions are addressed next: how many binary trees can explain a given character perfectly? Conversely, given a tree, how many characters

can you put on the leaves without forcing homoplasy?

Parsimony is discussed for binary characters, and connections with Menger's theorem are drawn—namely, how many characters will give a particular parsimony score (i.e., the number of edges in the tree for which the character must mutate). Some fun connections are then made with edit operations on trees which are not so surprising once you have seen them.

A very interesting section follows on *ancestral state reconstruction*. Parsimony assignment to the root is a kind of voting system with special cases such as the star tree (majority rule), balanced tree, and caterpillar. Again there are some interesting results here. Majority rule is trivial. In a caterpillar tree the two outliers can determine the state at the root. In balanced trees an extreme unfairness in outcome can also be obtained. Further interesting results are presented in Chapter 9. Finding *maximum parsimony trees* on multiple characters instead of single characters is then discussed. Finally there is a very brief section on *supertrees* and *short encodings*.

Chapter 6. Continuous Phylogenies and Distance-Based Tree Reconstruction. The important concept of *distances* between the objects on the leaves is considered. A central question is whether it is possible to reconstruct a tree with *edge lengths* assigned to all edges, given only the distances between leaves. Key to this is the *four-point condition*: a metric has a tree representation if and only if it satisfies the four-point condition. On rooted trees the corresponding object is the *ultrametric*. The *Gromov–Farris transform* takes a tree metric and gives an ultrametric for any arbitrary rooting of the tree. *Symbolic ultrametrics* are also discussed, although it is not clear what they are for. An obvious but important observation is made: If there is a perfect tree, the distance between strings of characters will be a tree metric, but the reverse might not be true. This is obvious since distance is a major reduction in data so you can clearly arrange characters that conflict, but whose effect cancels out in the distance function.

There follows a slightly discursive section on distances on genomes that should have been extended more fully. *Neighbor joining* (NJ) and *balanced minimum evolution* (BME) are discussed as two *distance-based tree-reconstruction methods*. The generalization to reconstruction from *partial distances* is also covered.

There is a short section on *indexed pyramids* and *Kalmanson metrics*, a very strange generalization of the four-point condition and a way of representing such a distance function using a weighted sum of splits. The section on the *geometry of the space of tree metrics* was most exciting and way too short. Here trees are embedded in the positive orthant of a high-dimensional Euclidean space and a series of interesting questions are addressed such as the connectedness of the space and the distance between two trees. Steel should have added a few sentences like "[...]T-X is a compact connected simplicial complex of dimension $n-4$" to give the book some appeal to the average rubber boot owning botanist.

Next follows an excellent review of *phylogenetic diversity* (PD). Everyone interested in biodiversity should read this chapter. The first measure of PD given is the total branch length of the tree relating the species set. Assume you are allowed to keep only k of the original n species. How do we find the subset of size k with the largest PD? A simple greedy algorithm is presented, then some alternative questions are posed: (i) $\Delta\mathrm{PD}(Y)$: find Y where the PD loss is the largest if you extinguish Y; (ii) *maximum minimum distance* (MMD): find the set such that the longest shortest edge from the lost species to the surviving ones is minimized. (iii) a weighted version of PD; (iv) PD based on splits instead of trees. Optimization of PD and diversity indices for rooted trees are considered.

Some differences between the rooted and unrooted cases are discussed, and it is proven that PD is unimodular: the sum of PD on two sets is larger than the sum of PD on their union and their intersection. An interesting problem discussed is that of some nature reservations that cost something to maintain and which have a budget. Which reservations should you keep to maximize PD while staying within budget? There follows a subsection on decomposing the PD for a set to contributions from single species. *Fair proportion* (FP) and *Shapley*

value are discussed and it is then proven that the expected Shapley value is identical to FP for the complete tree.

The two extensions of PD (PD over Abelian groups and abstract diversity theory) should have been either extended or turned into footnotes, since it is hard to see their relevance. Section 9.2.3 is interesting. Extinctions that kill off species randomly and their effect on PD are considered. This leads to two special problems: *Heightened evolutionary distinctiveness* and the *Noah's Ark problem*. In the last section extinctions are applied to trees generated by the Yule process or the birth-death process.

Chapter 7. Evolution on a Tree: Part One. In Chapter 7, processes on a phylogeny are investigated. The chapter discusses different models for the evolution of character states along a phylogeny, how the models are related, how they can be combined, and the conditions under which the phylogeny can be recovered from the character distributions (i.e., from sufficiently many sampled characters). A good review of *Markov chains* is given and then the theory is extended to *Markov processes on a tree*. This starts out very general, with nonhomogeneous Markov chains on trees where each branch has its own transition probability. In some sense this still seems overly restrictive, since the internal nodes are arbitrary relative to the process so why let them define different evolutionary regimes? If the determinant of the transition matrix is nonzero, then a distance measure can be defined— *logDet*—that obeys the four-point condition that we know characterizes the tree, so a tree is identifiable from its leaf distributions.

There are a number of natural questions about irreversible rate matrices (Q) that are not discussed. Reversible Q are easy to characterize: give each state an equilibrium rate and each edge a rate of flow. Can one format an intuitive decomposition of the nonreversibility? There is a very interesting discussion of commuting and *noncommuting rate matrices*. Then we dive into the basic models starting with the *equal input model* that for nucleotides was introduced by Felsenstein in 1981, though similar models had been used for proteins by Margeret Dayhoff around 1970. Their purpose was to get the equilibrium distribution right. Nick Goldman and Bjarne Knudsen devised equal output models using which a weighted sum of equal input and equal output matrices can be made to create intermediates, which might have been worth mentioning.

There follow some very short sections. First, a theorem is presented saying that in the Jukes–Cantor model on r states, the *maximum likelihood* (ML) lengths are extreme: paths between different states are maximal (infinite) and those between identical states are minimal (zero). This is interesting, but it applies for a single state; for multiple states the ML estimates are hopefully more reasonable.

This is followed by a section on *G-equivariant and group based models*. G-equivariant is a new concept, in which the model doesn't change under group permutation of the state space. The most important new models in this class are models that do not change if you swap strand. "Group based" means the state space is a group and the Kimura three-parameter model is the most famous example. A very nice overview figure relating different models is given. A fun observation is that the most general model here is "general time reversible with variable Q," that is, not time reversible if the Q's don't commute. The section on *phylogenetic mixture models* is very important and it is well written. Here the identifiability of the underlying tree can be a problem, in contrast with models where all positions evolve by the same process. The chapter ends with the *Hadamard story* and *Felsenstein Zone*, which are covered a bit too briefly given how frequently they are discussed elsewhere.

Chapter 8. Evolution on a Tree: Part Two. This chapter continues the theory and application of Markov processes on trees. Statistical methods for *estimating phylogenies and ancestral character states* are discussed. This leads to questions about how much data is required to reconstruct a tree.

Some preliminaries on metrics on distributions are covered, followed by a list of attractive *properties of ML* estimation in general and a set of identifiability definitions and results. Now comes some fun stuff about *convergence* of logDet, ML, and max-

imum parsimony (MP). LogDet can have large variance for finite data and MP will converge with great speed toward the right result (faster than ML), provided we have parameters where it converges toward the right result! There follow some important sections on information-theoretic results on *how much data is needed* to infer the correct placement of a long branch and on the reconstruction of an ancestral state. Some of these results are very surprising at first; for example, the amount of data needed to infer a tree grows as the logarithm of the number of taxa.

The section on so-called "*phylogenetic oranges*" describes phylogenies as Euclidian flakes embedded in higher-dimensional space that are glued together in even lower-dimensional flakes (corresponding to multifurcating phylogenies), and these have certain exact mathematical properties. This is a fun bit of geometry, but its implications eluded us. Similarly, the phylogenetic invariants (algebraic analysis) are charming, but the intuition behind them is hard to grasp.

The *infinite cluster random model* is interesting and an extension of what has been discussed before under the equal input model. Going to infinite states gives you the Crow and Kimura 1964 infinite allele model. There are two extensions of this: the ladder model (investigated by Kingman among others) used to describe protein electrophoresis data, and the infinite site model by Kimura and Otha in 1971 used for DNA data. These models clearly needed to be incorporated into Steel's algebraic invariant dual homotopic centralizing framework!

Chapter 9. Evolution of Trees. The chapter returns to the topic of the evolution of trees themselves and their relationship with underlying evolutionary processes ("phylodynamics"). A section on the *pure birth process*—the simplest tree model—conveys much intuition and interesting paradoxes. The distribution of the number of descendants of a lineage after time t is geometric with parameter exp(lambda*t), where lambda is the speciation rate. This simple result could be explained by a simple argument, though the author derives it from a more general point process construction later in the chapter. The subtle issues of conditioning on elapsed time or final number of species are then discussed. Again some interesting paradoxes appear, arising from the properties of the simple Poisson process. For example, the distributions of a random internal edge and an external edge are the same, which is surprising since the latter will continue to grow in the future before becoming the former. The reconstruction of ancestral states is discussed for models with only two states and fixed mutation rates. (The extension to a large number of states is quite relevant for applications but unfortunately not discussed.) The probability that any *ancestral-state reconstruction* method could be correct for large trees depends on a new relevant parameter, namely, the ratio of speciation and substitution rates. In general, if the speciation rate is less than four times the mutation rate, then random guessing is as good as any method. Surprisingly, above this threshold majority rule is able to recover the ancestral state, while MP loses information about the ancestral state faster than majority rule, and even above its threshold of 6 it shows poorer performances.

The more interesting case of *birth-death processes* is discussed next. These models are widely applied since most real histories have both types of events. This leads to the interesting addition of the "reconstructed" tree or process, where only the lineages that have living descendants at time t are considered. Important effects like the "pull of the present" and the "push of the past" deriving from this conditioning are discussed in a few pages. More discussion would have been nice, since this is very relevant for phylodynamics and the behavior of these processes is not always intuitive. The next section introduces the *coalescent point process*. This is an important process since it provides a simple characterization of birth-death processes conditioned on time and extant lineages, even allowing for time and age dependence of the parameters. The construction is clear, but there are more subtle connections made that eluded us. The classical *Kingman coalescent* is discussed briefly, since it can be found in great detail in any book on population genetics. Finally, there is an interesting and important discussion on the loss of *PD* and the

difference between Kingman coalescent and birth-death models with respect to branch lengths.

The rest of the chapter is devoted to the relationship between gene trees and species trees. This includes some very interesting results on the *anomalous gene trees* of Rosenberg and Degnan, where the most likely gene tree disagrees with the topology of the species tree. This is a challenging result reminiscent of the Felsenstein Zone. As in other places in the book, it seems that most of the combinatorial nitty-gritty has been left out. The Degnan–Rosenberg anomalies raise the question as to whether methods can be developed that can give the right species trees. The author discusses these issues briefly, as well as other relevant issues for recent research like concatenation and ML on gene trees. Coalescent models to embed gene trees within species trees are then discussed. Lateral gene transfer is also considered, where the coalescent event is substituted by a species-jump and then a coalescent. These problems are extremely important, since you can observe genes but are generally keen to make statements about the species tree, which is extremely hard to observe directly. Note that most of the chapter focuses on incomplete lineage sorting, which dominates recent research but is not the only source of discordance between genes and species trees.

Chapter 10. Introduction to Phylogenetic Networks. The book ends with a chapter on phylogenetic networks. These phylogenetic structures include several deviations from tree-like inheritance and play an important role in prokaryotes due to the high rates of lateral gene transfer, but they also appear in different contexts (e.g., hybridization in eukaryotes). The topic is clearly worth studying. In the literature on networks there is a trend toward a high ratio of concepts to real biological use; this trend is apparent also in the large number of definitions presented in this chapter, even though the author makes it clear that they are needed because of the increased complexity of phylogenetic networks compared to trees. In this respect, this chapter is a very useful guide to such complexity. Most of the results apply to binary trees or networks.

Implicit (unrooted) networks are discussed first: networks with single cycles (unicyclic networks), galled networks (that can have cycles, but no node can be a member of more than one cycle), split networks including the widely used Neighbor-net method, and median networks built from sequences and related to MP.

The rest of the chapter is devoted to the more interesting case of *explicit (directed) networks*, which represent evolutionary histories more closely. The important difference between tree vertices (with a single ancestor) and reticulation vertices (with multiple ancestors) is discussed. For mathematical purposes, it is useful to consider subclasses of networks with bounded complexity of reticulation, which can be realised in different ways described here: level-k networks, tree-child networks, tree-sibling networks, and reticulation-visible networks. For example, in *tree-child networks*, reticulation is limited by the fact that every internal vertex has a child that is a tree vertex. Temporal networks can be defined by extra conditions on the order of splits/reticulations, which, however, make sense only if all species survived and were sampled. More classes of networks with nice mathematical properties are presented (networks without redundant arcs, normal networks, regular networks). The chapter discusses the relationships between these networks, as well as the larger class of *tree-based networks* (networks obtained adding links to a tree), whose characterization is less intuitive than might be expected.

Finally, the relationship between trees and networks is unveiled. Removing reticulations, how many trees can be displayed by a network? Can a specific tree be displayed? Conversely, is it possible to reconstruct a network from the trees displayed, or from subnetworks, distances, or characters? What is the network that minimizes reticulation from a set of trees? These questions find their answers here.

There are many good things in this chapter, but it still feels a pity to ignore the ancestral recombination graph (the structure that describes the relationship of a set of genomes from a population or viruses). However, much of what is explained is extremely close to the ancestral recombination graph, which is already discussed in

other books.

Comparison to Other Books. Although phylogenetics is an older field, it has for the last 50 years been strongly tied to sequence data and the field of molecular evolution. Prior to this era, Willi Hennig's basic outline of a theory of phylogenetic systematics [3] was the major work in the field and was motivating in its attempt to formulate rigorous principles, despite using no statistics or algorithmics, for phylogenetic inference. It was expanded in 1966 into *Phylogenetic Systematics* [4]. Sokal and Sneath's *Principles of Numerical Taxonomy* [10] was both more exact and statistical. Sheila Embleton's *Statistics in Historical Linguistics* [1] was clearly focused on language and distance methods but could also be applied to sequence data. The first textbook fully focused on molecular evolution and phylogenetics was the undergraduate text by Li and Graur [7], *Fundamentals of Molecular Evolution*, that later was expanded into Li's *Molecular Evolution* [6]. In 2004 Felsenstein published his large *Inferring Phylogenies* [2], which is highly readable and also covers the history of phylogenetics. The same year Steel and Semple published *Phylogenetics* [9], which is a very appealing read for mathematicians, statisticians, and computer scientists. The following year Ziheng Yang published *Computational Molecular Evolution* [11], which in 2013 was expanded into *Molecular Evolution: A Statistical Approach* [12].

Besides these, there is a series of books with a very hands-on approach, instructing readers on how to navigate existing programs, such as *Molecular Systematics*, [5] by Hillis, Moritz, and Mable and *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing* [8] by Salemi, Vandamme, and Lemey.

How should these books be prioritized by a researcher who wants to get into phylogenetics? Well, it depends on the interests and background of the researcher. If the researcher is a mathematician, statistician, or computer scientist, Felsenstein [2], Yang [12], Steel (2016), and possibly Semple and Steel [9] would provide a sound basis. If the researcher is a biologist, Felsenstein [2], Yang [12], and Salemi, Vandamme, and Lemey [8] would be good choices. Felsenstein and Yang are on both lists, since Felsenstein provides an excellent background and Yang is closer to data analysis, which is after all the motivation for phylogenetics. The reason for discarding Steel's books for the biologist is that they are simply too mathematical. However, it would be useful if the insights from Steel's books were diffused as much as possible into the biological community.

Summary. Steel covers most of what might be considered relevant, but there are some important omissions. Some of these could have been included with the addition of 40–60 pages, but some topics could not have been included without extensively altering the scope of the book. In the former category we find the bootstrap, statistical alignment, recombination, and phylogenetic regression. In the category of topics that would have seriously changed the scope of the book, we find the evolution of complex characters (structures, networks, shapes, phenotypes...), selection, annotation, and MCMC.

Does the book fail on some accounts, or contain biases due to Mike Steel being a mathematician? It is easy to ask for the impossible, listing a series of topics and computational experiments that would significantly enlarge the book and require an extra 6 months' work from Mike Steel and potentially some additional computational assistance. But it is a review's obligation to be critical and to encourage an excellent book to become even better. Since Steel does such an excellent job of extracting the essence of algorithms and mathematical results, it is a pity that certain topics have been ignored.

Getting into mathematical phylogenetics by reading this book would probably be far faster than tracking down the articles that Steel has digested for us. Thus, the topics left out are seriously disadvantaged. However, the book is already 60% longer than it was supposed to be as part of this book series.

A more detailed review of this book can be found at `http://preview.tinyurl.com/steel16bookreview`.

[1] S. M. Embleton, *Statistics in Historical Linguistics*, Brockmeyer, Bochum, Germany, 1986.

[2] J. Felsenstein, *Inferring Phylogenies*, Sinauer Associates, Sunderland, MA, 2004.

[3] W. Hennig, *Grundzüge einer Theorie der Phylogenetischen Systematik*, Berlin, 1950.

[4] W. Hennig, *Phylogenetic Systematics*, University of Illinois Press, Urbana, IL, 1966.

[5] D. M. Hillis, C. Moritz, and B. K. Mable, *Molecular Systematics*, Sinauer Associates, Sunderland, MA, 1996.

[6] W.-H. Li, *Molecular Evolution*, Sinauer Associates, Sunderland, MA, 1997.

[7] W.-H. Li and D. Graur, *Fundamentals of Molecular Evolution*, Sinauer Associates, Sunderland, MA, 1991.

[8] M. Salemi, A.-M. Vandamme, and P. Lemey, *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*, Cambridge University Press, Cambridge, UK, 2009.

[9] C. Semple and M. A. Steel, *Phylogenetics*, Oxford University Press, Oxford, 2003.

[10] R. R. Sokal and P. H. A. Sneath, *Principles of Numerical Taxonomy*, W. H. Freeman, San Francisco, 1963.

[11] Z. Yang, *Computational Molecular Evolution*, Oxford University Press, Oxford, 2006.

[12] Z. Yang, *Molecular Evolution: A Statistical Approach*, Oxford University Press, Oxford, 2014.

Mathias C. Cronjäger
David Emms
Luca Feretti
John Hein
Oxford University