Phylogenetics.—Charles Semple and Mike Steel. 2003. Oxford University Press, New York. xiii+239 pp. $90.00.

Phylogenetics is a field within which we have seen fruitful interplay between several different branches of science, including biology, mathematics and computer science. Unfolded in this book is a story about the mathematical development of phylogenetics, told by two mathematicians who have made many significant contributions to this growing field. The authors' versatility in words, as well as in mathematics, makes reading this book altogether an enjoyable experience for the mathematically-inclined. Clarity permeates through the entire length of the book, with myriad illustrative examples and figures. Furthermore, exercises are provided at the end of each chapter to reinforce the reader's understanding. For ease of reference, the main text is preceded by a short section where some notation and mathematical preliminaries are provided. Although discrete mathematics is the predominant focus of the book, algorithmic and probabilistic aspects are also frequently considered.

To the less mathematically-inclined reader, this book might seem too dense and unreadable at first sight. They should not be intimidated by it, however; the pre-requisite for reading this book is actually minimal. Once the reader is accustomed to the formal style of the book, he or she should not have much trouble reading through the book, perhaps skipping the proofs if desired. After a while, lemmas, propositions and theorems may appear convenient for summarizing main results. To the less biologically-inclined reader, the authors' well-phrased explanation of underlying biological motivations should prove helpful.

Perhaps it would be worthwhile to draw comparison, albeit only briefly, with *Inferring Phylogenies* by Joseph Felsenstein (2004), another excellent book on phylogenetics which became available shortly after the book under review did. Although the two books share several common grounds in terms of topic, they significantly differ not only in style but also in perspective and in depth. Furthermore, *Inferring Phylogenies* addresses many more topics than *Phylogenetics*. For example, Felsenstein gives a good overview of the coalescent process and recombination, two widely-studied areas which have played a major role in statistical genetics. In contrast, *Phylogenetics* contains only a partial description of the coalescent process, and recombination is not mentioned at all. Nevertheless, *Phylogenetics* contains details and special topics that are not included in *Inferring Phylogenies*. In our opinion, the two books complement each other well; it would be ideal if they could be read in conjunction.

Despite the fact that the book under review is very good and unique, there are some potential criticisms that the reader could voice depending on his or her perspective and background. We here offer some suggestions. Phylogenetics has risen to great importance as a consequence of the growth of exact quantitative data, especially at the sequence level, over the last five decades. A sketch of the historical development of phylogenetics would have been appreciated, and would have indicated that this field is still growing and is in no way finished. The importance of phylogenetics is also due to the inherent tree-ness of life and to the availability of characters in species that reflect this tree-ness. Again, some discussion on this subject and the great success of tree reconstruction—its major influence in all of biology and in present projects of constructing the tree of life—would have been motivating for any reader of this book.

Any criticism regarding omitted topics could be a bit unjust, for one of the appealing qualities of this book is that it is concise, and any attempt to take such a criticism into account would inevitably lengthen the book. With that consideration in mind, we point out that some major areas of active research have been omitted from discussion in this book. For example, sequence alignment is hardly mentioned, although one of the authors of this book has done some groundbreaking work that has given new momentum to the field of statistical alignment on phylogenetic trees. In general, alignment of sequences is an important problem and often is a precursor to phylogenetic analysis. Recent developments have extended stochastic models of character evolution to consider the alignment problem by also including a stochastic description of insertion and deletion processes.

Although the authors' choice of topics closely reflects their interest, the book is in no way a mere eclectic collection of their previous publications. The book consists of eight very well-organized chapters that are skillfully sewn together to make up one coherent picture. Logical flow of the book is well thought out, with some chapters being smoothly interconnected. Relevant introductory materials are provided when necessary, and each chapter except for Chapter 1 ends with at least one specialist topic. Commonly used symbols are listed near the end of the book, which we found very useful at times. In what follows, we briefly summarize what each chapter contains.

An indispensable tool in phylogenetics is graph theory, and indeed a common theme which runs through all the chapters in the book is graph theory, in one form or another. Numerous useful terminologies and facts from graph theory are laid out in Chapter 1, appropriately setting the scene for the coming of fully-developed theories in later chapters. The so-called intersection graphs, which play an important role in later chapters in particular and in combinatorial biology in general, are also discussed in Chapter 1. The chapter closes with a brief look into three concrete biological applications, aside from phylogenetics, in which graph theory has proved useful. These examples well illustrate how mathematical concepts such as intersection graphs naturally arise in biological problems.

Chapter 2 introduces a class of fundamental objects called X-trees, which appear time and again throughout the book. Here, X denotes a finite set that labels certain vertices, and phylogenetic trees, which are of main interest to biologists, are a particular kind of X-trees. The main reason for considering such a general class of objects is that, as fully described later in the book, there exists a natural equivalence between X-trees and a certain kind of set system related to X. This equivalence has played a

key role in the mathematical study of phylogenetics.

Various types of X-trees—including rooted X-trees and ranked phylogenetic trees, the latter being commonly used in population genetics to represent coalescent processes—are defined and several relevant enumeration results are provided. Related to simple models of speciation and extinction in biology, two widely-studied models for randomly generating rooted binary phylogenetic trees are discussed. Three types of tree rearrangement operations on binary phylogenetic trees, which can be used to transform one tree to another, are also discussed in Chapter 2. These operations have many applications in phylogenetics, including finding maximum parsimony trees and reconstructing phylogenetic trees, to name a couple.

In Chapter 3, the aforementioned equivalence between X-trees and a certain type of set system—more precisely, a collection of pairwise compatible X-splits—is formalized as the "Splits-Equivalence Theorem" (Buneman, 1971). An X-split is a partition of the set X into two non-empty subsets, and two X-splits are said to be compatible if a certain condition is satisfied. As the authors elaborate through various topics in this chapter, the notion of X-splits, as well as that of their compatibility, is central to the mathematical study of phylogenetics. For instance, several well-known methods of finding consensus trees, useful for representing common characteristics of a collection $\mathcal{P}$ of X-trees, are based on determining X-splits shared by all or some X-trees in $\mathcal{P}$. Also described in Chapter 3 is an analogue of the Splits-Equivalence Theorem for rooted X-trees. Furthermore, the so-called Buneman graph, associated to a set of X-splits not necessarily pairwise compatible, is discussed in the specialist topic section.

The main topic of Chapter 4 is about characters, which can be roughly described as certain attributes—for example, morphological or genetic data—associated to species. More precisely, characters are functions from a non-empty subset of X into a set of states. The primary importance of characters lies in that they are what allow us to reconstruct phylogenetics trees. Convexity and compatibility of characters—the former being closely related to homoplasy-freeness in biology—are examined in this chapter using graph theoretical techniques. Biologically, a collection of characters is said to be compatible if all characters in the collection could have evolved on the same X-tree without encountering homoplasy. There exists a tight link between X-splits and a special type of character called "binary," and this link is highlighted in the chapter. In addition, this chapter addresses several important issues such as studying the algorithmic complexity of determining compatibility, dealing with the case in which a collection of characters is not entirely compatible, generalizing characters to account for ambiguity in data, and determining when there exists a unique X-tree associated to a compatible collection of characters.

Chapter 5 focuses on the widely-used method of maximum parsimony for inferring phylogenetic trees from characters. The main idea underlying this method is to choose the simplest possible explanation when a given collection of characters is not entirely compatible. This approach can be divided into two parts, namely to determine the parsimony score of a sequence of characters for a given tree and to find a maximum parsimony tree. Both of these topics are clearly addressed in the chapter. In its classical formulation, maximum parsimony concerns minimizing the number of changes in character states. Given a phylogenetic tree, computing the minimum number of changes and finding the most parsimonious explanation can be done using an efficient algorithm called the Fitch-Hartigan algorithm. This classical problem, as well as several interesting extensions of it, is described here in the authors' usual clear style.

What brings special light to this well-known topic, in our opinion, is the authors' ability to elaborate on insightful connections between the maximum parsimony problem and discrete mathematics, sometimes in the context of probabilistic questions. How combinatorial and graph theoretical analyses can be useful for phylogenetics is well demonstrated here. For example, the authors describe the translation of the classical parsimony problem into the Steiner problem for graphs and discuss how it can be used in connection with the Buneman graph, introduced in Chapter 3, to show that every maximum parsimony tree for a collection of distinct binary characters is displayed in the associated Buneman graph.

Chapter 6 concerns various questions that stem from combining the information contained in different trees to construct a single tree. Algorithmic aspects are emphasized in various places in this chapter. Two principal tasks of growing importance are to determine the compatibility of an arbitrary collection of semi-labelled trees, either all rooted or all unrooted; and, when a given collection of trees is not compatible, to construct a tree, called supertree, which best summarizes the information contained in the input trees. Both unrooted and rooted cases are examined in detail, and the contrast between the two cases highlighted in the chapter is very illuminating. The rooted case is shown to be more tractable for two reasons. First of all, for a set of unrooted phylogenetic trees, determining compatibility is NP-complete, whereas for rooted phylogenetic trees, there exists a polynomial-time reconstruction algorithm called BUILD (Aho et al., 1981), variation on whose theme leads to other useful applications. Secondly, there exists a supertree method for rooted phylogenetic trees which satisfies a certain set of desired properties, but the same does not hold for unrooted trees.

At the heart of tree reconstruction is the notion of a tree to be "identified" or "defined" by its subtrees. In the case of unrooted phylogenetic trees, subtrees which serve as elementary building blocks are quartet trees, defined as binary phylogenetic trees with exactly four leaves. The authors have appropriately chosen important questions for discussion, including finding the number of required quartet trees, using quartet trees for tree reconstruction, and determining which subsets of quartet trees associated to a tree define the tree. The problem of determining whether a collection of subtrees is definitive is also considered.

Chapter 7 marks an important place in the book, as

it is where edge lengths are considered for the first time. Given a phylogenetic tree with positive edge weights, a natural distance function can be defined on the set of species. Namely, the distance between any pair of species x and y can be defined as the sum of the edge weights in the path joining x and y. An important problem relevant to tree reconstruction is studying the converse of the above statement. That is, given an arbitrary dissimilarity map on X—determined from input data, for instance—it is of interest to know whether such a map can be represented by an X-tree with positive edge weights and, if so, to construct such a representation. A vast amount of mathematical work has been done on this topic, and it certainly helps to have important results gathered in one place, as the authors have done in this chapter.

In addition to the theoretical works related to determining when a dissimilarity map is a tree metric, several tree reconstruction methods, including the popular neighbour-joining method, are described in the chapter. These methods can be applied to arbitrary dissimilarity maps. The authors also provide a readable account of split decomposition theory, which has been implemented in *SplitsTree*, a computer software widely used by biologists (Dress et al., 1996). Furthermore, there are two specialist topics covered in this chapter, both of which have much biological, as well as mathematical, significance.

Chapter 8, the final chapter of the book, is about stochastic processes on trees describing the evolution of characters, a topic central to phylogenetics. Given a sequence of characters generated by such a stochastic process on a particular phylogenetic tree, how well various reconstruction methods can recover the underlying tree is an important question, in regard to both testing existing methods and devising new ones. Markov processes on phylogenetic trees—and, more generally, on graphs—are introduced, and probability distribution of characters, as well as that of dissimilarity, are examined for various models. Under certain assumptions, such distributions are shown to determine the underlying tree, up to isomorphism and the position of the root.

Although some important topics such as maximum likelihood and Bayesian methods are only very briefly mentioned, this chapter still carries much weight, for it contains, along with a clear exposition of the results described above, an accessible introduction to numerous advanced topics. Of particular usefulness is the authors' lucid and confluent description of the Hadamard transformation, rate variation across characters, and group-based models. Their clear explanation of the Felsenstein zone and phylogenetic invariants should also be useful to non-specialists.

The authors state in the preface that their intention is to provide "a reasonably self-contained overview of an expanding field." In our opinion, they certainly succeed in meeting that goal. The list of references compiled in the book by itself seems quite useful. All in all, this book should serve as an excellent mathematical introduction to phylogenetics for beginners and as a good reference for experts in the field.

We wish to point out, however, that this book is an outlier among books on phylogenetics, in the sense that it is not intended for the same type of audience as that for which most books and review articles—for instance, Felsenstein's book—on the subject are directed. This book is much more mathematical than other available books, and hence is perfect for discrete mathematicians who want to learn about the theory behind phylogenetics. There is still a room for a more intuitive book explaining and illustrating key concepts of phylogenetics for non-mathophobic biologists, in a similar vein as *Proofs from THE BOOK* (Aigner and Ziegler, 1998) and *The Book of Numbers* (Conway and Guy, 1996) have done for the fields of combinatorics and elementary number theory. The best qualified to write such a book in phylogenetics would also be Semple and Steel. Much of the combinatorics behind phylogenetics is in principle accessible to all, but the book under review will mainly be enjoyed by researchers with strong quantitative backgrounds and will be tough going for typical frog taxonomists.

In closing, we mention that updated errata for the book can be found in the authors' personal web-pages, which can be accessed through their department's webpage (http://www.math.canterbury.ac.nz/).

References

Aho, A.V., Sagiv, Y., Szymanski, T.G, and Ullman, J.D. 1981. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. SIAM Journal on Computing 10:405–421.

Aigner, M. and Ziegler, G. 1998. Proofs from THE BOOK. Springer-Verlag, Heidelberg.

Buneman, P. 1971. The recovery of trees from measures of dissimilarity. Pages. 387–395. *in* Mathematics in the Archaeological and Historical Sciences. (F.R. Hudson, D.G. Kendall, and P. Tautu, eds.) Edinburgh University Press, Edinburgh.

Conway, J.H. and Guy, R.K. 1996. The Book of Numbers. Copernicus, New York.

Dress, A.W.M., Huson, D., and Moulton, V. 1996. Analyzing and visualizing sequence and distance data using SplitsTree. Discrete Applied Mathematics, 71:95–109.

Felsenstein, J. 2004. Inferring Phylogenies. Sinauer, Sunderland, Massachusetts.

Yun S. Song and Jotun Hein. Department of Statistics, University of Oxford, Oxford, OX1 3TG, United Kingdom.