Systematic Biology Advance Access published October 8, 2012

Book Review

Syst. Biol. 0(0):1–3, 2012

DOI:10.1093/sysbio/sys082

Basic Phylogenetic Combinatorics.— Andreas Dress, Katharina T. Huber, Jacobus Koolen, Vincent Moulton, and Andreas Spillner. 2012. Cambridge University Press, Cambridge. xii+264 pp. ISBN 978-0-521-76832-0, £40.00, $65.00 (hardback).

Genomic data have become more and more important in biology, and methods that analyze these data are often based on mathematical ideas. Indeed, in recent years mathematics has become increasingly valuable for biology. The book *Basic Phylogenetic Combinatorics*, however, clearly shows that the converse is also the case. Evolutionary biology has inspired an exciting new field of mathematics that comprises many beautiful and challenging mathematical questions which have important real-life applications.

This book has not been written for biologists. After a short, clear preface in words, the main part of the book is full of mathematical notation, formulas, and proofs. Even for mathematicians, the book might not be as easy to read as the first word of the title suggests. Moreover, the book might not be of direct interest to biologists because it does not describe many methods or algorithms that can be immediately applied to biological data, nor does it describe many applications or refer to papers where such applications can be found. For this, you can consult the book of Felsenstein (2004).

This does not mean that *Basic Phylogenetic Combinatorics* has no relevance for biology. On the contrary, the book describes the mathematical theory that forms the foundation of computational phylogenetics. The theory described in this book has been the basis for numerous methods and algorithms that can be used by biologists. A good example is given at the very end of the book, where the QNet method is described. This method can be used to construct a phylogenetic "split network" from a collection of quartet trees. Such a network is a more informative representation of phylogenetic data than a tree because "reticulate" regions are used to visualize parts of the data that are not tree-like; and this is illustrated by an application to some *Salmonella* data.

Such split networks are used more and more often by biologists to display evolutionary data for which no well-supported phylogenetic tree exists. To give just one example from the literature, Silver et al. (2011) used split networks generated by the QNet method to study the evolutionary history of the bacteria *Aeromonas veronii*, which can be a pathogen for humans but can also be a symbiont for medicinal leeches. They used the constructed networks to identify which groups of strains evolved down a well-resolved tree and in which groups horizontal gene transfer is likely to have occurred. This was used to show that horizontal gene transfer can occur at high frequency even for strains that are adapted to distinct niches.

For mathematicians, especially those with an interest in combinatorics, this book is an excellent way to learn all about phylogenetic combinatorics, which can roughly be described as discrete mathematics related to phylogenetic trees and other discrete mathematical objects related to phylogenetics. In mathematical terms, the field concerns leaf-labeled trees and, more generally, leaf-labeled graphs. These structures are sufficiently general to have more applications than the ones in evolutionary biology, including applications to other evolutionary processes (e.g., language evolution) and other types of data analysis (e.g., voting patterns, word clouds). Nevertheless, the main application of phylogenetic combinatorics is, and might always be, evolutionary biology.

The book is almost completely self-contained. The first chapter describes the basic mathematics that is being used in the rest of the book. Then the phylogenetic combinatorics begins, starting with standard well-known theorems, and continuing all the way to some of the latest results in the field. The book does not just give an overview of results but provides complete and often elegant proofs of all of the theorems and lemmas. References are provided to give the reader the chance to read more about a specific subject, but it is never necessary to read these references to understand the messages in the book.

The book concerns 3 "encodings" of unrooted phylogenetic trees, namely, splits, quartets, and metrics. To explain what this means, let us first consider splits. A split is a division of the taxa (the labels of the leaves) into 2 groups. Each branch of a phylogenetic tree describes one split. That is, it divides the taxa into those that are on one side of the branch and those that are on the other side of the branch. Now, suppose that you are given all of the splits described by the branches of some phylogenetic tree. Then, it is possible to uniquely reconstruct this phylogenetic tree from those splits. This is what is meant by saying that splits are an encoding of phylogenetic trees. A quartet of a phylogenetic tree is the restriction of the tree to 4 of its taxa. Given all of the quartets of some phylogenetic tree, it is again possible to uniquely reconstruct the tree. Hence, also quartets form an encoding of phylogenetic trees. Finally, metrics (i.e. distances) are a third encoding of phylogenetic trees. If one knows the pathlength distance between each pair of

taxa in some phylogenetic tree, then one can uniquely reconstruct that phylogenetic tree from those distances. Chapter 2 of the book shows proofs of the (well-known) facts that each of these encodings gives rise to a unique phylogenetic tree.

None of this is directly usable in practice because a real biological data set hardly ever corresponds exactly to a tree. However, more interesting than the uniqueness itself is the fact that there are clean mathematical conditions that can be used to decide if a given collection of splits, quartets, or distances corresponds to a tree. Chapter 3 of the book describes these conditions and proves their sufficiency and necessity. These conditions have been well studied and used in several phylogenetic-tree reconstruction methods. Indeed, the 3 encodings correspond to 3 types of input data for which phylogenetic reconstruction methods exist: quartets are used in quartet-based methods (e.g., quartet puzzling), metrics in distance-based methods (e.g., neighbor joining), and splits are used to construct so-called split networks. It should be noted that the most accurate currently known methods for constructing phylogenetic trees (maximum likelihood and Bayesian methods) work directly with genetic sequences, and such sequence-based methods are not discussed at all in this book. Nevertheless, distance-based methods are often used to find a first solution which is subsequently improved (in a local-search approach) using sequence-based methods. In addition, "supertree" methods (related to quartets) will probably always be important for combining smaller phylogenies into larger ones, because the total number of species and strains is so large that it is unlikely that any method will ever be able to analyze them all simultaneously. Hence, quartet- and distance-based methods are still relevant, and so are split-based methods.

Chapters 4, 5, and 6 of the book describe how, given the splits, quartets, or distances (respectively) of some phylogenetic tree, this tree can be reconstructed. As mentioned above, this will hardly ever be possible in practice because a given data set will usually not be perfectly consistent with any phylogenetic tree. Reasons for this include possible errors in the data or the fact that the real evolutionary history might not be tree-like due to hybridization, recombination, or horizontal gene transfer events. For these situations, the book describes possible ways of constructing a network instead of a tree. Such networks are often called "data-display" networks because they are not based on a concrete model of evolution but are used to display the inconsistencies in the data, that is, the networks indicate where and how much the data deviate from being tree-like.

A topic not described by the book, apart from a short note at the end of Chapter 9, is what are called "evolutionary" (or "explicit") phylogenetic networks, that is, networks that explicitly describe a hypothetical evolutionary history. Such networks differ from data-display networks in the sense that evolutionary networks usually have a root, directions on their branches, and special nodes called "reticulations" which are used to model hybridizations, recombinations, and/or horizontal gene transfers. It is important to note that the data-display networks discussed in this book have shown to be a valuable tool for visualizing conflicts in data but do not aim to visualize explicit evolutionary scenarios. Evolutionary networks have not yet been studied as much as data-display networks, and directed graphs are in general not as well understood as are undirected graphs, meaning that more research in this direction is still necessary.

The last 4 chapters of the book deal with the most specialized topics. Chapters 7 and 8 show how the different encodings are related. For example, it is shown how one can directly obtain the quartets from a phylogenetic tree from the splits of that phylogenetic tree, without constructing the tree itself. The same is done for each pair of encodings, in both directions. These chapters are very abstract and mathematically elegant but are far away from possible applications. The last 2 chapters are slightly more concrete.

Chapter 9 discusses rooted trees. It is explained that each of the 3 encodings of unrooted phylogenetic trees has a rooted variant: clusters instead of splits, rooted triplets instead of quartets, and ultrametrics instead of metrics. A technique called the "Farris transform" is used to show that several results for unrooted trees also hold for the rooted variants. This is related to the rooted, evolutionary phylogenetic networks mentioned above.

Finally, Chapter 10 discusses some more ways of handling data that are not consistent with any phylogenetic tree. Generally, the 2 possibilities are to either construct a phylogenetic tree that is consistent with a fraction of the data (e.g., consensus trees) or to construct a phylogenetic network (e.g., by the QNet method).

In a certain sense, the book can be seen as a follow-up to the book by Semple and Steel (2003). Both books describe the mathematical foundations of phylogenetics but do not focus on computational aspects and applications. The book by Semple and Steel describes a somewhat wider range of topics, including (multistate) characters, maximum parsimony, probability, and Markov models, while Dress et al. concentrate on combinatorial aspects and give a deeper, completer, and more recent exposition of these aspects, including several previously unpublished results.

To summarize, *Basic Phylogenetic Combinatorics* does not provide an overview of all topics in mathematical phylogenetics but focuses on the discrete, abstract, and elegant subfield of phylogenetic combinatorics. This subfield is described from the basics up to advanced results in a concise and self-contained way, aiming at a mathematical audience. The book describes theory that has been, and in the future will more often be, used for designing phylogenetics-related methods and software. Moreover, the book shows how biology has inspired a whole new branch of exciting and beautiful mathematics that is interesting not only for its application in biology.

Anyone with an interest in discrete, combinatorial mathematics will enjoy reading this book.

References

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.

Semple C., Steel M. 2003. Phylogenetics. Oxford: Oxford University Press.

Silver A.C., Williams D., Faucher J., Horneman A.J., Gogarten J.P., Graf J. 2011. Complex evolutionary history of the *Aeromonas veronii* group revealed by host interaction and DNA sequence data. PLoS One 6: e16751.

Leo van Iersel, Centrum Wiskunde & Informatica, Science Park 123, 1098 XG Amsterdam, Netherlands; E-mail: l.j.j.v.iersel@gmail.com