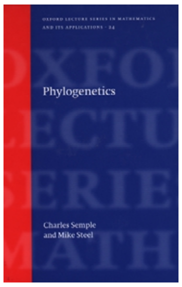
Phylogenetics as applied mathematics

Phylogenetics by Charles Semple and Mike Steel. Oxford University Press, 2003. £45.00 hbk (254 pages) ISBN 0 19 850942 1

Ziheng Yang

Department of Biology, University College London, Darwin Building, Gower Street, London, UK WC1E 6BT



Theoretical developments in phylogenetics are published in a variety of subject journals, including biology, applied mathematics, combinatorics and computer science. If you rely mainly on the biology literature for recent methodological advancements, as most readers of *TREE* probably do, what have you missed?

Well, *Phylogenetics* gives you an excellent opportunity to catch up. In eight chapters, this book provides a concise and lucid summary of the mathematics literature related to phylogenetics. The first three chapters introduce the basics of graph theory and definitions of trees, and describe tree rearrangement algorithms (e.g. nearest neighbor interchange, subtree pruning and regrafting) and Robinson and Fould's partition distance between trees. Compatibility and parsimony methods are discussed in Chapters 4 and 5, respectively. Chapter 6 is devoted to super-tree construction, the problem of combining subtrees for overlapping species into a super-tree. Chapter 7 describes tree reconstruction from a matrix of pairwise distances. The final chapter introduces Markov-process models of character evolution and mentions Hadamard representation, likelihood and Bayes analyses.

My only quibble with *Phylogenetics* is that it is mathematical but not statistical. Although the book proves many mathematical theorems concerning phylogenies, the analysis is deterministic and ignores the stochastic nature of the data or the sampling errors in estimates. Phylogenetic methods seem to be chosen for discussion because they are tractable analytically, rather than because they are useful for answering important biological questions. For example, compatibility methods are now rarely used in molecular phylogenetics, as are those based on invariants. Super-tree construction can be useful, but probably not until the uncertainties in estimated subtrees are accommodated. Models and methods that are well known to molecular systematists, associated with names such as Kimura, Hasegawa, Kishino, and so on, are not mentioned at all.

Nowadays, Cavalli-Sforza and Edwards's view that phylogeny reconstruction should best be considered a statistical estimation problem [1] is well accepted. In the past two decades, phylogenetic methodologies have been driven by the explosive accumulation of molecular sequence data. To me, the most exciting theoretical advances are those probabilistic models and statistical methods for analyzing molecular data to address fundamental biological questions (e.g. [2–5]). For example, given a sequence alignment, what is the best estimate of the underlying phylogeny? How reliable is the estimate or can we rule out alternative relationships? How can we estimate dates of species divergences? How should we combine sequence data from different gene loci to obtain reliable estimates of phylogenies and biological parameters? What can we learn about the evolutionary process of the gene; that is, is it dominated by mutation or selection? The book comes close to discussing some of those methods in the final chapter but ends right when it is getting interesting.

Phylogenetics is written at the graduate level and is intended for biologists as well as for mathematicians, statisticians and computer scientists. It must be a daunting task to accommodate such disparate groups of readers, and this book is more to a mathematician's liking. The exposition is formal even though the required mathematics is basic. Here, the authors could have helped the biologist reader by including a verbal rendering of important definitions and results. However, if you overcome the mathematical formalism and understand what the theorems are about, you will find many new ideas scattered among the pages. Programmers, in particular, will gain insights from reading this book; for example, into tree representations and tree-rearrangement algorithms.

Concisely and clearly written, *Phylogenetics* is a must-read for mathematicians or computer scientists who wish to do research in molecular phylogenetics, computational biology and bioinformatics. I hope the book will attract powerful mathematicians into this exciting area of research. It would be even nicer if they are persuaded or forced to focus on questions that biologists hope to answer with their data. Modern methods of molecular phylogenetics, such as likelihood and Bayes Markov chain Monte Carlo methods, involve sophisticated probabilistic modeling and pose huge statistical and computational challenges, especially with the increasing need to perform integrated analysis of multiple heterogeneous data sets. Those problems should provide ample opportunities to put mathematicians' prowess to good use. It might not be feasible to prove many theorems, because most of the problems are not tractable analytically, but the reward

Corresponding author: Ziheng Yang (z.yang@ucl.ac.uk).

is great when numerical methods help to answer fundamental biological questions.

References

1 Cavalli-Sforza, L.L. and Edwards, A.W.F. (1967) Phylogenetic analysis: models and estimation procedures. *Evolution* 21, 550–570

2 Swofford, D.L. *et al.* (1996) Phylogeny inference. In *Molecular Systematics* (Hillis, D.M. *et al.*, eds), pp. 411–501, Sinauer Associates

3 Yang, Z. and Bielawski, J.P. (2000) Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* 15, 496–503

4 Huelsenbeck, J.P. *et al.* (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294, 2310–2314

5 Whelan, S. *et al.* (2001) Molecular phylogenetics: state of the art methods for looking into the past. *Trends Genet.* 17, 262–272

0169-5347/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/S0169-5347(03)00194-0

Vital statistics

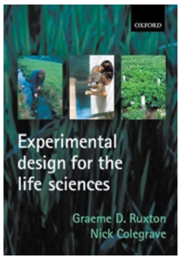
Experimental Design for the Life Sciences by G.D. Ruxton and N. Colegrave. Oxford University Press, 2003. £14.99 pbk (132 pages) ISBN 0 19 925232 7.

Modern Statistics for the Life Sciences by A. Grafen and R. Hails. Oxford University Press, 2002. £22.99 pbk (384 pages) ISBN 0 19 925231 9.

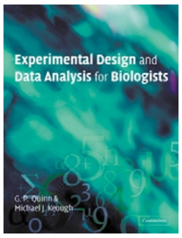
Experimental Design and Data Analysis for Biologists by G.P. Quinn and M.J. Keough. Cambridge University Press, 2002. £75.00 hbk (556 pages) ISBN 0 521 00976 6

Innes C. Cuthill

Centre for Behavioural Biology, School of Biological Sciences, University of Bristol, Bristol BS8 1UG, UK



OXFORD
Modern Statistics for the Life Sciences
Alan Grafen and Rosie Hails


Experimental Design and Data Analysis for Biologists
G. P. Quinn & Michael J. Keough
CAMBRIDGE

In February, I attended a workshop on the use and abuse of statistics in biomedical research. Sponsored by the Medical Research Council, it had representatives from all the main UK bioscience funders and learned societies. The conclusion of the meeting was clear: there is a crisis in the teaching of statistics in the biosciences. Most biologists, psychologists and medics know a very limited range of experimental designs and analyses; they frequently misapply their knowledge and are too scared, lazy or arrogant to talk to statisticians who might help them. Can these three books start to put matters right?

Ruxton and Colegrave's book is aimed primarily at the undergraduate biology market, and mathematically naïve undergraduates at that. I would doubt that there is anything in the book that would surprise a professional biologist, but this would be a good book with which to coax mathophobic students into good habits. The book is equation free and takes a very, very gentle approach. The philosophy is that 99% of biological research can be accomplished with relatively few experimental designs and that you do not need to understand any statistical theory to appreciate why things such as replication, independence of subjects and random allocation are good ideas. The book goes as far as factorial, blocked and split-plot designs but stops short of anything more sophisticated. It also stops short of telling you how to analyze the data that you have collected from your elegantly designed experiment. I can understand why: a desire to keep the book short and to concentrate on the principles rather than getting bogged down with explanations of how to interpret the voluminous output of most stats packages. However, I would urge Ruxton and Colegrave to reconsider this for the second edition. Statistics and experimental design are two sides of the same coin and you cannot really teach one without the other. To explain how to analyze the designs covered in the book, say in SPSS or Minitab, would not add very much text and would give one a very nice single text for teaching an introductory course. After all, modern stats packages enable you to tackle quite sophisticated statistics without pain and with only passing acquaintance of the theory behind the tests. This of course has its downside. Not only can assumptions be unknowingly violated but, through failing to get advice from a statistician, one could also miss out on alternative, more powerful designs and analyses. However, to get and implement effective advice from a statistician, one probably needs to know rather more than Ruxton and Colegrave's book provides, and more of the jargon that their book (for their market, sensibly) avoids. This is where Grafen and Hails, and Quinn and Keough, come in.

Grafen and Hails' book is unlike any other basic statistics text for biologists. It does not say much about t-tests and chi-squared tests, or anything about non-parametric tests; it teaches us instead about the general linear model. Sounds complicated, so why should a biologist take the plunge? All the common parametric tests can be expressed as general linear models; that is, a sum of linear effects (whose magnitude and significance we are interested in testing) and normally distributed 'error' or residual variation. Regression is the most familiar 'GLM', where the dependent variable is modelled as linear

Corresponding author: Innes C. Cuthill (i.cuthill@bristol.ac.uk).