Evolution, 60(2), 2006, pp. 421–423

STATISTICAL METHODS IN (MOLECULAR) EVOLUTION[1]

PETER BEERLI
School of Computational Science and Department of Biological Sciences
Florida State University, Tallahassee, Florida 32306-4120
E-mail: beerli@scs.fsu.edu

Received January 3, 2006.

Statistical Methods in Molecular Evolution, edited by Rasmus Nielsen, contains a wide survey of current research in molecular evolution. It is organized into sections—introduction, program ''tutorials,'' models, and inference—a setup that constitutes a gentle introduction to the topic for mathematically inclined readers. For practical biologists, the start might be somewhat more challenging, although the introduction is tailored to a mixed audience. All chapters are written by researchers with active research projects in the areas they write about. I address each chapter with a short comment.

Introduction.—The introductory material sets the stage for all further chapters. Without going into too much depth, the authors give a broad overview of topics such as Markov chain-based substitution models, likelihood concept, Markov chain Monte Carlo (MCMC) methods, and population genetic aspects of molecular evolution. (1) ''Markov Models in Evolution:'' Galtier, Gascuel, and Jean-Marie give a crash course on Markov models that will leave mathematicians happily humming along and many biologists struggling with the mathematical syntax. The discussion of population models of DNA, RNA, and protein sequence evolution is concise but lacks the presentation of the transition probabilities for some of the models. Readers who want to familiarize themselves with these models still need to read Felsenstein (2004) and Swofford et al. (1996). (2) ''Introduction to Applications of the Likelihood Function in Molecular Evolution:'' Buschbom and von Haeseler give an overview of the likelihood principle. Several examples of application of the maximum-likelihood principle—from simple one-parameter inferences to complicated many-parameter problems, such as finding the best tree given a set of sampled sequences—are given. The difficulties inherent in likelihood ratio testing receive too little attention. It would have been useful to read about difficulties with testing of hypotheses, taking into account boundary conditions of the parameters. For example, how should one test if a branch length in a phylogenetic tree is zero? And should this be used as a means of judging support for the tree? Given that this book will have a much higher profile than a single paper, coverage of such topics would have been helpful to many readers. (3) ''Introduction to Markov Chain Monte Carlo Methods in Molecular Evolution:'' Larget gives a brief introduction to MCMC sampling, using a Bayesian approach exclusively. The Gibbs sampler, a special case of the Metropolis-Hastings (MH) sampler, is explained in detail. Regarding phylogenetics and population genetics, the Gibbs sampler seems to be of less importance than the MH sampler. Important components of these samplers are the changes from one parameter value to the next; this is illustrated by changing a topology of a phylogenetic tree. The reversible jump MCMC (Green 1995), probably the most important improvement of the MH algorithm, is discussed with an example of sampling from two different substitution models. Larget ends the chapter by discussing the problems of convergence and how much confidence we should have in the results of a specific outcome of an MCMC run. Many convergence diagnostics are available, but Larget's ''crude diagnostic'' (p. 58), in my opinion, is superior to trace plots or many other statistics because the crude approach forces the user of the methods to understand the differences between the results. (4) ''Population Genetics of Molecular Evolution:'' Bustamante introduces the reader to population genetics concepts. At first, I thought that such a chapter was misplaced in the introduction to molecular evolution. However, Bustamante skillfully combines the advances in selection theory and its null model, the neutral theory, in population genetics with phylogenetic inference of selection pressure on amino acid changes. This chapter makes it evident that molecular evolution is not simply a new wrinkle in the study of evolution, but that it is the exploration of evolution with modern tools and is not restricted to tree-based thinking.

Practical approaches for data analysis.—This section of the book is probably the one that will age the fastest, because all chapters discuss specific computer programs and their capability to estimate parameters. All authors are still working on these programs, and so these programs will continue to evolve. (5) ''Maximum Likelihood Methods for Detecting Adaptive Protein Evolution'' Bielawski and Yang give a guided tour of estimation of selection pressure on phylogenies measured by ω, the ratio between rates of nonsynonymous and synonymous substitutions. The examples are helpful for readers who want to learn about PAML (phylogenetic analyses by maximum likelihood), but the discussion also touches on general issues with relative estimates of selection pressure among different phylogenetic lineages. (6) ''HyPhy: Hypothesis Testing Using Phylogenies'' Kosakovsky Pond and Muse deliver a tutorial to their program HyPhy. HyPhy is a very versatile phylogenetic framework that permits many tests and comparisons not available in other computer programs. This tutorial will help potential users of the program. (7) ''Bayesian Analysis of Molecular Evolution Using MrBayes:'' Huelsenbeck and Ronquist do not intend to describe all possible options in MrBayes but work toward their example analysis of a mixture model with many different substitution models. Their fine example of substitution mod-

[1] *Statistical Methods in Molecular Evolution*. Rasmus Nielsen, ed. 2005. Springer, New York. xii + 508 pp. HB $89.95, ISBN 0-387-22333-9.

els and transition probabilities is easy to translate into a class lecture on models of molecular evolution. The discussion of likelihood versus Bayesian inference and the discussion of model misspecification are somewhat brief. The sections on model choice and the estimation of divergence time highlight the strengths of Bayesian approaches when dealing with complicated mixture models. (8) ''Estimation of Divergence Times from Molecular Sequence Data:'' Thorne and Kishino give an overview of the difficulties in estimating divergence dates when allowing for variable rates among different evolutionary lineages. This chapter is a must-read for anyone thinking about estimating divergence times.

Models of molecular evolution.—A subset of possible models used in molecular evolution is described. These models cover a wide range of interests, from estimation of mutation rate parameters, gene rearrangement, and protein structure to phylogenetic inference that takes into account differences among sites in the sampled sequences. (9) ''Markov Models of Protein Sequence Evolution:'' Dimmic gives a concise overview of the development of protein substitution models. He describes the early attempts by Dayhoff to count substitutions using a parsimony-like framework to generate a static rate of change matrix; attempts to use likelihood to estimate parameters of this rate matrix by Hasegawa and colleagues; and recent advancements made through application of parameter-rich hidden Markov models that take into account structural information. The descriptions of the models are short but enable readers to decide which model they want to investigate further. (10) ''Models of Microsatellite Evolution:'' Calabrese and Sainudiin describe models of microsatellite evolution and present an impressively long list of models. Most of the literature describing or using the more complicated models seems to be centered on completely sequenced species. It is evident that the simple ladder model of Ohta and Kimura does not describe the data very well; models that allow for mutation rates depending on repeat number fit better. One would hope that program writers will incorporate some of these more sophisticated models. (11) ''Genome Rearrangements:'' Durrett discusses Hannenhalli and Pevzners inversion model for two chromosomes and extensions of this model. The algorithm for finding the minimal number of inversions to transform one chromosome into another is described in depth. This chapter is very technical. The description of formal models explaining inversions is certainly useful, but I had hoped for less depth and more breadth. (12) ''Phylogenetic Hidden Markov Models:'' Siepel and Haussler give a brief introduction to hidden Markov models that account for variable substitution patterns along sequences that are phylogenetically related to each other (phylo-HMM). They describe concepts theoretically and provide examples. One gets the impression from this chapter that any problem involving trees and sequences can be solved as a phylo-HMM, although it seems that sometimes the technicalities are overwhelming—as in the example of allowing for coestimation of alignment and phylogeny, in which the authors consider inferences with more than three sequences to be very difficult.

Inferences of molecular evolution.—This section is heterogeneous. It contains some inferences based on real data but most often describes examples of a new inference method. This organization is fine with me, but readers looking for applications of the methods and models described in earlier sections will be disappointed. Perhaps the section should have been titled ''More Models and Error Assessment of Inferences in Molecular Evolution.'' (13) ''The Evolutionary Causes and Consequences of Base Composition Variation:'' McVean discusses the difficulties in understanding the differences of base composition among species or different genomic regions. The introduction to substitution models can be easily skipped when one has read the book up to this chapter, but it is certainly a welcome refresher for readers browsing this book haphazardly. McVean introduces a simple population genetic model that uses two states—for example, GC versus AT or preferred versus unpreferred codon. The model connects selection pressure with observed frequencies and allows us to make statistical inferences. The chapter about phylo-HMM suggests a possible improvement of this crude model, as McVean points out. (14) ''Statistical Alignment: Recent Progress, New Applications, and Challenges:'' Lunter, Drummond, Miklos, and Hein describe their method of multiple sequence alignment, which is an implementation and further development of Thorne, Kishino, and Felsenstein's model that jointly takes into account possible alignments of sample sequences and a mutation model. It seems that this work is similar in scope to the alignment problem hinted at in the chapter by Siepel and Haussler, also using a phylo-HMM to coestimate sequence alignment and phylogeny. Lunter et al. do not intend to solve the HMM analytically but revert to a Bayesian framework, using MCMC for the inference. This seems to work for small trees with about 10 sequences. (15) ''Estimating Substitution Matrices:'' Yap and Speed discuss several ways that amino acid substitution parameters can be estimated by maximum likelihood from pairs of sequences or from many sequences, assuming a tree structure. (16) ''Posterior Mapping and Posterior Predictive Distributions:'' This chapter by Jonathon Bollback could easily have been two chapters, given how different the topics are. Mapping of characters on phylogenetic trees has a long tradition and was until recently dominated by maximum parsimony-based approaches. Statistical approaches, placement of ancestral states using maximum likelihood or, as Bollback describes, Bayesian posterior mapping, offer more robust estimation of ancestral states and their uncertainty. The section about predictive distributions highlights the use of Bayesian approaches in a hypothesis framework. For many problems, we cannot assume that the distribution of the test statistics is known, and in a likelihood framework, we would resort to bootstrapping. With parametric bootstrapping, we take the maximum likelihood estimate to generate many new datasets, and these are the base of the test statistic values. This ignores the uncertainty of the estimate itself. Predictive distributions remedy this problem by generating new datasets and test statistic values from the posterior distribution of the Bayesian inference of the data. Bollback discusses pros and cons of this emerging method of hypothesis testing in molecular evolution. (17) ''Assessing the Uncertainty in Phylogenetic Inference:'' Shimodaira and Hasegawa present a more standard view (compared with Bollback) of assessing the confidence that we should have in a specific, inferred phylogenetic tree. They discuss nonpara-

metric bootstrap, several testing methods (for example, the use of parametric bootstrap to evaluate confidence of substitution model parameters), model selection tests, and problems with multiple comparisons.

I applaud that the whole volume takes the position that statistical treatment of evolution is not only appropriate but preferred; all chapters discuss model-based inference of the evolutionary processes. The philosophical dispute over whether maximum parsimony or statistical approaches are the correct way to infer phylogenetic relationships is still ongoing; that discussion often ignores imperfection of the data at hand and the fact that any estimate should be accompanied by a discussion of the uncertainty. Without an explicit model, it is often difficult to meaningfully express this uncertainty, but statistical/probabilistic approaches have a notion of error built in.

Statistical Methods in Molecular Evolution shows innovative ways to analyze problems not only in molecular evolution but also in all parts of evolutionary biology. Concepts described in this book can be easily transferred to other areas of research besides the ones discussed if one is willing to phrase the problem in a rigorous mathematical model framework. In this sense, I read this book not as a narrow treatise for researchers interested in molecular evolution but as a seminal work for all researchers interested in questions related to evolutionary inference. In several chapters, examples include models of quantitative characters, such as morphology data. Obviously, the approaches developed for inferences on molecules apply to more complicated systems; it will be only a matter of time until all inferences in evolutionary biology can be done using probabilistic analyses. For example, it would be interesting to see whether Bayesian mappings of behavioral characters or morphological patterns on phylogenetic trees give better results than the currently more common parsimony-based mappings. *Statistical Methods in Molecular Evolution*, seen as a subset of evolutionary study by some, is ready to leave the box and be considered as *(Statistical) Methods in the Study of Evolution*.

Molecular evolution is still focused on phylogenies—that is, their accurate representation for large sets of taxa using increasingly complex mutation models—or on estimates of parameters of these mutation models. Many chapters in this book emphasize species comparisons: for example, estimation of phylogenies, selection pressure, or mutation parameters. Only a few chapters address population genetic problems directly. Mathematical population genetics, having a longer history than mathematical phylogenetics, was centered for many years on ''frequency-based'' approaches. Only 25 years ago, Kingman developed the coalescence theory (see Kingman 2000). This theory allows us to calculate probabilities on ancestral relationships among sampled individuals (genealogies). Coalescence theory facilitates the merger of population genetics theory with phylogenetic approaches, but still, there are mostly two camps: phylogeneticists and population geneticists. Only a few people are moving freely between them. Rasmus Nielsen is certainly one of these researchers, and his work so far has merged many population genetic and phylogenetic aspects of biological research under the umbrella of molecular evolution. Although Nielsen did not contribute a chapter to his book, his work permeates all its chapters. This book gives an overview of his interests and current achievements in molecular evolution. In short, this book should be on your bookshelf.

Literature Cited

Felsenstein, J. 2004. Phylogenetic inference. Sinauer Associates, Sunderland, MA.

Green, P. J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82: 711–732.

Kingman, J. F. 2000. Origins of the coalescent. 1974–1982. Genetics 156:1461–1463.

Swofford, D., G. Olsen, P. Waddell, and D. Hillis. 1996. Phylogenetic inference. Pp. 407–514 *in* D. Hillis, C. Moritz, and B. Mable, eds. Molecular systematics. Sinauer Associates, Sunderland, MA.

Book Review Editor: D. Futuyma