
Preface

Studies of evolution at the molecular level aim to address two major questions: reconstruction of the evolutionary relationships among species and investigation of the forces and mechanisms of the evolutionary process. The first is the realm of systematics, and is traditionally studied using morphological characters and fossils. The great utility and easy availability of molecular data have made molecules the most common type of data used for phylogeny reconstruction in most species groups. The second question concerning the mechanisms of molecular evolution is studied by estimating the rates of nucleotide and amino acid substitutions, and by testing models of mutation and selection using sequence data.

Both areas of research have experienced phenomenal growth in the past few decades, due to the explosive accumulation of genetic sequence data, improved computer hardware and software, and development of sophisticated statistical methods suitable for addressing interesting biological questions. By all indications, this growth is bound to continue, especially on the front of data generation. Phylogenetic analysis has entered the genomic age, with large data sets consisting of hundreds of species or sequences analysed routinely. The debate of morphology versus molecules is largely over; the values of both kinds of data are well appreciated by most researchers. The philosophical debate concerning parsimony versus likelihood is ongoing but appeared to become less acrimonious. Much exciting progress has been made to develop and implement powerful statistical methods and models, which are now used routinely in analysis of real data sets.

The time appears ripe to summarize the methodological advancements in the field, and this book is such an attempt. I make no effort to be comprehensive in the coverage. There is hardly such a need now, thanks to recent publication of Joseph Felsenstein's (2004) treatise, which has discussed almost everything relevant to phylogenies. Instead I take the view that molecular evolutionary analysis, including reconstruction of phylogenies and inference of the evolutionary process, is a problem of statistical inference (Cavalli-Sforza and Edwards 1967). Thus well-established statistical methods such as likelihood and Bayesian are described as standard. Heuristic and approximate methods are discussed from such a viewpoint and are often used to introduce the central concepts, because of their simplicity and intuitive appeal, before more rigorous methods are described. I include some discussions of implementation issues so that the book can serve as a reference for researchers developing methods of data analysis.

The book is written for upper-level undergraduate students, research students, and researchers in evolutionary biology, molecular systematics, and population genetics. It is hoped that biologists who have used software programs to analyse their own data will find the book particularly useful in helping them understand the workings of the methods. The book emphasizes essential concepts but includes detailed mathematical derivations, so it can be read by statisticians, mathematicians, and computer scientists, who would like to work in this exciting area of computational biology.

The book assumes an elementary knowledge of genetics, as provided, for example, by Chapter 1 of Graur and Li (2000). Knowledge of basic statistics or biostatistics is assumed, and calculus and linear algebra is needed in some parts of the book. Likelihood and Bayesian statistics are introduced using simple examples and then used in more sophisticated analyses. Readers who would like a systematic and comprehensive treatment of these methods should consult many of the excellent textbooks in probability theory and mathematical statistics, for example DeGroot and Schervish (2002) at the elementary level, and Davison (2003), Stuart *et al.* (1999), and Leonard and Hsu (1999) at more advanced levels.

The book is organized as follows. Part I consists of two chapters and introduces Markov-process models of sequence evolution. Chapter 1 discusses models of nucleotide substitution and calculation of the distance between a pair of sequences. This is perhaps the simplest phylogenetic analysis, and I take the opportunity to introduce the theory of Markov chains and the maximum likelihood method, which are used extensively later in the book. As a result, this chapter is probably most challenging for the biologist reader. Chapter 2 describes Markov-process models of amino acid and codon substitution and their use in calculation of the distance between two protein sequences and in estimation of synonymous and nonsynonymous substitution rates between two protein-coding DNA sequences. Part II deals with methods of phylogeny reconstruction. Parsimony and distance methods are discussed briefly (Chapter 3), while likelihood and Bayesian methods are covered in depth (Chapters 4 and 5). Chapter 5 is an expanded version of the chapter in *Mathematics in Phylogeny and Evolution* edited by Olivier Gascuel (Oxford University Press, 2005). Chapter 6 provides a review of studies that compare different phylogeny reconstruction methods and covers testing of trees. Part III discusses a few applications of phylogenetic methods to study the evolutionary process, such as testing the molecular clock and using the clock to estimate species divergence times (Chapter 7), and applications of models of codon substitution to detect natural selection affecting protein evolution (Chapter 8). Chapter 9 discusses basic techniques of computer simulation. Chapter 10 includes a discussion of current challenges and future perspectives of the field. A brief review of major phylogenetics software packages is included in Appendix C. Sections marked with an asterisk * are technical and may be skipped.

Example data sets used in the book and small C programs that implement algorithms discussed in the book are posted at the web site for the book: <http://abacus.gene.ucl.ac.uk/CME/>. It will also include a list of errors discovered since publication of the book. Please report errors you discover to the author at z.yang@ucl.ac.uk.

I am grateful to a number of colleagues who read earlier versions of chapters of this book and provided constructive comments and criticisms: Hiroshi Akashi (chapters 2 and 8), Adam Eyre-Walker (chapter 2), Jim Mallet (chapters 4 and 5), Konrad Scheffler (chapters 1, 5, and 8), Elliott Sober (chapter 6), Mike Steel (chapter 6), Jeff Thorne (chapter 6), Simon Whelan (chapter 1), and Anne Yoder (chapter 6). Special thanks go to Karen Cranston, Ligia Mateiu and Fengrong Ren, who read the whole book and provided many detailed suggestions. Jessica Vamathevan and Richard Emes were victims of my experiments when I tested some difficult passages. Needless to say, all errors that remain are mine. Thanks are also due to Ian Sherman and Stefanie Gehrig at Oxford University Press for initiating this project and for valuable support and patience throughout it.

Ziheng Yang
London
March 2006