
Preface

The field of molecular evolution is devoted to elucidating the processes generating variation within and between species at the molecular level. It addresses the fundamental question of why all life on Earth looks as it does from the perspective of molecular and evolutionary biology. Molecular evolution arose as a scientific field in the 1960s after protein sequences from multiple species first became available. In the 1970s, the first journal exclusively devoted to this field arose, and today it molecular evolution dominates the literature on evolutionary biology. Since the appearance of large-scale genomic data, the field of molecular evolution has emerged as one of the major scientific pillars in the analysis of genomic data, especially when it comes to data from multiple species.

The field of molecular evolution relies heavily on statistical theory. Usually, researchers only have access to DNA data, or other molecular data, from extant species. From such data they try to make inferences regarding past evolutionary processes. This inference problem is fundamentally statistical in nature, and it is not surprising that a large body of literature on statistical methods in molecular evolution has emerged. The statistical problems encountered in molecular evolution are often rather non-standard because the underlying statistical models usually involve superimposing stochastic processes along the edges of a tree. Several interesting and peculiar algorithmic and statistical problems arise from these models. While many books contain excellent coverage of the specialized, but important, area of phylogenetic inference (estimation of trees) from molecular data, there are no books that provide an introduction to the more general area of statistical methods in molecular evolution. With the publication of this book, we hope to rectify this problem.

The first four chapters of the book provide a general introduction to the area. The first chapter, by Galtier and his colleagues, introduces the models of DNA sequence change usually applied in molecular evolution. Evolution does not remember past states except by current form, and Markov models have therefore been the natural choice for statistical models to describe the evolution of DNA sequences. The first chapter provides an introduction to these

models and sets the stage for the subsequent chapters describing how the models are used for statistical inferences. The second chapter, by Buschbom and von Haeseler, provides an introduction to the use of the likelihood function in molecular evolution. In addition to introducing some basic statistical concepts for the uninitiated biologist, it also provides an introduction to the computational aspects involved in calculating sampling distributions based on the Markov models discussed in the first chapter. The third chapter, by Larget, provides an introduction to the use of Markov chain Monte Carlo (MCMC) methods in molecular evolution. Methods based on MCMC are receiving an increasing amount of attention and the third chapter provides an introduction to this area. Chapter 4, by Bustamante, provides an introduction to population genetic theory with special emphasis on areas of relevance for molecular evolution. The classical mathematical theory underlying studies of molecular evolution is population genetic, and much research in molecular evolution cannot be fully understood without an appreciation of population genetic theory.

The second section in the book contains four chapters written by the authors of some of the most important statistical computer packages used in the study of molecular evolution. These chapters discuss practical statistical approaches for analyzing DNA sequences and molecular data. Chapter 5, by Bielawski and Yang, discusses methods for detecting natural Darwinian selection using the program *Paml*. *Paml* is possibly the most commonly used computer program for analyzing models of molecular evolution. Chapter 6, by Pond and Muse, discusses a recently developed versatile computer package, *HyPhy*, for analyzing models of molecular evolution. Readers interested in developing and analyzing new models of molecular evolution may want to take advantage of this computer package. Chapter 7, by Huelsenbeck and Ronquist discusses Bayesian inference in molecular evolution based on the popular computer program *MrBayes*. The use of Bayesian methods in molecular evolution is quite new but has already had a tremendous impact on the field, largely due to the availability of *MrBayes*. Chapter 8 is written by the authors of the computer program *Multidivtime*, Kishino and Thorne, and it discusses statistical issues relating to the molecular clock. The molecular clock assumption is that the rate of molecular change has been constant through evolutionary time. This assumption has been used extensively in the literature to date evolutionary events, but on numerous occasions it has also been shown to be not invalid. The main focus of Chapter 8 is to discuss methods for dating evolutionary events when the molecular clock assumption is not met.

The third section introduces other models of molecular evolution beyond the basic nucleotide-based Markov chain models that were the main focus of Chapter 1. Chapter 9, by Dimmic, discusses models of protein evolution. Such models are important not only because they can be used for making inferences regarding protein evolution and function but also because the substitution matrices estimated using these models are important in alignment algorithms. The issue of estimating substitution matrices is revisited in Chap-

ter 15, and the issue of alignment is dealt with in depth in Chapter 14. In Chapter 10, Calabrese and Sainudiin discuss models of microsatellite evolution and statistical inferences based on these models. Microsatellites are small repeated patterns of DNA and are used extensively in many genetic studies because they are highly variable. The models used to analyze this type of data are fundamentally different from most other models used to analyze DNA sequence data. Chapter 11, by Durrett, discusses methods and models for analyzing whole-genome evolution incorporating rearrangements such as inversions and translocations. Although these models are still in their infancy, they have become highly relevant with the recent availability of large-scale genomic data. Chapter 12, by Siepel and Haussler, provides an introduction to the use of Hidden Markov models (HMMs) in the study of molecular evolution. Such models are very important when certain properties of the evolutionary process are thought to vary among positions along the DNA sequence. The use of HMMs for statistical alignment is also discussed in Chapter 14.

The last section of the book contains five chapters that further detail methods of inference in molecular evolution. In Chapter 13, McVean relates the Markov models of molecular evolution discussed in most chapters of this book to the population genetic models discussed in Chapter 4 in the context of variation of nucleotide composition among species. The frequency of different nucleotides is known to vary among species. This observation is interesting from the perspective of evolutionary biology, and it is also highly relevant to our choice of models for analyzing molecular evolution. In Chapter 14 on statistical alignment, Lunter, Drummond, Mikls, and Hein explore the relationship among evolutionary models, trees and the problem of aligning DNA or protein sequences. In Chapter 15, Yap and Speed discuss the estimation of substitution matrices for use in alignment problems. The last two chapters, Chapter 16 by Bollback and Chapter 17 by Shimodaira and Hasegawa, discuss issues related to hypothesis testing and model choice in molecular evolution. Bayesian methods have recently gained much popularity in the area of molecular evolution, leading to a debate regarding choice of statistical methodology not dissimilar to the discussions that have occurred in many other areas of applied statistics. Chapter 17 argues for the use of certain frequentist procedures for the tree estimation problem and discusses problems with Bayesian procedures, while other chapters of the book (e.g., Chapter 7) provide more optimistic views of the use of Bayesian methods in molecular evolution. Chapter 16 discusses the use of posterior predictive distributions for statistical inferences, in addition to providing a review of recent methods for estimating the history of mutations from DNA sequence data.

This book provides a comprehensive review of the many interesting statistical problems arising in molecular evolution provided by leading researchers in the field. It is intended for researchers and students from the statistical and biological sciences alike. For the statistician, the book will provide an introduction to an exciting area of application that often has been overlooked by statisticians. For the biologist, the book provides an introduction to the theory

underlying many of the methods they use in their daily research. Several of the chapters, including the four introductory chapters, are also highly suitable as texts for advanced undergraduate or graduate-level courses in molecular evolution.

Copenhagen, April 2004

Rasmus Nielsen