
Preface

“Mathematics seems to endow one with something like a new sense.”

Charles Darwin

The goal of population genetics is to understand how genetic variability is shaped by natural selection, demographic factors, and random genetic drift. The stochastic evolution of a DNA segment that experiences recombination is a complex process, so many analyses are based on simulations or use heuristic methods. However, when formulas are available, they are preferable because, when simple, they show the dependence of observed quantities on the underlying parameters and, even when complicated, can be used to compute exact answers in a much shorter time than simulations can give good approximations.

The goal of this book is to give an account of useful analytical results in population genetics, together with their proofs. The latter are omitted in many treatments, but are included here because the derivation often gives insight into the underlying mechanisms and may help others to find new formulas. Throughout the book, the theoretical results are developed in close connection with examples from the biology literature that illustrate the use of these results. Along the way, there are many numerical examples and graphs to illustrate the main conclusions. To help the reader navigate the book, we have divided the sections into a large number of subsections listed in the index, and further subdivided the text with bold-faced headings (as in this Preface).

This book is written for mathematicians and for biologists alike. With mathematicians in mind, we assume no knowledge of concepts from biology. Section 1.1 gives a rapid introduction to the basic terminology. Other explanations are given as concepts arise. For biologists, we explain mathematical notation and terminology as it arises, so the only *formal* prerequisite for biologists reading this book is a one-semester undergraduate course in probability and some familiarity with Markov chains and Poisson processes will be very useful. We have emphasized the word *formal* here, because to read and under-

stand all of the proofs will require more than these simple prerequisites. On the other hand, the book has been structured so that proofs can be omitted.

What is in this book?

Chapter 1 begins with the theory of neutral evolution in a homogeneously mixing population of constant size. We introduce and study the discrete-time Wright-Fisher model, the continuous-time Moran model, the coalescent, which describes the genealogy of a nonrecombining segment of DNA, and two simplified models of mutation: the infinite alleles and infinite sites models. Based on these results, Chapter 2 introduces the problem of testing to see if observed DNA sequences are consistent with the assumptions of the “null model” underlying the theory developed in Chapter 1.

Chapters 3 through 6 confront the complications that come from relaxing the assumptions of the models in Chapter 1. This material, which filled two chapters in the first edition, has doubled in size and contains many results from the last five years. Chapter 3 introduces the ancestral recombination graph and studies the effect of recombination on genetic variability and the problem of estimating the rate at which recombination occurs. Chapter 4 investigates the influence of large family sizes, population size changes, and population subdivision in the form of island models on the genealogy of a sample. Chapter 5 concerns the more subtle behavior of the stepping stone model, which depicts a population spread across a geographical range, not grouped into distinct subpopulations. Finally, Chapter 6 considers various forms of natural selection: directional selection and hitchhiking, background selection and Muller’s ratchet, and balancing selection.

Chapters 7 and 8, which are new in this edition, treat the previous topics from the viewpoint of diffusion processes, continuous stochastic processes that arise from letting the population size $N \rightarrow \infty$ and at the same time running time at rate $O(N)$. A number of analytical complications are associated with this approach, but, at least in the case of the one-dimensional processes considered in Chapter 7, the theory provides powerful tools for computing fixation probabilities, expected fixation time, and the site frequency spectrum. In contrast, the theory of multidimensional diffusions described in Chapter 8 is more of an art than a science. However, it offers significant insights into recombination, Hill-Robertson interference, and gene duplication.

Chapter 9 tackles the relatively newer, and less well-developed, study of the evolution of whole genomes by chromosomal inversions, reciprocal translocations, and genome duplication. This chapter is the least changed from the previous edition but has new results about when the parsimony method is effective, Bayesian estimation of genetic distance, and the midpoint problem.

In addition to the three topics just mentioned, there are a number of results covered here that do not appear in most other treatments of the subject (given here with the sections in which they appear): Fu’s covariance matrix for the site frequency spectrum (2.1), the sequentially Markovian coalescent (3.4), the beta coalescent for large family sizes (4.1), Malécot’s recursion for

identity by descent and its study by Fourier analysis (5.2), the “continuous” (or long-range) stepping stone model (5.5–5.6), Muller’s ratchet and Kondrashov’s result for truncation selection (6.4), approximations for the effect of hitchhiking and recurrent selective sweeps (6.5–6.7), the Poisson random field model (7.11), fluctuating selection (7.12), a new approximate formula for the effect of Hill-Robertson interference (8.3), and a new result showing that the subfunctionalization explanation of gene duplication is extremely unlikely in large populations (8.6).

Having bragged about what I do cover, I must admit that this book has little to say about computationally intensive procedures. Some of these methods are mentioned along the way, and in some cases (e.g., Hudson’s composite likelihood method for estimating recombination rates, and the Kim and Stephan test) we give some of the underlying mathematics. However, not being a user of these methods, I could not explain to you how to use them any better than I could tell you how to make a chocolate soufflé. As in the case of cooking, if you want to learn, you can find recipes on the Internet. A good place to start is www.molpopgen.org.

Mens rea

In response to criticisms of the first edition and the opinions of a half-dozen experts hired to read parts of the first draft of the second edition, I have worked hard to track down errors and clarify the discussion. Undoubtedly, there are bugs that remain to be fixed, five years from now in the third edition. Comments and complaints can be emailed to rtd1@cornell.edu. My web page www.math.cornell.edu/~durrett can be consulted for corrections.

Interdisciplinary work, of the type described in the book, is not easy and is often frustrating. Mathematicians think that it is trivial because, in many cases, the analysis does not involve developing new mathematics. Biologists find the “trivial” calculations confusing, that the simple models omit important details, and are disappointed by the insights they provide. Nonetheless, I think that important insights can be obtained when problems are solved analytically, rather than being conquered by complicated programs running for days on computer clusters.

I would like to thank the postdocs and graduate students who in recent years have joined me on the journey to the purgatory at the interface between probability and biology (in my case, genetics and ecology): Janet Best, Ben Chan, Arkendra De, Emilia Huerta-Sanchez, Yannet Interian, Nicholas Lanchier, Vlada Limic, Lea Popovic, Daniel Remenik, Deena Schmidt, and Jason Schweinsberg. I appreciate the patience of my current co-authors on this list as I ignored our joint projects, so that I could devote all of my energy to finishing this book.

As I write this, a January (2008) thaw is melting the snow in upstate New York, just in time so that my wife (and BFF) Susan can drive my younger son, Greg, back to MIT to start his fourth semester as a computer scientist/applied mathematician. My older son David, a journalism student in the Park School

at Ithaca College, and I still have two weeks before classes start. Lying back on the sofa proofreading and revising the text while the cats sleep by the fire, it seems to me that academic life, despite its many frustrations, sure beats working for a living.

Rick Durrett

Several figures included here come from other sources and are reprinted with permission of the publisher in parentheses. Figures 3.6, 3.7, and 3.8, from Hudson (2001), Figures 6.3 and 6.4 from Hudson and Kaplan (1988), and Figure 6.6 from Hudson and Kaplan (1995) (Genetics Society of America). Figure 8.7 from Lynch and Conrey (2000) (AAAS). Figure 4.3 from Cann, Stoneking and Wilson (1987) (Nature Publishing Company).



<http://www.springer.com/978-0-387-78168-6>

Probability Models for DNA Sequence Evolution

Durrett, R.

2008, XII, 431 p., Hardcover

ISBN: 978-0-387-78168-6