Quantifying Life: A Symbiosis of Computation, Mathematics, and Biology

Dmitry A. Kondrashov

The University of Chicago Press
Chicago and London

Contents

Chapter 0

Introduction

> *"What is a man," said Athos, "who has no landscape? Nothing but mirrors and tides."*
>
> —Anne Michaels, *Fugitive Pieces*

0.1 What is mathematical modeling?

A *mathematical model* is a representation of some real object or phenomenon in terms of quantities (numbers). The goal of modeling is to create a description of the object in question that may be used to pose and answer questions about it without doing hard experimental work. A good analogy for a mathematical model is a map of a geographic area: a map cannot record all the complexity of the actual piece of land, because the map would need to be size of the piece of land, and then it wouldn't be very useful! Maps, and mathematical models, need to sacrifice the details and provide a bird's-eye view of reality to guide the traveler or the scientist. The representation of reality in the model must be simple enough to be useful, yet complex enough to capture the essential features of what it is trying to represent.

Since the time of Newton, physicists have been very successful at using mathematics to describe the behavior of matter of all sizes, ranging from subatomic particles to galaxies. However, mathematical modeling is a new arrow in a biologist's quiver. Many biologists would argue that living systems are much more complex than either atoms or galaxies, since even a single cell is made up of a mind-boggling number of highly dynamic, interacting entities. This complexity presents a great challenge and fascinating new questions.

New advances in experimental biology are producing data that make quantitative methods indispensable for biology. The advent of *genetic sequencing* in the 1970s and 1980s has allowed us to determine the genomes of different species, and in the past few years next-generation sequencing has reduced sequencing costs for an individual human genome to a few thousand dollars. The resulting deluge of quantitative data has answered many outstanding questions and has also led to entirely new ones. We now understand that knowledge of genomic sequences is not enough for understanding how living things work, so the burgeoning field of *systems biology* investigates the interactions among genes, proteins, or other entities. The central problem is to understand how a network of interactions among individual molecules can lead to large-scale results, such as the development of a fertilized egg into a complex organism. The human mind is not suited for making correct intuitive judgements about networks comprised of thousands of actors. Addressing questions of this complexity requires quantitative modeling.

0.2 Purpose of this book

This textbook is intended for a college-level course for biology and pre-medicine majors, or for more established scientists interested in learning the applications of mathematical methods to biology. The book brings together concepts found in mathematics, computer science, and statistics courses to provide the student a collection of skills that are commonly used in biological research. The book has two overarching goals. The first is to explain the quantitative language that often is a formidable barrier to understanding and critically evaluating research results in biological and medical sciences. The second is to teach students computational skills that they can use in their future research endeavors. The main premise of this approach is that computation is critical for understanding abstract mathematical ideas.

These goals are distinct from those of traditional mathematics courses that emphasize rigor and abstraction. I strongly believe that understanding mathematical concepts is not contingent on being able to prove all of the relevant theorems. Instead, premature focus on abstraction obscures the ideas for most students; it is putting the theoretical cart before the experiential horse. I find that students can grasp deep concepts when they are allowed to experience them tangibly as numbers or pictures, and those with an abstract mindset can generalize and add rigor later. As I demonstrate in part 3 of the book, Markov chains can be explained without relying on the machinery of measure theory and stochastic processes, which require graduate-level mathematical skills. The idea of a system randomly hopping between a few discrete states is far more accessible than sigma algebras and martingales. Of course, some abstraction is necessary when presenting mathematical ideas, and I provide correct definitions of terms and supply derivations when I find them to be illuminating. But I avoid rigorous proofs and always favor understanding over mathematical precision.

The book is structured to facilitate learning computational skills. Over the course of the text, students accumulate programming experience, progressing from assigning values to variables in Chapter 1 to solving nonlinear Ordinary differential equations (ODEs) numerically by the end of the book. Learning to program for the first time is a challenging task, and I facilitate it by providing sample scripts for students to copy and modify to perform the requisite calculations. Programming requires careful, methodical thinking, which facilitates deeper understanding of the models being simulated. In my experience teaching this course, students consistently report that learning basic scientific programming is a rewarding experience, which opens doors for them in future research and learning.

It is of course impossible to span the breadth of mathematics and computation used for modeling biological scenarios. This did not stop me from trying. The book is broad but selective, sticking to a few key concepts and examples that should provide enough of a basis for a student to explore a topic in more depth later on. For instance, I do not go through the usual menagerie of probability distributions in Chapter 4 but only analyze the uniform and the binomial distributions. If one understands the concepts of distributions and their means and variances, it is not difficult to read up on the geometric or gamma distribution if one encounters it. Still, I omitted numerous topics and entire fields, some because they require greater mathematical sophistication, and others because they are too difficult for beginning programmers (e.g., sequence alignment and optimization algorithms). I hope that you do not end your quantitative journey with this book!

I take an even more selective approach to the biological topics presented in every chapter. The book is not intended to teach biology, but I do introduce biological questions I find interesting, refer to current research papers, and provide discussion questions for you to wrestle with. This requires a basic explanation of terms and ideas, so most chapters contain a broad summary of a biological field, such as measuring mutation rates, epidemiology modeling, hidden Markov models for gene structure, and limitations of medical testing. I hope the experts in these fields forgive my omitting the interesting details that they spend their lives investigating, and trust that I managed to get the basic ideas across without gross distortion.

0.3 Organization of the book

Each chapter in the textbook is centered around a mathematical concept, along with models, biological applications, and programming. This multipronged approach provides a diverse set of teaching tools: motivational questions from biology can be formalized using mathematical terms, solved for simple cases on the board, and then demonstrated in more complex manifestations using the programming language R. Each chapter contains enough material for a week of learning and includes various assignments. The mathematics sections contain simple practice problems for the corresponding mathematical skills, the programming sections contain either debugging exercises or simple programming assignments, and the biological modeling sections contain discussion questions intended to stimulate students to think about assumptions and limitations of the models (and they frequently require students to read and digest a research paper). Each chapter ends with multi-question computational projects that walk students through implementing and investigating a computational model for a biological question.

Part 1 of the textbook (Chapters 1–5) starts with elementary mathematical ideas: variables and parameters, basic functions and graphs, and descriptive statistics. These simple concepts pair well with rudimentary programming steps that are introduced concurrently. Despite the conceptual simplicity, the first attempts at writing and executing code are invariably difficult for students, so I find this combination pedagogically sound. More advanced students can treat the first three chapters as review, but those who have never written code before are advised to focus on the programming exercises. Chapters 4 and 5 are less elementary, and students may encounter something new in the realms of probability distributions and estimation through sampling.

Part 2 of the book (Chapters 6–9) concerns relationships between two variables, both categorical and numerical. This is a largely data-driven part of the course, but it also introduces crucial theoretical concepts that are used later, particularly conditional probability and independence. I present the standard chi-squared test for independence and then warn students about misuse of p-values in the chapter on Bayesian thinking. The ideas of linear regression are familiar to most students at this level, but few are acquainted with correlation at a more than perfunctory level. The last chapter of this part delves into nonlinear fitting using logarithmic transformations and its applications.

Part 3 of the book (Chapters 10–13) is an introduction to Markov models divided into four chapters. The story progresses from describing models with transition matrices and flow diagrams to recursive calculation of probability distribution vectors, then to stationary distributions and finally to describing dynamics using eigenvalues and eigenvectors. The level of mathematical sophistication jumps considerably, and so do the computational expectations. Students learn to generate simulated strings of Markov states and then to repeat the simulations to generate entire data sets evolving over time.

Part 4 of the book (Chapters 14–17) addresses one-variable dynamical systems. The first chapter analyzes linear discrete-time equations and their solutions; the next one graduates to linear differential equations and their solutions, which build on the discrete-time ideas. We then move to graphical analysis of nonlinear ODEs, and finish with a look at the crazy behavior and chaos in nonlinear discrete-time models.

A one-semester (or one-quarter) course based on this book can be designed in several ways. The first two parts of the book provide the necessary foundation for the next two, both mathematically and in programming skills, but parts 3 and 4 are essentially independent. One could teach a reasonable course based on either parts 1, 2, and 3, or parts 1, 2, and 4. Another option is to omit the last chapter of each part (Chapters 5, 9, 13, and 17), because they contain more advanced topics than the rest and are designed to be skipped without any detriment to the flow of ideas. I should note that with the exception of part 4 (actually only the last three chapters), none of the rest use any concepts from calculus, so one could design a course for students with shaky or nonexistent knowledge of calculus. For an audience with greater mathematical maturity, one could power through part 1 in 2–3 weeks and be able to go through most of the textbook in a semester.

A course based on this textbook can be tailored to fit the quantitative needs of a biological sciences curriculum. At the University of Chicago, the course I teach has replaced the last quarter of calculus as a first-year requirement for biology majors. This material could be used for a course without a calculus prerequisite that a student takes before more rigorous statistics, mathematics, or computer science courses. It may also be taught as an upper-level elective course for students with greater maturity who may be ready to tackle the chapters on eigenvalues and differential equations. My hope is that it may also prove useful for graduate students or established scientists who need an elementary but comprehensive introduction to the concepts they encounter in the literature or that they can use in their own research. Whatever path you traveled to get here, I wish you a fruitful journey through biomathematics and computation!