

Preface

This book provides an introduction to two important aspects of modern biochemistry, molecular biology, and biophysics: computer simulation and data analysis. My aim is to introduce the tools that will enable students to learn and use some fundamental methods to construct quantitative models of biological mechanisms, both deterministic and with some elements of randomness; to learn how concepts of probability can help to understand important features of DNA sequences; and to apply a useful set of statistical methods to analysis of experimental data. The availability of very capable but inexpensive personal computers and software makes it possible to do such work at a much higher level, but in a much easier way, than ever before.

The Executive Summary of the influential 2003 report from the National Academy of Sciences, “BIO 2010: Transforming Undergraduate Education for Future Research Biologists” [12], begins

The interplay of the recombinant DNA, instrumentation, and digital revolutions has profoundly transformed biological research. The confluence of these three innovations has led to important discoveries, such as the mapping of the human genome. How biologists design, perform, and analyze experiments is changing swiftly. Biological concepts and models are becoming more quantitative, and biological research has become critically dependent on concepts and methods drawn from other scientific disciplines. The connections between the biological sciences and the physical sciences, mathematics, and computer science are rapidly becoming deeper and more extensive.

Quantitative approaches have become particularly prominent in the large-scale approaches of systems biology and its associated high-throughput techniques: bioinformatics, genomics, proteomics, metabolomics, cellomics, etc. High levels of quantitation are also needed in some of the more biophysically oriented aspects of biochemistry, molecular and cellular biology, physiology, pharmacology, and neuroscience.

The increasing use of quantitation at the frontiers of modern biology requires that students learn some basic quantitative methods at an early stage, so that they can build on them as their careers develop. To deal with realistic biological problems, these quantitative methods need to go beyond those taught in standard courses

in calculus and the elements of differential equations and linear algebra—courses based mainly on analytical approaches—to encompass appropriate numerical and computational techniques. The types of realistic biological problems that contemporary science is facing are generally too large and complex to yield to analytical approaches, and usually specific numerical answers are desired, so it makes sense to go directly to computational rather than analytical mathematical answers.

Modern molecular and cellular biology also demands increasingly sophisticated use of statistics, a demand difficult to meet when many life science students don't take even an elementary statistics course.

To add significant instruction in computational and statistical methods to an already overcrowded biology curriculum poses a challenge. Fortunately, modern computer tools, running on ordinary personal computers, enable very sophisticated analyses without requiring much analytical or programming knowledge or effort. In essence, this is a “black box” approach to quantitative biology; but I would argue that using a set of black boxes is better than not using quantitative tools at all when they would substantially enhance the results of biological investigations. The challenge, then, is to make students—and more mature scientists—aware of the appropriate black boxes, their capabilities, and the steps needed to access those capabilities. This will require a small amount of programming and an even smaller amount of analytical manipulation.

In this book I show how to use readily available computer tools to formulate quantitative models and analyze experiments in a way that measures up to the standards of biology in the 21st century. In particular, I show how to use the free, open-source software program R in a variety of biological applications.

R is a free software environment for computer programming, statistical computing, and graphics. The R web site emphasizes statistical computing and graphics, which it does superlatively well; but R is also a very capable environment for general numerical computer programming.

The characteristics of R that make it a good choice on which to build quantitative expertise in the biochemical sciences are:

- It runs on Mac OS, Windows, and various Linux and Unix platforms.
- It is free, open-source, and undergoing continual (but not excessive) development and maintenance. It is an evolving but stable platform that you will be able to rely on for many years.
- It has a wide variety of useful built-in functions and packages, and can be readily extended with standard programming techniques.
- It has excellent graphics.
- Its capabilities are very similar to excellent and widely-used but expensive commercial programs such as Matlab.
- If needed for large, computationally demanding projects, R can be used to interface with other, speedier but less convenient programming languages.
- Once you learn its (fairly simple) syntax, it is easier and more efficient than a spreadsheet.

- It has many sample datasets, which help with learning to use the program.
- It is widely used in statistics, and is increasingly used in biological applications. See particularly the Bioconductor project primarily based at the Fred Hutchinson Cancer Research Center. Bioconductor is “an open source and open development software project for the analysis and comprehension of genomic data.”

Because of these capabilities, R can serve as your basic quantitative, statistical, and graphics tool as you develop your career.

Useful references

The book that comes closest to this one in emphasizing the numerical and programming capabilities of R, rather than mainly its statistical capabilities, is *Introduction to Scientific Programming and Simulation Using R* by Owen Jones, Robert Maillet, and Andrew Robinson, Chapman & Hall/CRC (2009) [36]. It has particular emphasis on probability and stochastic simulation.

Most of the books that teach how to use R (or its progenitor S) do so in the context of its use as a program for doing statistics. Statistics will be only one of our foci in this book, but one or more of these books may be useful for reference.

- *Introductory Statistics with R*, by Peter Dalgaard, Springer (2002) [15]
- *Using R for Introductory Statistics*, by John Verzani, Chapman & Hall/CRC (2005) [65]
- *Modern Applied Statistics with S*, 4th ed. by W.N. Venables and B.D. Ripley, Springer (2002). (The standard advanced reference.) [64]
- *Data Analysis and Graphics Using R: An Example-Based Approach* (Cambridge Series in Statistical and Probabilistic Mathematics), 2nd ed., by John Maindonald and John Braun, Cambridge University Press (2007) [43]

Several books use R in a biological context. An excellent online text at a level similar to this book is K. Seefeld and E. Linder. *Statistics Using R with Biological Examples* [55].

An advanced introduction is *Computational Genome Analysis: An Introduction* (Statistics for Biology & Health) by Richard C. Deonier, Simon Tavaré, and Michael S. Waterman, Springer (2005) [18]. The emphasis of this book is mostly on bioinformatics.

R Programming for Bioinformatics by Robert Gentleman, Chapman & Hall/CRC (2008) [26] deals with programming in R, with bioinformatics examples. It assumes a good deal of prior programming experience.

Analysis of Phylogenetics and Evolution with R by Emmanuel Paradis, Springer (2006) [49] has a useful introduction and some intermediate to research-level examples in both DNA and organismal phylogenetics and evolution.

Stochastic Modelling for Systems Biology by Darren J. Wilkinson, Chapman & Hall/CRC (2006) [67] uses some R code in its treatment of systems biology.

Bioinformatics and Computational Biology Solutions Using R and Bioconductor (Statistics for Biology and Health) edited by Robert Gentleman, Vincent Carey, Wolfgang Huber, Rafael Irizarry, and Sandrine Dudoit, Springer (2005) [27] explores the Bioconductor packages, especially as applied to analysis of microarray data. A compact book that is perhaps more suitable as an introduction to this material is *Bioconductor Case Studies* edited by Florian Hahne, Wolfgang Huber, Robert Gentleman, and Seth Falcon, Springer 2008 [31].

In my development of this book and course, I have drawn heavily on *Computer Simulation in Biology: A BASIC Introduction*, by R.E. Keen and J.D. Spain (1992) [39]. This book, which appears to be out of print, uses BASIC, an earlier and less capable computer language than R; but it has a good selection of topics and computer simulation examples for an introductory course.

Of the many recent books on mathematical and computational biology, these two fall closest to my philosophy, in their selection of topics and in emphasizing computational rather than analytical approaches:

- *Computational Cell Biology* edited by Christopher Fall, Eric Marland, John Wagner, and John Tyson, Springer (2002) [24]
- *Mathematical Models in Biology: An Introduction*, Elizabeth S. Allman and John A. Rhodes, Cambridge University Press (2004) [2]

Acknowledgments

I thank Professors David Bernlohr and Paul Siliciano for having arranged the opportunity to teach the course, Biochemistry 4950, that led to this book. I am grateful to the students in the course for giving feedback about the material and pointing out puzzlements and inconsistencies. I am particularly grateful to Geteria Onsongo for noting many typos.



<http://www.springer.com/978-1-4419-0084-5>

Computer Simulation and Data Analysis in Molecular
Biology and Biophysics

An Introduction Using R

Bloomfield, V.

2009, XVI, 321 p., Hardcover

ISBN: 978-1-4419-0084-5