Statistical Methods in Bioinformatics: An Introduction
Gregory R. Grant and Warren J. Ewens
Springer Verlag, New York; ISBN 0 38795 229 2; 476pp. US$79.95; 2001

This book consists of 14 chapters and 4 appendices; each chapter ends with a series of exercises. Chapters 1 and 2 are devoted to probability theory including definitions, discrete and continuous distributions, and random vectors. Chapters 3 and 8 deal with statistics, in particular estimators; likelihood and Bayesian analysis; and entropy and hypothesis testing, including sequential analysis. Chapters 4, 7 and 10 move on to stochastic processes such as Markov chains, random walks and more advanced topics within Markov process theory, such as Markov chain Monte Carlo (MCMC; Gibbs sampling, the Hastings–Metropolis algorithm . . .). Chapters 11 and 12, the last of the theory chapters, are on hidden Markov models and computationally intensive methods, respectively.

The theory thus introduced is then applied in the intervening chapters. Chapter 5, ‘The Analysis of One DNA Sequence’, covers shotgun sequencing and an analysis of the occurrence of words in sequences. Chapter 6, ‘The Analysis of Multiple DNA or Protein Sequences’, covers basic alignment algorithms and similarity matrices used in evaluating matched amino acid pairs. Chapter 9 goes through, in great detail, the statistics behind BLAST. Chapter 13, ‘Evolutionary Models’, covers classical Markov models of nucleotide substitutions, first in a discrete time model, then in a continuous time framework. Chapter 14, ‘Phylogenetic Tree Estimation’, covers the fundamental principles and algorithms used in tree estimation, such as distance, parsimony and likelihood.

There are several appealing sides to this book. First of all, it is self-contained and all the theory is introduced with a specific biological application in mind, which should increase the motivation for a biologist to get through the theoretical chapters. Secondly, the exposition of the theory is logical and approachable, going from probability and stochastic processes into statistics. Many interesting topics are dealt with, often in a way that, in itself, is illuminating and enriching. Thirdly, one gets a clear idea about how useful and important statistics and probability theory are to bioinformatics, and that bioinformatics involves much more than large databases and programs. Finally, it has been written in conjunction with a course, which makes it well worked through, effectively pedagogical and accessible to any student who seeks insight into theoretical aspects of bioinformatics. The list of problems at the end of each chapter is both of practical and of theoretical relevance. In essence, the book provides the reader with the statistical and probability theory necessary to understand, in detail, statistical issues in sequence analysis.

The title of the book gives the impression that it covers more than it does. Concerning the biological applications addressed, it is almost entirely limited to sequence analysis, while bioinformatics today is a lot more. A more precise title would have been ‘Basic Statistics and Probability Theory with a View Towards Sequence Analysis’. The book would have been enriched if there
had been chapters on subjects such as
gene mapping, coalescent theory and
gene expression data, since most
bioinformaticians will, at some stage, be
confronted with these issues. This is
especially surprising since Ewens is a
central contributor to some of these
fields. The first reaction to the book is that most
of its material could be found elsewhere
in three basic textbooks on probability
theory, statistics and stochastic processes,
while much of the bioinformatics part
could be found in Durbin et al.\(^1\) After
closer consideration, however, this book
is very thorough on the problems it
selects, which makes it worth serious
study.

A few sections could have benefited
from more careful elaboration. Examples
are the sections on ‘bootstrapping’ and
‘distance measures on trees’. As to the
first, the authors provide an interpretation
of the bootstrap that many in the fields of
bioinformatics and molecular biology
could learn from. The frequent
misinterpretation of the bootstrap in
journal papers and elsewhere really makes
this a topic that deserves more space and
investigation in the book, especially
because the bootstrap has such widespread
applications. In the section on distance
measures, the four-point metric is not
mentioned at all, despite its much higher
relevance in phylogeny estimation than
the ultra-metric that is treated in detail. It
would also add to the value of the book if
answers to the exercises were provided on
a suitable web page.

There are various smaller things we
find quite peculiar. In Chapter 14,
‘Phylogenetic Tree Estimation’, all
subheadings begin with ‘Tree
Reconstruction’, but nowhere is the
essential difference between estimation
and reconstruction pointed out.
Nowhere in Section 5.5, on r-scans, do
r-scans seem to be defined. Chapter 12,
‘Computationally Intensive Methods’, is
misleading. It deals with the bootstrap,
permutation tests and multiple testing
procedures. Few of these are, today,
really computationally intensive, in
contrast to many MCMC techniques
which are. Finally, the figure captions are
in general non-informative, which is a
drawback.

This book is ideal for a statistics module
in the countless MSc courses in
bioinformatics that have been initiated all
over the world. The book might be
slightly demanding here and there, but it
covers the basic theory needed to get
first-hand insight into the statistical aspects
of bioinformatics. Only a mathematically
talented biologist will feel at ease going
from completely basic probability theory
via sequential testing theory to see this
applied in the explanation of the BLAST
program all within the time-span of a few
months.

In summary, this is a very timely book
that, for many, could be rewarding
reading. It has been a widespread
misconception that bioinformatics was
mainly about the use of computer science
in the biosciences. Statistics has an equally
important role to play and the book
demonstrates this by selecting topics that
all owe a great deal to statistics.

Jotun Hein, Professor of Bioinformatics
Carsten Wiuf, Research Associate,
Department of Statistics,
Oxford University,
South Parks Road, Oxford, UK

Reference
1. Durbin, R., Eddy, S. R., Krogh, A. and
Mitchison, G. J. (1998), ‘Biological Sequence
Comparison’, Cambridge University Press,
Cambridge.

The Shattered Self: The End of
Natural Evolution
Pierre Baldi
MIT Press; ISBN 0 26202 502 7;
245pp; 2001

Bioinformaticians will be most familiar
with the work of Pierre Baldi from his
numerous publications and from