Book reviews

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 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

'Bioinformatics: A Machine Learning Approach', the book coauthored with Soren Brunak and now in its second edition. Rather further down the Amazon bestseller list, and a long way behind a market leader such as 'Self Matters: Creating Your Life From The Inside Out', you'll find 'The Shattered Self', Baldi's essay on the ethical, philosophical and practical implications of the current IT and biotechnology revolutions. It addresses such uncomfortable ideas as *in vitro* birth, organ-farming, genetically modified organisms, human cloning and artificial intelligence. For both IT and biotechnology, he asserts that 'The only way to appreciate what could happen is to understand what is happening now and project it into the future'. He adopts Brunak's new coinage 'fiction science' to describe this extrapolation but this can be easily subsumed by the existing term 'science fiction' which already comprehends H. G. Wells's Dr Moreau modifying organisms across the species barrier and Arthur C. Clarke installing a global network of geostationary satellites. Presumably 'fiction science' is used to distance its material from the sort of futuristic fiction which is often illustrated with awesome ladies sitting astride winged horses and wielding halberds. In 'science fantasy', however, we have an existing and wholly adequate phrase to describe this genre. By contrast, Baldi's book is illustrated with intriguing and attractive pictures of conjoined and identical twins and two-headed snakes. With the very broad canvas that he has undertaken to paint, none of us is going to agree with everything that Baldi predicts, but his book will help us to think about what the future portends.

Baldi enriches his book with a number of interesting order of magnitude calculations. I have been a sucker for these since hearing about Enrico Fermi accurately estimating the power of the atomic bomb explosion at Alamogordo by standing up and letting some scraps of paper fall into the blast wave. Here we learn that the number of neurons in the human brain (10^{12}) is comparable to the grains of sand on a beach; that the number of humans who have ever lived is only twice the current number; that more than half the scientists who have ever practised are alive today; and that the information received by an average brain during a lifetime is about equivalent to, but is rapidly being outstripped by, the hard-drive capacity of all the computers currently on the planet.

The author and his readers are on firmest ground when the future of information technology is addressed. With Moore's Law doubling computer power every 18 months and the Internet doubling every six months, some pretty impressive calculations will be possible. Hopefully the best brains in the planet will be able to do something more exciting with these resources than playing chess, storing pictures of naked people and making available more audio-visual entertainment than we can access in a lifetime.

As with computers, so with current biotechnology: there is a predictable squandering of talent on the utterly trivial. They are standing on the shoulders of giants and they create . . . a dead pet cloning service. Now any bereaved pet-owner with a spare US$250,000 can choose to have parts of their pet cryogenically preserved and cloned by 'Genetics Savings and Clone'. The name tells me all that I need (and rather more than I want) to know about the company.

For the future of biotechnology, Baldi needs longer time-scales and becomes both bolder and less credible in his predictions: 'within at most a few hundred years it will be possible to produce any human being with any given genome specified on a computer'. For me, there are two problems with this sort of statement. The first is that normal science rarely proceeds by small incremental steps for such a long time. A paradigm shift or a technological breakthrough is bound to come along to shorten the time-scale. The second is that, in contrast to building computers, biological systems are probabilistic rather than deterministic in their development. In particular, they are chaotically sensitive to initial conditions and random fluctuations. Thus, while I can believe it may become possible to create a human being, it will prove impossible to replicate a specific human being. A bit of Brownian motion in the zygote and, whoops, Cyrano has a cutely retroussé nose and no story to tell. Indeed we know this already from our experience of monozygotic, but emphatically not *identical*, twins. So the problem of self and self-identity in a world swamped with clones and spare body parts remains much the same as it has been since Socrates was at the Symposium.

On the other hand, technology has delivered us situations that could not have been imagined, let alone agonised over, by any reasonable Athenian up until the present generation. Baldi's response to most of these issues, and I applaud him in this, is to accept and develop the principal of utilitarianism – the greatest good for the greatest number – by asserting that there are very few absolutes in ethical matters. In the difficult areas of abortion, cloning, stem cells and genetically modified organisms we must take it one step at a time and deal with the problems as they occur. If those steps have to come faster than we'd like (and progress will relentlessly assure that they will) then we are bound to make some poor decisions. The adaptability of humans is such, however, that we will learn from these decisions and do better next time. There is no future, both literally and metaphorically, in trying to stop the deluge of technological advance. Indeed, the most heartening aspect of the book is Baldi's boundless optimism for the future. So another excellent reason for reading it is because it will be good for the morale.

Andrew T. Lloyd
INCBI, the Irish EMBnet Node

Bioinformatics: The Machine Learning Approach

Pierre Baldi and Søren Brunak
MIT Press; 2nd edn; ISBN 0 262 02506 X; 400pp; US$49.95/ £34.95 (hbk); 2001

Although this second edition of a classic book is of general interest to any bioinformatician, its main audience is anyone with a keen interest in machine learning. Ranging from optimisation techniques to neural networks, from hidden Markov models (HMMs) to grammars and linguistics, most, if not all, relevant topics are covered in this book.

My personal problem with books such as this is that they usually presuppose an unhealthy appetite for mathematics in their readers; which I definitely have not got! In this case the maths is presented clearly and in chewable amounts. Moreover, the appendices contain concise introductions to statistics, information theory, graphical (Bayesian) networks up to HMM technical detail. 'The Machine Learning Approach' is a highly practical book, though presenting an appropriate amount of detail or the theory behind the practical applications.

Chapter 1 contains the obligatory introduction to molecular biology. Fortunately the focus is on the information content of biological sequences, and does not try to provide a crash course into molecular biology, as is so often the case in bioinformatics books, although it still contains some interesting biology of a more general nature. Did you know, for example, of the existence of crosses between lions and tigers, called ligers and tigrons (like mules and hinnies the name differs depending on the sex of the male parent)? I didn't.

Chapters 2, 3 and 4 lay down the framework for machine learning methods. The first two chapters explain Bayesian probability theory, while Chapter 4 introduces the algorithms commonly used, such as dynamic programming, expectation maximisation (EM), Markov chain Monte Carlo methods, simulated annealing and genetic algorithms.

After having laid down these foundations, we arrive at the core of this book, in which the basic theory is applied to real world methods and problems. These following chapters cover diverse subjects such as neural networks (Chapters 5 and 6), HMMs (Chapters 7 and 8), graphical modelling (Chapter 9), phylogeny (Chapter 10) and grammars and linguistics (Chapter 11). Chapter 12, new in the second edition of this book, is devoted to the analysis of DNA microarray data.

What I particularly liked about these chapters is the inclusion of a plethora of examples to accompany and exemplify the theory. Neural network examples include protein secondary structure prediction, signal peptide sites, gene finding and splice sites. The examples described in the HMM chapters include a number of protein applications such as protein classification, detection of G-protein coupled receptors in expressed sequence tag (EST) databases, signal peptides and signal anchors. In the DNA and RNA field, topics such as gene finding, splice site, intron and exon prediction, and prediction of promoter regions are covered.

Chapter 9 moves on to more exotic models, such as hybrid models in which HMMs are combined with neural networks. The applications again include protein secondary structure prediction and gene finding.

The odd one out in my view is the chapter on phylogeny. When looked upon in the light of Bayesian probabilistic models of evolution, it fits in with the general concept of the book. But describing phylogeny reconstruction purely in the light of probability theory brings the subject down to the mere mechanics of tree construction. This does not do justice to such a broad and complex field of research as phylogenetics.

Chapter 12, on DNA microarrays and gene expression, is still a bit thin for a subject that is widely considered an important field of research in bioinformatics. See, for example, the December 2001 issue of *Briefings in Bioinformatics* (Vol. 2, No. 4). It is also a pity that methods such as kernel methods and SVMs (support vector machines) have been tucked away in an appendix, and have not received more attention. I would have liked to see these topics covered in more detail and, as in the other book chapters, accompanied by some relevant examples.

Chapter 13 'Internet resources and public databases', the final chapter, is somewhat of a disappointment. As the great Dutch soccer-legend and homemade philosopher Johan Cruyff once said: 'every advantage 'as its drawback'. The same principle applies here. Of course it is always dangerous to start compiling lists of links to servers, software and databases, as it will never be complete. For example consider the references to the obsolete SRS 5 server in Heidelberg (why not use SRS 6 at the EBI?) or to the NRL_3D database (which has not been updated for ages, and the link even does not exist anymore). Fortunately there is a reference to the web page[1] where most of these links were taken from, maintained at Brunak's Center for Biological Sequence Analysis. But even this site suffers from the same problems as the book does. It puzzles me for example why the reference of the WhatIf program by Gert Vriend (previously EMBL, now CMBI) points to the HGMP in Hinxton, UK. Likewise, the link for Terri Attwood's PRINTS database is still to UCL, while the database has actually been in Manchester for over three years.

Beside these minor points of criticism, having second editions of books such as 'The Machine Learning Approach' and Baxevanis and Ouelette's 'Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins' (reviewed in *Briefings in Bioinformatics*, Vol. 2, No. 4) is another proof that the field of bioinformatics has finally come to a state of maturity. Will we live to see the first edition of this book become a collector's item? I think not: it is not as if we are collecting firsts of R. L. Stevenson's 'Treasure Island'. For a field that is as young as bioinformatics is, however, it may be considered a classic.

Jack Leunissen
CMBI, EMBnet
The Netherlands

Reference

1. URL: http://www.cbs.dtu.dk/biolink.html