Bioinformatics: The Machine Learning Approach

Book · January 2001

CITATIONS
972

READS
3,694

2 authors:

Pierre Baldi
University of California, Irvine
417 PUBLICATIONS **28,456** CITATIONS
SEE PROFILE

Søren Brunak
University of Copenhagen
605 PUBLICATIONS **82,563** CITATIONS
SEE PROFILE

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