Theoretical Population Biology **58**, 77–78 (2000)
doi:10.1006/tpbi.2000.1468, available online at http://www.idealibrary.com on IDEAL®

BOOK REVIEW

R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis*: *Probabilistic Models of Proteins and Nucleic Acids*, Cambridge Univ. Press, Cambridge, UK, 1998.

This is one of the more rewarding books I have read within this field. It expanded my knowledge of several subjects and I have frequently used it for pedagogical purposes.

The recent growth in the number of determined biological sequences (and also secondary and tertiary structures of nucleotides and proteins) is well known and has created the amorphous fields of Bioinformatics, Biological Sequence Analysis, and Computational Biology, which, depending on taste and interests, can be made to cover or overlap with subject matters as diverse as methods of molecular evolution, molecular population genetics, genome analysis, and structural analysis.

Methods of molecular evolution include those for estimating phylogenies, characterizing the processes governing the evolution of sequences, and reconstructing ancestral sequences, organisms, or biochemical systems. Molecular population genetics can be viewed as molecular evolution at the population level, and includes estimating population structure (size, geography, and history), selection, ancestral sequences, and evolutionary events. Genome analysis implies the use of genetic analysis to find genes and regulatory signals, assign function to genes by sequence comparison, and locate genes influencing quantitative characters and diseases. Last, structural analysis includes the prediction (or partial prediction) of high-order structures from sequence data, comparisons of structures, molecular dynamics, docking, and much more.

All these fields have been growing quickly relative to most other fields of science, and genome analysis and structural analysis must be expected to grow enormously in coming years due to the completion of the genome projects and the medical potential of these fields.

Bioinformatics is an emerging scientific discipline driven by the flood of sequence data and the increased computational capabilities of computers coupled to the increased ability to model biological structures and their dynamics. The homology displayed by sequences and biological structures makes the efficient transfer of information between researchers ever more relevant, as they often study different variants of the same phenomena.

Expectations for this rising field vary enormously. Many molecular biologists regard bioinformatics as little more than a service that generates useful databases and programs which perform well-defined tasks. Some biocomputation enthusiasts envision that most molecular biology by the year 2010 will be done by computer, while many present-day molecular biologists are scornful of such views.

The recent emergence of this field makes it a jungle to the newcomer: there is no academic tradition related to the field, nothing that could be called a textbook has existed, until recently, and articles in the field are of highly fluctuating quality. There is no accepted bioinformatics curriculum, so bioinformaticians are a mixture of computer scientists, statisticians, physical chemists, and evolutionary and structural biologists who have changed subject matter to bioinformatics.

New books, journals, and proceedings dedicated to these subjects are now appearing several times each year. Good books describing this field perform an important service. Two volumes (183 in 1990 and 226 in 1996) in the series Methods in Enzymology, edited by Russell Doolittle, are very good introductions to this field from an application viewpoint. Dan Gusfield's "Algorithms on Strings, Trees and Sequences" (1997) is probably the best volume on algorithmic aspects. The only book that has dared to dip into the whole field is "Calculating the Secrets of Life" (Eds. Lander & Waterman, 1995). Unfortunately, the book was much too short to do any part of the field full justice. Li's "Molecular Evolution" (1997) has emerged as the standard textbook within its field. However, we are still waiting for a good book on the statistical analysis of sequences.

The book under review has been written by four authors who have all contributed substantially to the field. Richard Durbin, with a background in mathematics, is head of the Bioinformatics Division at the Sanger Centre, and thus is very close to the actual application of biological sequence analysis. Sean Eddy is a computer scientist working in St. Louis, Missouri. Graeme Mitchison is a mathematician working at the Laboratory of Molecular Biology, Cambridge, England. Anders Krogh is a physicist working at the Centre for Biological Sequence Analysis in Denmark and is especially known for his textbook on neural networks with Hertz and Palmer (1991) and his use of hidden Markov models (HMMs) in alignment problems (1994).

The authors say in the preface that "This is a subjective book by opinionated authors. It is not a tutorial on practical sequence analysis." This is true and moreover a necessity for a book that has been written by active participants in the field who do not want to be turned into full-time textbook writers.

The book comprises the following chapters:

i. An introduction, which includes a short argument in favor of Bayesian statistics.

ii. A chapter on pairwise alignment, in which several fundamental sequence algorithms, global/local alignment, scoring matrices, insertion–deletion costs, and significance of scores are discussed.

iii. Markov chains and hidden Markov models, which provides a very nice basic introduction to the concepts.

iv. Pairwise alignment using HMMs, which also includes an interesting presentation of how to use finite state automata to solve a series of alignment problems.

v. Profile HMMs; for sequence families, which shows how to align many sequences using a profile HMM and how to test if a sequence belongs to a sequence family described by a profile HMM.

vi. Multiple sequence alignment methods, which presents basic algorithms solving this problem, computational speed-ups, and different ways to score a multiple alignment (with and without a phylogeny).

vii. Building phylogenetic trees, presenting concepts relating to trees, including mainly parsimony and distance methods in phylogeny reconstruction and a section on trees and alignments that could have been in an earlier chapter.

0040-5809/00 $35.00

viii. Probabilistic approaches to phylogeny, which discusses the probability of data as a function of the tree and the evolutionary process, a Metropolis–Hastings method for investigating the tree-space, and statistical alignment and so-called tree-HMMs.

ix. Transformational grammars, which provides a very nice and concise description of the Chomsky hierarchy, stochastic analogues, and their applications in sequence analysis.

x. RNA structure analysis, again a very nice and concise treatment, with very good descriptions and exemplifications of the basic ideas dealing with this. At some points it is somewhat technical. More empirical descriptions on the importance and achievements would have been nice.

xi. An appendix on probability.

This book fulfills a very useful function and provides a shortcut to understanding for the reader, who would otherwise have to consult the primary literature. It is also quite up-to-date. It dives into the actual principles of methods, in contrast to just describing output produced by programs. A nice feature is the exercises: these are often illuminating, although at various places, I do not get the answers it says I should get. More exercises (possibly with answers) would have been even better.

The presentation is uneven. I found the chapter on grammars (Chap. 9) and the first half of the RNA chapter (chap. 10) excellent, but the last half of the RNA chapter completely impenetrable. I never understood the basic idea behind tree-HMMs (Chap. 8), despite several attempts.

There are a series of errors—these must be confusing, especially for more biologically-oriented persons. It is a help that they are listed at a web site (http://www.genetics.wustl.edu/eddy/publications/cupbook_errata.html).

Areas with which the authors are less familiar get very scant treatment. This is especially true for areas related to evolutionary biology. For example, the coalescent, a very central concept in mathematical and molecular population genetics, has been relegated to a probability measure on trees without any explanation. Also, models of molecular evolution are not treated in sufficient depth. The distinction between synonymous and nonsynonymous is not described at all, although it is fundamental in the evolutionary analysis of any gene. This level of treatment is regrettable, since most biological sequence analysis stands on the shoulders of molecular evolution and population genetics. These fields have been ignored by molecular biologists during the last two decades, largely because of their reliance on mathematics and statistics. It would be unfortunate if this tradition were to be carried on by the more mathematically literate bioinformaticians.

I would have liked a discussion of the relationship between HMM-alignment and evolution. Parsimony alignment reconstructs evolution by postulating history in terms of the insertion–deletions and substitutions needed to evolve the first sequence into the second sequence. Two nucleotides that match are interpreted as having descended from the same ancestor nucleotide. In practice, HMM-alignments are interpreted the same way, but is there any justification for that? An HMM in this context has a certain architecture and defines a probability measure on sequence space. A given alignment is a function of its architecture, but different architectures can yield the same probabilities on sequences. How then is the architecture with optimal alignment properties determined?

Their choices of references are often arbitrary. For instance, they refer not to Thorne *et al.*'s (1991) paper, which puts the foundation of alignment on a stochastic model of substitutions and insertion–deletions, but instead to a follow-up paper from 1992. The latter paper mainly addresses a technicality, concerning incorporating longer insertion–deletions, that does not seem very satisfactory.

Last, some (hopefully) constructive criticisms: A description of biological applications and examples, in which biological sequence analysis has contributed markedly, and a discussion of important subjects of rising importance, such as comparisons of structures and gene finding, might have been included.

In a field that evolves so quickly, a description of what the authors think will be needed a few years from now would have been of great interest. Durbin and Eddy, at least, must be in a position to identify problems in biological sequence analysis that need to be solved and to discuss what is needed in order to make optimal use of the data generated by a project such as the *C. elegans* genome project.

My overall evaluation is that this book is very good and a must for active participants in the field. In addition, it could be particularly useful for molecular biologists, at least those who are not suffering from mathophobia. Its main assets are the newness and freshness of the approach and the fact that concepts and problems are discussed. I much hope that it will be revised regularly and that the authors will expand the fields covered.

REFERENCES

Doolittle, R. Ed. 1990. "Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences," Methods in Enzymology, Vol. 183, Academic Press, San Diego.

Doolittle, R. Ed. 1996. "Computer Methods for Macromolecular Sequence Analysis," Methods in Enzymology, Vol. 266, Academic Press, San Diego.

Gusfield, D. 1997. "Algorithms on Strings, Trees, and Sequences," Cambridge Univ. Press, Cambridge, UK.

Hertz, J., Krogh, A., and Palmer, R. G. 1991. "Introduction to the Theory of Neural Computation," Addison–Wesley, Reading, MA.

Krogh, A. M., and Brown, *et al.* 1994. Hidden Markov models in computational biology: Applications to protein modeling, *J. Mol. Biol.* **235**, 1015–1531.

Lander, E. S., and Waterman, M. S. 1995. "Calculating the Secrets of Life," National Academy Press, Washington, DC.

Li, W.-H. 1997. "Molecular Evolution," Sinauer, Sunderland, MA.

Thorne, J. L., Kishino, H., and Felsenstein, J. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences, *J. Mol. Evol.* **33**, 114–124.

Thorne, J. L., Kishino, H., and Felsenstein, J. 1992. Inching toward reality: An improved likelihood method model of sequence evolution, *J. Mol. Evol.* **34**, 3–16.

Jotun Hein
Institute of Biological Sciences
University of Aarhus
Building 540, office 128
Ny Munkegade
DK 8000 Aarhus C, Denmark

Printed in Belgium