BOOKS

Electric genes

Terry Gaasterland

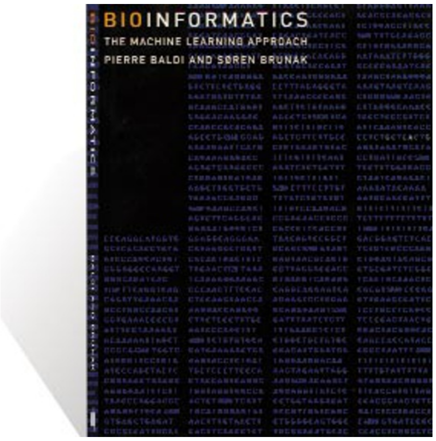
Bioinformatics: The Machine Learning Approach

Pierre Baldi and Søren Brunak
1998 MIT Press, 351 pages, $40 hardcover

The wealth of molecular biology data pouring forth from genome sequencing projects is now being augmented by high-throughput measurements of DNA, RNA, and protein expression levels in varying cells and tissues, the fledgling structural genomics initiative, and by unceasing incremental contributions from experimental laboratories around the world. This data is complex and detailed, and mixes digitized "knowledge" with information-bearing data—encoded patterns for protein molecules, their genetic regulation, and their intercoordination.

For any bioinformatics student, whether initially trained in computation, mathematics, genetics, biochemistry, physics, or biophysics, the challenge is to find a way to connect computational techniques to biologically meaningful problems in the context of this data. The student must identify tractable problems and then solve them computationally. To the extent that applicable methods exist and the goal is clear, a bioinformatician may be seen as an engineer building bridges between computation and biology. However, to the extent that the student seeks to create new methods to explore previously unmined data and to create insights, the bioinformatician is a new breed of scientist.

This book is a guide for the pure computational scientist or pure biologist who wants to become fluent in the computational details of molecular bioinformatics. *Bioinformatics: The Machine Learning Approach* contains lucid presentations of computational material reinforced with real-life examples of applications in molecular biology. The mathematical rigor of this volume is balanced by clear textual explanations.



The book lays out the computational foundations for addressing biological problems, and shows the techniques for solving them, starting with the basics of probabilistic modeling and inference, and progressing to fundamental machine learning algorithms, including dynamic programming, gradient descent, Markov chain and Monte Carlo methods, simulated annealing, and evolutionary and genetic algorithms. It then moves on to explain neural networks, hidden Markov models, stochastic grammars, and fundamental methods for phylogenetic reconstruction of evolutionary relationships, such as parsimony and maximum likelihood.

Throughout the book, the authors use real examples from problems encountered in molecular biology in the context of the ever-growing biological sequence and structure databases. Accessible both to the computer scientist and to the mathematically inclined biologist, it gives a refreshing context for problems such as RNA-folding, signal peptide identification, cleavage site prediction, functional annotation of genes, and classification and clustering of protein sequences.

The use of machine learning techniques to classify and cluster data is well suited to processing molecular data. An organism's genome contains immediate information—it encodes every possible component of a particular cell. At the same time, it records subtle differences between a specific type of organism and its very close relatives; in the context of other genomes, it yields information about the relative evolution of species.

One aspect of genome interpretation is to discover, label, and connect all of the encoded parts and to catalog the degree to which parts are lost, modified, or conserved from one type of organism to the next. Another is to discern the tiny differences between the genomes of closely related organisms—for example, single nucleotide polymorphisms, larger polymorphisms, variable regions, and microsatellite variations. A third aspect is predicting the conditions under which the differences between organisms will manifest as strength, advantage, or opportunity for collaboration. Machine learning techniques are particularly suited for facilitating these types of investigations.

For example, comparing sites that vary within a cluster of protein sequences of common function from many different bacterial organisms with the variation within a particular bacterial species can reveal the degree to which that species has explored the full space of known variants for that molecule. This, in turn, can lead to insights about the degree to which that protein is essential (and thus highly conserved), or possibly indicate the cumulative effect of impeded DNA repair systems within particular strains of the species.

With this work, Baldi and Brunak have provided a sound foundation for the process of classifying and interconnecting the hierarchy of parts encoded by genomic sequence data and their variability. Not only is the book appropriate for students new to this intersection between computation and biology, it will also prove useful for long-time workers on "classic" problems in computational molecular biology. The book has a continuity from beginning to end that helps a reader to develop an understanding of machine learning techniques and how to apply them to molecular biology. Together with *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology,* by Dan Gusfield; *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids,* edited by Richard Durbin and written with S. Eddy, A. Krogh, and G. Mitchison; and *Calculating the Secrets of Life: Applications of the Mathematical Sciences in Molecular Biology*, a collection edited by Eric S. Lander and Michael S. Waterman, this book is one of four indispensable books for the bioinformatician's library. ///

Terry Gaasterland is assistant professor and head of the laboratory of computational genomics, The Rockefeller University, 1230 York Avenue, New York, NY 10021 (gaasterland@rockefeller.edu).