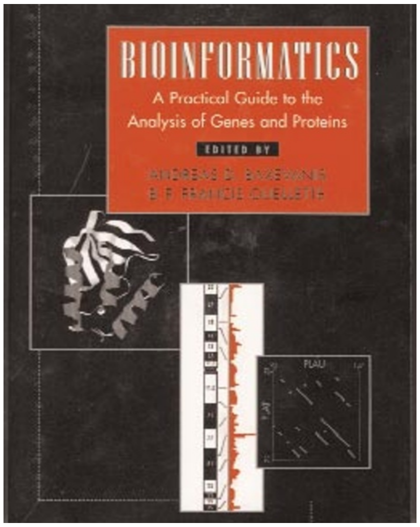
A guide for the bioinformatics novice

Terry Gaasterland

BIOINFORMATICS: a practical guide to the analysis of genes and proteins, edited by Andreas D. Baxevanis and B. F. Francis Ouellette. Published by John Wiley & Sons, Inc., New York, New York 10158, USA; 1998. 370 pages, US $ 59.95. ISBN 0-471-32441-8.

The novice user of bioinformatics tools needs a guide that answers several fundamental questions — what are these tools designed for and what can they do; what are their limitations; how does one access them, and where can one find further information. For each of the basic subfields of bioinformatics, *Bioinformatics: a practical guide to the analysis of genes and proteins*, edited by Andreas D. Baxevanis and B. F. Francis Ouellette, provides a survey, a list of world wide web addresses (URLs), and a list of monographs and reviews to which the reader may go for further information. Each chapter covers fundamental definitions and makes no assumptions about prior knowledge.

The most comprehensive chapters contain surveys of physical mapping databases (by L. Stein at Cold Spring Harbor Laboratory), predictive methods for finding genes (by J. Fickett at SmithKline Beecham Pharmaceuticals), and techniques for phylogenetic analysis (by M. Hershkovitz and D. Leipe at the National Center for Biotechnology Information (NCBI)). Useful chapters that are truly aimed at the novice user cover details about Genbank database entries (F. Ouellete), getting started with the GCG package (B. Butler), and types of methods for comparing query sequences with a large number of known sequences (G. Shuler). A very practical chapter guides the reader through the maze of database partitions in Genbank and how to submit new sequences to the NCBI (J. Kans and F. Ouellette).



In general, the book gives only one or two examples of bioinformatics software for each type of analysis and these examples tend to center on tools available through the NCBI. Consequently, many widely used tools are absent or covered incompletely. The reader should regard this book as an introduction to the types of tools available for sequence analysis, not as a comprehensive compilation. For example, there is just a brief review of the PredictProtein server (http://www.bioc.columbia.edu/predictprotein/) in the context of secondary structure prediction but it does not cover the other facets of the tool such as prediction of buried or exposed residues, hydrophobicity, or sequence variability with respect to secondary structure. Descriptions of protein structure homology modeling tools and databases are either lacking or only mentioned briefly: the SwissModel suite of tools receives only a few sentences; Modeller (http://guitar.rockefellar.edu/modeller), MoDBase (http://pipe.rockefeller.edu/modbase) and threading tools are absent. The book stops short of describing genome analysis (for example, MAGPIE, http://magpie.rockefeller.edu) and tools for integrating output from those that operate at the gene level (for example, the URL for BCM Search Launcher is listed but not discussed). The chapter on gene-finding points to the need for these systems.

The book provides a broad overview of the basic tools for sequence analysis. It is a good starting point for the reader who wants to learn about the types of tools used in bioinfomatics and how to get started. For biologists approaching this subject for the first time, it will be a very useful handbook to keep on the shelf after the first reading, close to the computer.

The more inquisitive reader could follow up with *Bioinformatics: the machine learning approach* by P. Baldi and S. Brunak, which describes in a straightforward style machine learning-based methods for prediction of protein and genomic features. Other books that delve more deeply into the technical details of bioinformatics methods include *Biological sequence analysis: probabilistic models of proteins and nucleic acids* by R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, which focuses on how the sequence alignment and pattern extraction tools work; and *Algorithms on strings, trees and sequences* by D. Gusfield, which gives in depth technical detail on string and tree processing algorithms.

Terry Gaasterland is in the Laboratory of Computational Genomics, The Rockefeller University, 1230 York Avenue, New York, New York 10021, USA.
email: gaasterland@rockefeller.edu