Book Review

Problems and Solutions in Biological Sequence Analysis

Mark Borodovsky, Svetlana Ekisheva
Cambridge University Press, Cambridge, UK; 2006; ISBN: 0521612306; Paperback, 360 pp.; $37.80.

Since its first publication in 1998, *Biological Sequence Analysis* (BSA) by Durbin *et al.* has become a standard textbook for teaching bioinformatics. Part of the resources offered by BSA to understanding the underlying principles of bioinformatics is a set of workable exercises left to the interested reader to solve. Given the multidisciplinary background of bioinformatics, solving these problems requires integrating knowledge from various fields including genetics and molecular biology as well as mathematics and computer science. This is not an easy challenge for the numerous bioinformatics students who come with different abilities from a wide variety of educational backgrounds. Therefore I was glad to see this book offering not only step-by-step solutions to the problems presented in BSA but also to a large set of additional bioinformatics problems.

From the first introductory chapter dealing with basic notions of probabilities to the more technical final chapter on more complex probabilistic concepts used in bioinformatics, this book follows exactly the structure of BSA. All problems presented in BSA are meticulously solved, and so are additional problems on topics concerning pairwise and multiple sequence alignments, hidden Markov models (HMMs) and building phylogenetic trees. The solutions often contain generalizations of particular cases and interesting remarks that draw the reader's attention to specific results important for the analysis of biological sequences. Most problems have analytical solutions, but some of them are algorithmic examples. In such cases, the less inexperienced computer programmer can benefit from the Web Supplemental Materials including Perl and C++ implementations of those algorithms. As in BSA each chapter is concluded with a 'Further reading' section with more up-to-date bibliographical references illustrating theoretical advances and research directions for the topics presented.

I think that some of the most valuable parts of this book are the frequent theoretical introductions that broaden the fundamentals presented in BSA. The derivation of substitution matrices is briefly discussed in BSA. This is a critical step in scoring pairwise alignments, and I commend the authors for treating it thoroughly here. More attention is given to the PAM family in which case an ample discussion on the derivation of the amino acid substitution scores is followed by illustrative examples first in the context of a stationary Markov model, and later on in the context of an evolutionary model too. The statistical significance of pairwise alignments is addressed with exercises and theory topics describing the statistics of high-scoring pairs and the distribution of the length of the longest common word between random sequences. Theory and problems dealing with the selection of a particular probabilistic model and the estimation of its parameters expand the HMM chapter from BSA. More details are given for parameter estimation in the case of profile HMMs where a discrimination method for weighting training sequences is presented. Multiple sequence alignments are important not only to estimate the parameters of profile HMMs, but also in many other cases including phylogenetic inferences. I was glad to see the authors add three more exercises to the only one given for this topic in BSA. Probabilistic interpretations are abundant throughout the book and I found especially useful the treatment of realistic molecular evolution models that allow for insertions and deletions as well as variable rates at different sequence sites.

This book is an excellent companion of the BSA book. Its mathematics is not simple but it would certainly help in developing the mathematical ability of the students studying BSA. By introducing detailed theoretical fundamentals this

book extends BSA's domain and goes beyond a simple collection of bioinformatics problems with accompanying solutions. I recommend it to anyone interested in a more in–depth understanding of the bioinformatics of DNA and protein sequence analysis.

Mihaela Pertea
Assistant Research Faculty,
Center for Bioinformatics and Computational Biology,
University of Maryland,
College Park, MD, USA
E-mail: mpertea@umiacs.umd.edu