Book reviews

An Essential Guide to the Basic Local Alignment Search Tool: BLAST

Ian Korf, Mark Yandell and Joseph Bledell
O'Reilly Associates, Sebastopol, CA; ISBN 0 596 00299 8; 311 pp.; US$39.95 (pbk); July 2003

This book contains five parts: a one chapter introduction, three chapters of theory, five on practice, three on 'Industrial Strength BLAST', reference manuals for NCBI-BLAST and WU-BLAST, plus five appendices. The introduction walks a novice user through a preliminary search via the NCBI BLAST website and explains how to interpret the output. The theory provides some background on molecular biology, the purpose and mechanism of sequence alignment methods, and evaluating alignments and significance measures resulting from database searches. The practical section provides an in-depth discussion of the BLAST suite of programs, detailed explanation of the BLAST output file format, cursory description of alignment statistics, and tips, hints and recipes for building protocols and pipelines out of the BLAST suite of programs. The final content chapter, 'Industrial Strength BLAST', provides an in-depth explanation on how to install BLAST, curate databases and perform effective and efficient searches by applying software and hardware optimisations. The remaining sections (reference and appendices) provide reference and code materials to facilitate advanced use of BLAST.

The current generation of computational biologists enjoy the enviable position of being in such demand that they cannot fulfil the biological community's need for their services. Thus, effective training of new scientists, with the skills to leverage the sophisticated tools that have been developed for performing fast yet sensitive sequence database searches, has become imperative for computational biology to remain a growing, and useful, science. Although there are several books that provide a good introduction to the science behind bioinformatics-oriented computational biology ('Biological Sequence Analysis' by Durbin, Eddy, Mitchison and Krogh being the seminal piece), and a recent collection of scripting language-oriented practical applications texts (there are two O'Reilly books covering Perl programming for bioinformatics, plus my personal favourite, 'Genomic Programming in Perl' by Rex Dwyer), no single text provided an integrated view of the science behind alignment searching and the most popular program for doing that, BLAST. This book aims for that, and I believe it succeeds well enough to recommend that this book be essential reading material for advanced undergraduates taking an upper division bioinformatics course, and for graduate students just starting in a computationally oriented biological field of study. It may even be appropriate for advanced students and professors who need to 'bone up' on their BLAST skill set.

The initial introductory chapter walks the reader through a straightforward and elementary search on the NCBI BLAST website. A *HOX* gene from coelacanth is searched against the nr database and the resulting hits are inspected. The book does a fine job of introducing the novice user to the task of interpreting a BLAST report. Immediately following this introductory example, the book has a soft introduction to the central dogma of molecular biology, the atomic structure of

nucleic acids and proteins, and the genetic evolution. This clearly didactic sort of introduction is necessary to prepare a novice reader to make intelligent decisions when they attempt to discern biological information from their BLAST search output.

The following two chapters introduce sequence alignment and the theory behind the statistical measures used to evaluate sequence similarity, respectively. The alignment chapter gives clear, worked-out examples of both Needleman–Wunsch and Smith–Waterman, something I find lacking in some other bioinformatics texts (proper treatment of edge conditions is often completely missing). A working example Perl program that demonstrates dynamic programming can be used as a reference by students who wish to understand the mechanics of sequence alignment methods and is a welcome addition. The chapter on alignment statistics introduces the relevant concepts, without delving too deeply into some of the more obscure but important issues behind the theory.

The five practical chapters perform a very important service. Perhaps other readers who have independently learned to set up and use BLAST locally will appreciate the value of the information presented in these chapters, because it is sadly lacking (or hidden) in the standard BLAST documentation. The first chapter walks the reader through the basic BLAST programs, how the BLAST algorithm achieves its speed-up (fortunately, a short introduction to time and memory complexity earlier in the book as prepared the reader to appreciate the algorithm), and where to go to learn more details. The next few chapters perform the tedious chore of completely explaining the standard BLAST output file format, how to interpret BLAST statistics, and crucial tips for improving the speed and sensitivity in BLAST searches. The tips are all worthwhile, although only some may be applicable to any particular laboratory or researcher, depending on their interests. The book even provides a short section on 'How to lie with BLAST statistics', which is of questionable value, although it does demonstrate the importance of database size in E-value computations. The remaining textual contact section contains three chapters. The first contains all the information required to download and install BLAST and sequence databases locally. The next provides critical advice on how to maintain BLAST databases – as the authors say, 'one of the most neglected yet important aspects of using BLAST'.

The remaining sections, BLAST Reference, and the Appendices, have the feel of 'needing to fill the book out a little'. The reference contains little more than the manual pages for the NCBI-BLAST and WU-BLAST programs, while the appendices contain information that probably belonged directly in the content of the book: description of the various NCBI-BLAST sequence alignment output formats, tables of values computed on the similarity matrices used for nucleotide and protein scoring schemes, and source code for a couple of Perl utility scripts.

In all I heartily recommend this book be read by anybody who does not know BLAST but has a need to use it. It provides a firm introduction to both the basic molecular biology and computer science underlying BLAST, and does so in a didactic and low-nonsense style. The 'meat' of this book is in its clear explanations of knowledge that previously had to be learnt from other BLAST practitioners or gleaned from extensive searching of old mailing lists and obscure documentation pages.

David Konerding