Book reviews

Bioinformatics, Biocomputing and Perl: An Introduction to Bioinformatics Computing Skills and Practice

Michael Moorhouse and Paul Barry

John Wiley and Sons Ltd; ISBN 0 470 85331 X; 506 pp.; 2004

Bioinformatics is a tool biologists need to make sense of new types of methods involving large quantities of data. The thrust to develop bioinformatics comes from biologists, not from computer scientists. Take away bioinformatics and computer scientists will start looking elsewhere for ways of applying and developing string or graph theory, or whatever they are working on at the time. The biologists, on the other hand, would be in deep trouble. It is therefore surprising that so few biologist are fluent in bioinformatics.

Bioinformatics ought to be taught as part of the curriculum to all biologists exactly in the same way as statistics, biochemistry, anatomy or population genetics. There will be some who will hate it and do their utmost to find a branch of biology to hide from it (it will be hard), but a large majority will decide to make the effort and will find it useful. Then, of course, there are those who cannot have enough of it – and they will be joining the ranks of us bioinformaticians.

The new book by Michael Moorhouse and Paul Barry is one of the better answers to this common problem of what to teach to biologists about bioinformatics. The premise of the book is that to do bioinformatics you have to learn to program. In that, it differs from – and builds on – the hugely successful book by Jean-Michael Claverie and Cedric Notredam 'Bioinformatics for Dummies', which uses existing programs and web sites to do all the work. This is of course a useful approach in the beginning as makes it easier to start learning the methods and concepts of bioinformatics, but nevertheless, if you want to start using bioinformatics to something new, you will just have to start programming.

The programming in this context does not have to be anything exceedingly complicated. A dabbling student of bioinformatics needs mostly to automate the management of textual data, which includes storing and displaying it without graphical user interfaces. In 'Bioinformatics, Biocomputing and Perl' the tool is perl, the programming language most closely resembling natural ones in its messiness and unabashed use of shortcuts. Although no one uses perl to teach programming to computer scientists, perl is the most commonly used language in bioinformatics. This must have something to do with the historical nature of biological sciences, which leads to proliferation of exceptions at the expense of rules.

Moorhouse and Barry teach the basics of perl before moving to more biologically oriented topics. In that they differ from older books by James Tisdall ('Beginning Perl for Bioinformatics' and 'Mastering Perl for Bioinformatics'), who gives a very thorough coverage of perl programming language using bioinformatics examples. The Moorhouse and Barry book is less exhaustive, but no less educational. The authors do not cover advanced topics of perl programming, such as objects, or the intricacies of biological sequence data management, nor do they try to give all the details of every topic discussed in the book. The subtitle of the book is 'An Introduction to Bioinformatics Computing Skills and

Practice' and they really mean it. The idea is to give the student of biology in the later stages of his/her studies enough tools to start tackling most computational problems. More details can be found in specialised books once a general understanding of the field has been gained.

Topics covered in the book, in addition to perl programming basics, are sequence and protein structure data files, using relational databases, creating web pages from programs, creating graphics, and installing and using programs locally.

The authors' experience in teaching shows clearly throughout the book. The material must have been developed and polished in running numerous courses using it. It is clearly the aim of the authors that their book should be used as course material by others. This is made even easier by providing lecture slides freely on the book's web site.

The book is laid out in clearly structured textbook style. Chapters are punctuated by maxims that stress the important point just covered. These maxims are then repeated at the end of the chapter together with pointers to further reading and exercises. The tone of the text is friendly but to the point. The use of excessive wit, which might have felt refreshing a few years ago in programming language guides but grew tiresome after endless repetition, has thankfully been restricted to footnotes.

Although the authors claim that their aim is 'not to turn the scientist to a programmer', they do explain carefully not only the semantics and syntax of programming and also reasoning behind customs as well as pointing out widely recognised good programming practices. The only topics I missed from the book were the use of formal tests to monitor the program's performance during its development, and a mention of perl's Plain Old Documentation POD. They probably do not belong to the authors' personal programming style.

I found only one glaring and potentially confusing error in the book: On page 341 it says 'accession numbers between three [primary nucleotide sequence] databases are quite different...'. Since many, including bioinformatics professionals, have had a hard time dealing with all the different IDs in the EMBL-Bank, GenBank and DDBJ sequence databases, it is best to clarify this here. All sequence databases have their own local identifiers, typically on the first line. The accession numbers are the identifiers that are shared between databases and are guaranteed to be stable. They are especially useful in their full format with version number added (eg L07488.2), which uniquely identifies the sequence. There really should be a maxim in the book: 'Always use accession numbers in preference to all other sequence identifiers. If you know the version number, you are future proof.'

My other minor gripes about the book are that the only example of shell scripting (p. 393) is in tcsh, which is really not recommendable in any kind of serious scripting, and not in bash, which is the natural sh -compatible shell of Linux. Chapter 13.3 wastes half a page exploring different ways of finding if a perl module is installed on the system when a simple onliner 'perl -e "use DBI"' should have done the trick – as already shown on p. 150. The last chapter on Bioperl gets the syntax wrong on how to view online documentation (p. 451); the correct syntax is 'perldoc Bio::SearchI...'.

The above mistakes are really few and far between in this clear and well-edited book. They also show that the focus is on using Linux and open source tools. The examples can be run on other operating systems but the primary target platform is Linux, paving the way to using large scale UNIX-based servers.

Books about bioinformatics have proliferated during the past couple of years. Most try to explain what bioinformatics is, some deal with specific areas within bioinformatics, but very few so far have tried to take bioinformatics to budding biologists. Here we have a book that makes it easy.

Heikki Lehväsaiho
E-mail: heikki@ebi.ac.uk
The author is a Staff Scientist at the European Bioinformatics Institute and a core developer of Bioperl

An Introduction to Mathematical Methods in Bioinformatics

Alexander Isaev
Springer Verlag, New York; ISBN 3 540 21973 0; 294 pp.; €49.95; 2004

When the draft of the human genetic sequence was published in 2001, articles in the popular press excitedly proclaimed discovery of the 'Book of Life'. The authors of one of the academic articles announcing this feat were more sober, writing 'in principle, the string of genetic bits holds long-sought secrets of human development, physiology, and medicine. In practice, our ability to transform such information into understanding remains woefully inadequate' [*Nature*, Vol. 409 (2001), pp. 860–921].

Transforming data into understanding is the primary goal of informatics, indeed of all science; informatics is motivated by the modern requirement that we 'drink from the firehose'. The automated sequencing machines, the immensity of a genome, the abundance of species call for new methods of statistical inference as we seek to read that Book. The drive to bring this information to our students, to young scientists early in their careers, has inspired Professor Isaev to write this book, which surveys mathematical methods currently used to understand the working of the genome.

This book consists of 9 chapters, divided into two parts. Chapters 1–5 are devoted to sequence analysis (alignment, Markov chains and hidden Markov models, protein folding and phylogenetic reconstruction); chapters 6–9 provide mathematical background for the statistical methods used in the first part. It is based upon courses the author has presented to undergraduates at Australian National University.

This is a hugely ambitious undertaking; a young man reaching for the sky. Entire volumes have been devoted to each of these topics, which remain intensely active areas of research. Professor Isaev has done a great service to his students by introducing them to these fields while they are still being explored, while there are still many bright nuggets to be unearthed. The book is presented within Springer Verlag's Universitext series of mathematics, 'not necessarily textbooks for classroom use, concepts that are still being thought about, presenting material in a way that mathematicians can learn from.' On those terms, it is a great success.

I found the presentation uneven. Some topics (eg parameter estimation for HMMs) were explained at great length, while others (eg the first serious example, of sequence alignment) were flung down in a tangle. On the whole, the text seems to have been rushed into print without revision after contact with students. It would require a great deal of explanation, I believe, for all but the most serious of students, and a great deal of patience for all but the most enthusiastic of instructors. It is nonetheless a valuable addition to the literature, for I know of no text, even at the graduate level, which successfully achieves so ambitious a mission.

David Hart

Bioinformatics: Sequence and Genome Analysis, second edn

Edited by David Mount, University of Arizona, Tucson
CSHL Press, Woodbury, NY; 0 87969 712 1; 600 pp (approx.), illus., appendices, index; $89; August 2004