Book Review

Computational Biology: Unix/Linux, Data Processing and Programming

Röbbe Wünshiers
Springer-Verlag, Berlin, Heidelberg, Germany; 2004;
ISBN 3-540-21142-X; 284 pp.; Paperback;
UK £20.50

The aim of this book is to 'give exactly the sort of introduction into Unix, Unix-based operating systems and programming languages that will be a key competence for experimentally working molecular biologists'.

The book, written for complete beginners — not necessarily biologists, consists of four parts, and contains 12 chapters, an appendix, 19 figures and 12 tables. It is well laid out, with clear boxed examples of screen content and scripts. Ten short sets of practical exercises are included (with answers).

Part one introduces the basic components used by the book – Linux, shell programming, sed, awk and perl.

Part two gives an introduction to the history of Unix and Linux before discussing the relationship between Unix, Linux and Mac OSX, and some options for running Unix-like systems without actually installing Linux. A page is dedicated to a list of bioinformatics packages available for Linux, but its usefulness is limited, since their actual function is not covered within the book.

The real content is concentrated in parts three and four. Part three – 'Working with Unix/Linux' starts with an introduction to finding your way around a Unix system, including the basics of working with files and directories. Shells are introduced, before useful sections on using inbuilt shell commands to manipulate data and processes.

Chapter 8 introduces shell scripting, starting with input and output methods, and variables, before moving on to flow control. Examples illustrate common tasks such as selecting files for archive and removing unwanted spaces in text. Unfortunately, these are not always sufficiently rigorous for real use (eg a script to detect DNA content uses an expression which only matches lower-case letters). The last two chapters of the section cover regular expressions for pattern matching and the sed stream editor.

The contents of this part follow a logical flow, starting with the basics and moving on to more complicated concepts. Minor niggles include a few typos and the use of small pieces of example code referring to concepts before they are covered in the text.

The contents of Chapter 5 — 'Installing BLAST and ClustalW' — are less easy to fathom. Given the title of the book 'Computational Biology', it is reasonable to expect that some mention would be made of at least a few of the core programs used in bioinformatics and/or computational biology. Sadly, the book limits itself to this single chapter, together with the aforementioned list of biosciences software. The object here appears to be to use these programs as examples of how to install (and compile) other people's programs on a Linux platform, rather than illustrating their function. The author covers this at such a trivial level that it is difficult to see the benefits of including this chapter at all. Like many bioinformatics programs, BLAST uses environmental variables as a way of telling the program where to find key information. Unfortunately, these are ignored in the described installation. Instructions describe building a searchable database inside the software directory itself, at best a recipe for confusion if the reader wishes to make any real use of the program. In a similar vein, the instructions for ClustalW include compilation using a makefile. This is a very useful concept to understand but, unfortunately, it is glossed over here. Finally, the reader is instructed to run ClustalW in a way that circumvents its inbuilt menu system and builds an alignment purely on default settings. This gives a somewhat misleading view of the capabilities of the program.

Part 4 devotes 58 pages to the awk scripting language before moving on to perl for the remainder of the book. The author uses awk as an introductory programming language and illustrates a number of basic programming concepts such as reading input from different sources, variables and print statements. The later chapters on perl cherry-pick features not available within awk, such as hashes, scalar variables and complex string manipulations.

The content of the book seems somewhat skewed. The author has made a brave decision to spend a large proportion of the book teaching the awk scripting language. This would seem to move in a different direction to the majority of newer practitioners, who encourage the jump to learning perl earlier without resorting to heavy use of awk. The introduction specifically mentions the common problem of manipulating file formats required by different programs, but no section defines any of the commonly used biological file formats. This would make file conversions more straightforward and could be usefully located within the appendix. It was encouraging to see a mention of the Bioperl extensions, however brief, although sections on cgi scripting and MySQL would also have been useful.

Since the book encourages the user to attempt scripting and later programming tasks for themselves, it would be helpful to discuss the concept of layout and paths in more depth, perhaps as a specific chapter. Saving programs in consistent places, using links and bin directories are all useful concepts that can stop a machine from becoming hopelessly disorganised.

An extension of the chapter, covering installation of third-party code, would also have been useful, even though a detailed discussion of make and debugging is outside the remit of this book. This topic is seldom seen in books and would empower the reader to attempt to install some of the many excellent Open Source programs available.

Inbuilt Unix and shell commands can handle a range of data processing tasks but many bioinformatics textbooks do not include them, and yet few people first venturing into computational biology from the bench would turn to a pure Unix textbook for ideas. This book makes a strong attempt to cover an often neglected area, although with only partial success. There is a clear need for a book of this type, to bridge the knowledge gap in a palatable manner. The existing book 'Developing Bioinformatics Computing Skills' (Gibas & Jambeck; O'Reilly & Associates Inc., 2001) occupies a somewhat similar niche, although it places a higher emphasis on bioinformatics analyses.

This will be a useful introductory book for researchers who are not afraid to do some exploration on their own and who want to learn the basics of moving data around efficiently before moving into programming. An enthusiastic reader will, however, soon wish to refer to other textbooks to fill some of the gaps. This book should prove to be a helpful adjunct to the many textbooks dedicated to perl, Unix and bioinformatics tools.

Sarah A. Butcher
Centre for Bioinformatics
Imperial College London
London, UK