Reviewer: Antony Unwin
Augsburg University

Bioconductor Case Studies

Anyone analyzing microarray and other modern biological data will surely have come across the **Bioconductor** project and its associated packages. This work has extended R substantially and is an important tool for research. However, it is all still fairly new, so people have to teach themselves, or find someone who can give them a course on it. The book under review is basically the notes for such a course, given by some of the main people involved in the development of **Bioconductor**. Each chapter covers a different task that you might have to carry out as part of your analysis, providing sample code to show you how it can be done and lots of exercises for the reader to try it out for themselves. Answers to the exercises are provided at the end of the book. All the code, including solutions to the exercises, is available for downloading on the Web and—and this is well worth mentioning—it all runs straight out of the box. (It's a pity the URL given in the book is incorrect and one or two packages were initially tricky to find. On the other hand you'd think someone with my experience would know that it is simpler and more reliable to use Google than to depend on printed URLs.)

As is to be expected from notes from a course where chapters are written by different groups of authors, the material is not always fully consistent and there is a certain amount of repetition. The same code for reading and organizing the common dataset used appears, with a few slight variations, some nine times. This is a small price to pay for the benefits of concentrating on a common real dataset, though it is not really in the spirit of the book, which is more about providing good code to handle repetitive tasks.

Needless to say, this book does not cover any theory in depth, that is not its aim. The authors reasonably assume that their readers should either have a solid statistical grounding already or be prepared to study it elsewhere. The book describes various analyses, provides the code for them and discusses the output. This makes for an easy read and anyone who works through the book will gain confidence that they can carry out analyses on their own data.

The discussion of analyses is generally sound and practical. In particular the interpretation of the results of clustering is more sensible than you often see. The graphics are mostly

fine, but there are some exceptions: Figure 2.2 of boxplots by groups uses a colour shading for no particular reason and fails to mention that the group sizes range from 2 to 27463; Figure 3.1 is a QC plot with too much unreadable detail; Figure 3.12 is overlabelled; The histogram in Figure 6.1 does not show the high values in the data (though they show up in the histogram I was able to draw with the R code from the Web page); Figures 6.3 and 8.2 both plot scatterplots of values against index numbers for two groups, when side by side boxplots would permit better comparisons; Figure 10.8 is a messily overlabelled biplot; many scatterplots use the R default of open circles, though the ones using just points look better and are easier to decipher; Figure 15.4 should be a histogram, not a barplot. None of these are major criticisms and the graphics in other statistics books are doubtless not much better, it is just disappointing that graphics standards are not generally higher.

Is there anything the authors might add to improve the book? Apart from brushing up the graphics, it would be useful to add an index of which commands were used where, and the code on the Web should be referenced by book section rather than just listed in an order of "chunks".

This book is strongly recommended for learning more about **Bioconductor**.

Reviewer:

Antony Unwin
Department of Computer-Oriented Statistics and Data Analysis
Augsburg University
DE-86135 Augsburg, Germany
E-mail: unwin@math.uni-augsburg.de
URL: http://stats.math.uni-augsburg.de/