Journal of Statistical Software

Reviewer: Jan de Leeuw
University of California at Los Angeles

R Programming for Bioinformatics

Robert Gentleman
Chapman & Hall/CRC, Boca Raton, Florida, 2009.
ISBN 978-1-4200-6367-7. 314 pp. USD 69.95 (P).
http://www.bioconductor.org/pub/RBioinf/

This is a very excellent book. As far as I can see there is only one single thing wrong with it, and that is its title. The book is about R programming for *Absolutely Everybody*. The bioinformatics component is small, and not at all essential to the rest of the book. It would be unfortunate if the title did mislead some potential buyers and readers, because I think this is actually *the best handbook* on R programming that is currently available. *Fundamentals of R Programming* or *R Programming for Scientists* would have been a better title.

Gentleman is eminently qualified to write about R, since he is one of the creators of the original R system (Ihaka and Gentleman 1996). It is very important to have a programming text by someone who is intimately familiar with the nooks and crannies of the system, because he has has been involved in its development from the start. In addition, Gentleman is one of the key players in the **Bioconductor** project (http://www.bioconductor.org/, Gentleman *et al.* 2004) which undoubtedly explains the otherwise mysterious emphasis on bioinformatics in the title. As the old Dutch saying goes: Whose bread I eat, whose song I sing.

Chapter 1 is a short four-page introduction. Chapter 2 (62 pages) is *R Language Fundamentals.* This is mostly standard material, also to be found in books on using R, although additional attention is paid to replacement functions, exception handling, evaluation, and scope. Chapter 3, *Object-Oriented Programming in R* (53 pages), is an excellent and authoritive discussion of both S3 and S4 object-oriented programming (OOP) and their interaction. Chapter 4, *Input and Output in R* (27 pages), is mostly standard material, with some information on file formats and connections. Chapter 5, *Working with Character Data* (39 pages), has very good information on formatting, parsing, regular expressions, matching, and sorting. This is material one does not find in the usual "Statistics with R" books. I very much liked Chapter 6, *Foreign Language Interfaces* (29 pages). It has up-to-date information on the C and Fortran language interfaces and the R application programming interface. Chapter 7, *R Packages* (19 pages) covers the package management system. Chapter 8, *Data Technologies* (45 pages), discusses database technologies and interfaces, and the use of XML. It shows some interesting applications, mostly from bioinformatics. Finally Chapter 9, *Debugging and Profiling* (28 pages), discusses debugging by browsing and tracing and profiling both computation

and memory management. Taken together, the nine chapters provide an indispensable handbook for R programmers, and an excellent textbook for a graduate course in R programming.

Let's make a brief comparison with the competition. There are two other books on R programming, both also a must-have for all R programmers. Venables and Ripley (2000) is still very useful, but it is beginning to show its age. In particular, it was written at the time that S-PLUS was still a serious alternative to R, and consequently it had to jump from one dialect to another all the time, which impedes its flow. The book by Chambers (2008), reviewed in this journal by Ellis (2009), is a gourmet dinner, but it is not a textbook. It is perfect for dipping into many times, and for candlelight reading. I very much like the comparison Chambers makes with Jacques Pépin's *La Technique*. Gentleman's book is much more like *Mastering the Art of French Cooking*. Chambers has at least as much information, for example, on OOP and foreign language interfaces, but the information is presented very differently, in a more conceptual, literary, and leisurely way. You definitely need them both.

References

Chambers JM (2008). *Software for Data Analysis*. Springer-Verlag, New York.

Ellis B (2009). "Review of 'Software for Data Analysis'." *Journal of Statistical Software, Book Reviews*, **29**(4), 1–2. URL http://www.jstatsoft.org/v29/b04.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J (2004). "**Bioconductor**: Open Software Development for Computational Biology and Bioinformatics." *Genome Biology*, **5**, R80. URL http://genomebiology.com/2004/5/10/R80.

Ihaka R, Gentleman R (1996). "R: A Language for Data Analysis and Graphics." *Journal of Computational and Graphical Statistics*, **5**, 299–314.

Venables WN, Ripley BD (2000). *S Programming*. Springer-Verlag, New York.

Reviewer:

Jan de Leeuw
University of California at Los Angeles
Department of Statistics
Los Angeles, CA 90095-1554, United States of America
E-mail: deleeuw@stat.ucla.edu
URL: http://gifi.stat.ucla.edu/