Journal of Statistical Software

April 2005, Volume 14, Book Review 1. http://www.jstatsoft.org/

Reviewer: Thomas Boyle
University of Washington

The Analysis of Gene Expression Data: Methods and Software

Giovanni Parmigiani, Elizabeth S. Garrett, Rafael A. Irizarry, Scott L. Zeger (eds.)
Springer, New York, 2003.
ISBN 0-387-95577-1. 455 pp. $95.00.
http://astor.som.jhmi.edu/hex/pgiz.html

Almost a decade has passed since its introduction and use of microarrays for studying gene expression seems to be flourishing. Since it is a statistical and software intensive area, this book has a scope well adapted to the field. The opening chapter contributed by the editors is an excellent overview of the field and the threads that make it up. All the contributed chapters receive honorable mention somewhere therein. However, the review does not establish a true structure for the content of the book.

There are hints in the preface: "...goal is to provide guidance to practitioners in deciding which statistical approaches and packages ... and correctly interpreting the results."; "... most packages are directed at microarray data analysts with master's level training in [quantitative fields]"; "A minority of more advanced chapters are intended for doctoral students and researchers." Of the 18 contributed chapters, this reviewer would break things down as: 8 chapters including the opening chapter are well aligned with the goal of helping practitioners decide; 3 chapters move into the "Master's level" category, and 7 seem more oriented toward advanced researchers. This is a satisfactory breakdown, but when it comes to deciding on methods and software, the presentation would have been improved by including a chapter that compared and contrasted the several approaches presented. The decision process one is left with having been exposed to somewhere between 7 and 18 packages is still a daunting one.

There is a definite hands-on character to all the presentations. Generally, each contributer selected a different but interesting set of data to demonstrate the approach. Six chapters describe packages written in the R programming language and use its command line format to present the examples. This might present a problem for readers not familiar with that package – or maybe just the incentive they need to get on board. Three of the packages are web applications. The remainder are a mix of JAVA, windows executables, etc.

The only comprehensive single package described is **MAExplorer**. This is JAVA based with a genuine open source license. A package with similar scope from TIGR is described in the

opening chapter but does not have a contributed chapter devoted to it. Of course, the full set of R packages are comprehensive in aggregate but many practioners might not want to provide the glue that is needed to pull it together and if they are predominently biologists, there would be a user interface barrier way higher than that provided by a comprehensive package. Similarly, there is a user barrier presented by approaching the analysis as the sequential application of a set of disparate applications.

The **SNOMAD** web application is intriguing to contemplate since it professes to be a set of R applications wrapped by Perl CGI scripts into a web application. The range of the program is limited to the early steps in the analysis pipeline but the chapter does an outstanding job of describing these. One would like to see the web interface offered for local installation especially if it is readily extensible so as to include the full range of the R packages available.

The chapter on **DRAGON** presents an aggressive approach to integrating experimental work with the growing databases and literature on genes. This is a web based application which accesses a **MySQL** database. The database structure is clearly presented. The whole appears to be a powerful tool for accessing annotation information on a gene list found using other tools described in the book.

Topics covered in the advanced chapters include: bayesian methods, SAM false discovery rate control, identifying significant genes from the set with low expression, and relevance networks. The authors have provided in a compact package lots of food for thought on this important area.

Reviewer:

Thomas Boyle
University of Washington
Department of Genome Sciences
Seattle, WA 98185
E-mail: biowolp@u.washington.edu