

Book Review

Statistical Methods in Bioinformatics: An Introduction. Warren J. Ewens and Gregory R. Grant (2005), Second Edition, pp. 596, Springer Science + Business Media, New York. ISBN 0-387-40082-6

The first edition of this book (2001) has been widely adopted as introductory textbook on probability and statistics in bioinformatics courses all around the world. This has led to greater awareness among computer scientists and biologists that statistics – as the science of data analysis – has a lot to offer for bioinformatics. In particular, statistics provides many versatile and systematic tools for the efficient study of the processes governing genomic data, and it also offers a framework for assessing the many ad-hoc algorithmic procedures common in bioinformatics.

In comparison with the first edition of Ewens' and Grant's popular book the second edition has been expanded from 476 pages to 596 pages. This increase of nearly 25 percent in volume is mainly owed to two major changes. First, a completely new chapter (#13) about the analysis of gene expression data has been added, focusing on methods for detecting differential expression and on the problem of large-scale multiple testing. Second, there is now a very strong emphasis on statistical methods, rather than solely on probability theory. Consequently, all chapters dealing with statistical inference have been reorganized and greatly extended. As a result, the old chapter (#8) "Classical estimation and hypothesis testing" is now split in the two separate parts (#8) "Classical estimation theory" and (#9) "Classical hypothesis testing theory" – the latter now also includes a brief introduction into ANOVA and multivariate testing. The old chapter (#12) "Computationally intensive method" has been completely removed and all its content was merged into other chapters. Furthermore, the chapter "Probability theory (ii): many random variables" (#2) has been completely restructured.

With this quite substantial overhaul Ewens and Grant set out to address some of the criticism that has been voiced concerning the first edition. Specifically, a perceived shortcoming of the first edition was that it mainly dealt with probability models and their application to sequence analysis, and that it neglected inference procedures and analysis of non-character genomic data.

This second edition aims to rectify this. On the one hand, the inclusion of the additional chapter describing statistical approaches for analyzing microarray experiments considerably broadens the range of applications discussed in the book, and brings it (nearly) up-to-date with currently employed methods in expression analysis. On the other hand, the "learning from genetic data" aspect of bioinformatics is now much better accounted for by the extended discussion of estimation and hypothesis testing.

While these efforts are very laudable, it is still easy to identify quite a few "white spots" on Ewens' and Grant's map of statistical bioinformatics. For instance, given its widespread use the basic principles of MCMC should have been explained, perhaps in a (sub)chapter on sampling from distributions. Furthermore, the book contains very little about regression and classification. With a single additional chapter some basic methodology could have been covered, including regularization and model selection. All this would have naturally provided the basis for 1–2 further chapters on important current application such as on population genetics (SNP models, HAPMAP project) and on multivariate analysis (i.e. clustering and classification and perhaps graphical modeling) of transcriptome and proteomics data.

It is a pity that Ewens and Grant stopped short of extending the scope of the textbook to those areas of current bioinformatics research where statistics is now of utmost importance (i.e. functional genomics and population genetics). Perhaps this will be on offer in a third edition? Meanwhile, despite the outlined deficiencies, the second edition remains a reliable and highly informative resource that definitely belongs on the shelves of every aspiring bioinformatician.

Finally, it is also noted that Ewens' and Grant's book is perhaps not the ideal starting place for statisticians and other researchers with a modeling background to look for problems in bioinformatics – this purpose will be served much more adequately by the "Handbook of Statistical Genetics" (Balding et al., 2003).

References

Balding, D. J., Bishop, M., and Cannings, C., Eds. (2003). *Handbook of Statistical Genetics*, Second Edition. John Wiley & Sons.

Korbinian Strimmer*
Department of Statistics
University of Munich
Ludwigstr. 33
D-80539 Munich, Germany

* e-mail: korbinian.strimmer@lmu.de, Phone: +49 89 2180-3225