Genet. Res., Camb. (2003), **82**, *pp.* 151–153.
Printed in the United Kingdom

Book Reviews

DOI: 10.1017/S0016672303216438

Microarray Gene Expression Data Analysis: A Beginners Guide. H. C. Causton, J. Quackenbush and A. Brazma. Blackwell Publishing. 2003. 160 pages. ISBN 1405106824. Price £34.99 (paperback).

In the 1930's and 40's the development of a sound statistical framework for dealing with genetic analysis, particularly the large quantities of data collected from agricultural studies, laid the foundations for our current understanding of population genetics. The subsequent rise of molecular biology, with its reductionist focus on single genes and how they work at the molecular level, produced a generation of biologists (myself included!) who are not quite at home with the application of the statistical methods for analyzing variance or organizing very large datasets. This is unfortunate, since the current revolution in experimental approaches to 'whole-genome' biology, exemplified by the use of DNA microarrays, relies on the appropriate use of good statistical methods. Indeed, many of the approaches, such as the analysis of variance methods, developed for exploring complex agricultural data sets are directly applicable to genome-scale data; the populations being analyzed are now mRNAs instead of wheat fields. A few months ago I reviewed a volume that covered the application and uses of DNA microarrays; at the time I lamented the lack of any serious treatment of the issues relating to experimental design and data analysis in the book, advising budding experimentalists to make friends with a statistician. This new book aims to assist the mathematically naive in understanding how and why particular statistical tests are used as well as describing some of the common approaches to experimental design and data analysis. It also serves to introduce the increasing number of statisticians and mathematicians, who are becoming attracted by complex biological systems, to the issues surrounding the analysis of large-scale biological data.

Written by well-regarded figures in the microarray field, the book is divided into three major sections dealing with experimental design, data processing and meta-analysis. These are sandwiched between a brief introduction and a useful appendix listing recommended sources of public domain software, a far superior source of analytical tools than some of the ludicrously priced commercial offerings. The authors do a first class job of discussing each of the issues and cover most of the up to date approaches currently employed. With the number of papers published on 'novel methods for microarray data analysis' appearing to increase exponentially, this is no mean feat!

The chapter on experimental design is perhaps the most basic, covering a variety of issues ranging from array design to the experimental framework that should be considered when asking a particular biological question. The types and uses of controls and the use of replicates are discussed along with the important issue of how to deal with biological variability. Perhaps this chapter could have been improved by including some detailed examples of good experimental design from the published literature to help illustrate the points being made, but this is only a minor complaint. Armed with the background in this chapter there is no real excuse for a poorly designed microarray experiment.

One of the most intensive areas of research in the microarray field today is in data extraction and normalization. How can we relate the information in a computer file containing a collection of pixel intensities to the expression of a gene in a particular sample? Furthermore, how can we then process the often very noisy data such that we can make comparisons between different experiments to illuminate the biology under investigation? The chapter on data processing addresses these issues with a detailed discussion of the issues surrounding spot quantitation and data normalization. The chapter is fairly comprehensive, though disappointingly does not cover the more recent developments in variance stabilization normalization, a robust method for data analysis becoming increasingly popular due to its ability to deal equally with experimental variance across a wide range of spot intensities. Nevertheless, the chapter does a good job of describing the issues surrounding normalization and should provide a good platform for understanding the complexities of this vital step in the data processing pathway.

The fourth and largest chapter deals with data mining and the use of statistical methods to organize and explore gene expression data. Covering virtually all aspects of supervised and unsupervised analysis methods in use today, this chapter gently introduces the reader to the mathematical rational behind clustering and component analysis. This is not a trivial task, the principles underlying the mathematical treatment of large datasets and their relationships are not simple, and the authors do a commendable job of familiarizing biologists with the analytical tools that are most commonly encountered in the software packages they will undoubtedly use.

Overall this is an excellent book, it is well referenced and, to my mind, covers the vast majority of issues an experimenter needs to consider when venturing into the world of microarray data analysis. The book fills a clear gap in the field, providing a rigorous overview of the often confusing (for me at least) data analysis issues that most books on microarrays avoid or treat in a cursory way. I would say it is essential reading for any laboratory or researcher active in this rapidly evolving field and is recommended for the mathematician or statistician who is interested in the field or who has been persuaded by their biologist colleague to help them with their analysis. I urge the authors, however much they may dread the prospect, to consider a second edition in a year or two. The field is moving rapidly and while this volume will always provide the basics, a contemporary review of the state of play can only be a good thing.

STEVEN RUSSELL
Department of Genetics
University of Cambridge

DOI: 10.1017/S0016672303226434

The Origin of Species Revisited. DONALD R. FORSDYKE. McGill-Queen's University Press. 2001. 275 pages. ISBN 0-7735-2259-X. Price £37.95 (hardback).

This book has two purposes. One is to portray the largely forgotten ideas of George Romanes on 'physiological selection' as a major contribution to our understanding of speciation. The other is to present Donald Forsdyke's own ideas on variety of evolutionary topics, with an emphasis on speciation, but also including such questions as the evolution of dominance and dosage compensation. The connection between the two is that Forsdyke believes that his ideas on speciation are a molecular version of Romanes'. He also believes that he has hit on the true explanation of speciation, which has eluded the community of evolutionary biologists. Judging from the barbed comments scattered throughout the book, Forsdyke clearly has little respect for this community, especially theoretical population geneticists. Fisher, Haldane and Wright are dismissed with the comment 'their approach, and that of their followers, was largely genetical, with mathematical and rhetorical overtones which sometimes tended to obscure rather than enlighten' (p. 89).

Needless to say, this belief is likely to be viewed with scepticism by mainstream evolutionists, especially since there is a vast literature on the evolutionary genetics of speciation which is barely mentioned by Forsdyke (there is only one reference to Dobzhansky and two to Mayr, as opposed to 32 to Forsdyke), which has led to very different conclusions from his. After reading this book, I am still not entirely clear what Romanes had in mind by 'physiological selection', but it seems to mean much the same as reproductive isolation. The quotations from Romanes on pp. 52–53 indicate that he regarded this as a much more important evolutionary principle than natural selection, which no doubt explains the hostility towards his ideas displayed by Huxley and Wallace. It is also quite unclear from these quotations whether Romanes viewed reproductive isolation as a by-product of evolutionary divergence in allopatry, or whether he envisaged some kind of sympatric speciation. If Forsdyke's account of Romanes' ideas is accurate, it is difficult to disagree with the remarks of Ernst Mayr (cited on p. 214) that 'Romanes … made no clear separation of geographical and reproductive isolation … and often dealt with speciation as if it was the same as natural selection.'

What about Forsdyke's own ideas? He focuses on hybrid sterility as the key problem in speciation. This in itself seems unfortunate, since there are many different types of isolating barriers which can separate good species. Indeed, comparative work by Coyne and Orr (1997) shows that behavioural isolation in *Drosophila* often evolves more quickly than postzygotic isolation, and that full sterility of both male and female hybrids comes relatively late. Furthermore, as has been pointed out by Kliman *et al.* (2001), the mechanism proposed by Forsdyke is in contradiction to a large body of data (none of which is mentioned in his book). His idea is that hybrid sterility arises as a result of evolutionary divergence in GC content of silent and synonymous DNA sequences. This is claimed to lead to sterility of hybrids between parents whose GC content has sufficiently diverged, although the precise mechanism involved is never spelt out. No evidence is presented that such divergence is indeed causally involved in hybrid sterility, and genetic studies that point to individual genes contributing to reproductive isolation are ignored. Studies of codon usage show that the mean GC contents at synonymous sites are often nearly indistinguishable between related species (Kliman *et al.*, 2001). This difficulty is recognised by Forsdyke, who suggests that GC content first diverges and then converges, thereby rendering his theory untestable.

In fact, application of a little mathematics shows than random drift or selection are likely to produce only a very slow change in the GC content of a genome; the proportion of the genome that is GC has a denominator of the order of the genome size, unless the states of different sites are highly correlated. The variance in GC content between individuals in a population is thus of the order of the variance in GC content at an individual site, divided by the genome size. This is clearly an almost vanishingly small quantity; since both drift and selection operate at speeds that are limited by the within-population variance (as was well-known to the despised trio of Fisher, Haldane and Wright), it will take many tens or hundreds of millions of generations for GC content to be significantly changed in evolution. This is inconsistent with the relatively recent dates of divergence of many closely related species. Forsdyke's theory is thus not just untestable, it is unworkable.

BRIAN CHARLESWORTH
Institute of Cell, Animal and Population Biology
The University of Edinburgh

References

Coyne, J. A. & Orr, H. A. (1997). "Patterns of speciation in Drosophila" revisited. *Evolution* **51**, 295–303.

Kliman, R. M., Rogers, B. T. & Noor, M. A. F. (2001). Differences in (G+C) content between species: a commentary on Forsdyke's "chromosomal viewpoint" of speciation. *Journal of Theoretical Biology* **209**, 131–140.