Anyone with an interest in discrete, combinatorial mathematics will enjoy reading this book.

References


Relatively simple sequencing techniques first came into practical use in the mid 1980s, with PCR-based technology becoming a basic research tool by 1989. I well remember that one of my student office mates was working on bacterial DNA sequencing for his Ph.D. in the early 80s. At the end of his time, he repeated the entire previous 3 years of laborious laboratory work in 3 weeks with the new sequencing technology. He also learned computer programming during his Ph.D. tenure, so that he could write a program that folded the translated protein sequence into its 3-dimensional configuration. He was thus at the forefront of both molecular genetics and bioinformatics, whereas I was merely a very traditional ecology student.

This means that I watched modern molecular genetics and bioinformatics arrive; and it seems to me from studying this that bioinformatics has always had one very unique property, which I had learned as an ecologist, because much of the mathematical focus was then on how little data biologists needed in order to answer experimental questions, not on how much. The asymptotic properties of statistics (i.e. as sample size approaches infinity) were of theoretical interest but of little practical relevance to most experimentalists.

All of that changed with the molecular revolution. Phenotypic information gave way to genotypic information, with the genic information first being gene sequences, then multiple gene sequences, and finally genome sequences. At each step of this inexorable trend, the volume of data increased dramatically, ostensibly providing us with more information to answer experimental questions. Sheer volume of data, we were told, and still are being told, would finally reveal the truth. Nothing could make a scientist more starry eyed.

Sadly, the bioinformaticians seem to have forgotten the hard-won experiences of the biostatisticians. Most importantly, the Old Timers had worked out the difference between random error and systematic error, also known as stochastic variation and bias, respectively. Increasing amounts of data are, indeed, expected to be a valid means of dealing with random error. For example:

The massive amount of genomic sequence data that is now available for analyzing evolutionary relationships among...placental mammals reduces the stochastic error in phylogenetic analyses to virtually zero. One would expect that this would make it possible to finally resolve controversial branches in the placental mammalian tree (Hallström and Janke 2010, p. 2804).

Mind you, there is an exponential relationship between sample size and precision, so that doubling the precision of an estimated quantity requires the sample size to be squared, which leads to rapidly decreasing return for effort. Any sample size beyond \(n=30\) is, for practical purposes, little different from \(n=\infty\). Nevertheless, random error can be dealt with by brute force.

The same cannot be said for bias. By its very nature, systematic error remains unchanged in response to data volume. A biased small sample will say exactly the same thing as a biased large sample, if the source of bias is the same. Only thoughtful experimental design can deal with systematic error—bias must be actively avoided in the data-collection protocol. Systematic error can also be dealt with, to some extent, in the data-analysis protocol, but only if the precise nature of the bias is known a priori, which is rarely the case. Quantity is not, in itself, a problem, but it is also not a panacea; and it cannot be used as a substitute for quality in the computational study of molecular evolution.

This simple truth has not gone unnoticed in the recent flood of genomic analyses. Indeed, it has finally reached the attention of the high-profile journals, who have started to bemoan what is being recognized as poor-quality data analyses in genomic studies (e.g. Check Hayden 2012; Editorial 2012a, 2012b; MacArthur
2012). Sheer volume of data is apparently revealing the truth of bioinformatic ignorance rather than any larger truth. Overzealous data mining is seen to have replaced carefully performed experimental analyses, and bioinformatics is in danger of disappearing under the weight of the very data deluge that the practitioners seem to be so keen on.

Perhaps the ultimate problem is 2-fold. Biologists, who usually know about experimental quality, would rather cuddle koalas than learn much about mathematics; and the computational scientists know little about experimental quality and its application to data analyses. Thus, neither group can help the other, and they have fallen back on the forlorn hope that quality will take care of quality.

The 2 large volumes being reviewed here are intended, in some sense, to address this dual issue. They introduce nonbiologists to the world of genome research and introduce biologists to recent developments in the statistical methodology of genome analysis. Unfortunately, if quality is our criterion, then these volumes do the former job much better than they do the latter one.

One certainly cannot criticize this work on the grounds of quantity: there are 38 chapters arranged into 8 sections, written by a total of 86 people, and summing to 1033 pages of information. Even the price displays a considerable interest in quantity. As a compendium of current information, the volumes are unsurpassed. The authors have taken the task of introducing their subjects seriously, and there is barely a dud chapter among them.

The chapters are presented in a consistent mini-review format, often with links to further online resources. The literature citations cover many decades, thus providing a balanced view of each topic rather than a simple summary of current fashions. The coverage of topics is comprehensive, and up-to-date to some time in early 2011. As with any collaborative work, there is little obvious coherence to the collection of chapters, although the general trend is from introductory topics, through the collection of data, then through genome evolution, phylogenomics, populations, and finally on to data handling. The focus is rarely biological after the first few chapters, with the issues instead being seen as mostly computational and technical. This defines the intended audience, which is not really someone who wants to know about the biology of genomes—the emphasis is on the “informatics” rather than the “bio.”

The presentation varies somewhat between the chapters, most notably in the reference formatting, and some of the authors seem to be writing in Computereese rather than in plain English. The production quality is good, and there is the occasional use of color.

Most importantly, however, the issue of quality is unfortunately overlooked by far too many of the authors. For example, the critical commentary of MacArthur (2012) notes:

> the genomics community must take responsibility for establishing standards for the generation, quality control and statistical analysis of high-throughput data generated using new genomic technologies (a model that has generally worked well, for instance, in genome-wide association studies) (p. 429).

This assertion is reflected accurately in these volumes, because the only authors to directly address the quality of their data analyses are Besenbacher et al. in their chapter on genome-wide association mapping (Volume 2, Chapter 11). For the other authors, quality is either barely mentioned or simply assumed. The authors tell us about analyses, models, algorithms, protocols and databases, but not much about quality control with respect to any of these topics. We are usually told which programs exist (e.g. "BioNode has over 200 bioinformatics and statistical software packages, of interest to evolutionary biology" p. 529 of Volume 2), and presumably someone is using at least some of them, but whether any of them are any good for any specified purpose remains entirely unclear.

Of immediate interest to the readers of Systematic Biology is also the issue of whether there is any phylogenetic context to current genome studies. The answer presented by this work is: “not much, outside of bacteriology.” The genomes sequenced to date have by and large been model organisms, whose basic phylogenetic relationships are already known. These 2 volumes are about abundance of characters not abundance of taxa. As such, biological systematics plays almost no role in any of the chapters.

Moreover, in phylogenetics, it is yet to be shown that genomic data tell us much that we did not know before. The estimated relationships among humans, chimpanzees, and gorillas did not change as a result of genome sampling (Galtier and Daubin 2008), for example, nor did those of malaria species (Kuo et al. 2008) nor those of mammal superorders (Hallström and Janke 2010). In all 3 cases, the relationships were just as complex after the genome sequencing as before—perhaps the resolution of controversial branches in our trees has not occurred as a result of increased access to character data. In this sense, a small sample of representative gene sequences should reveal just as much of the genealogical truth as will a genome-wide sample. Indeed, elementary sampling theory tells us that a small sample has a number of advantages over a large one (e.g. Cochran 1977), including time and money savings.

Intriguingly, there is almost nothing about phylogenetic networks anywhere in the 2 volumes. This has become a controversial point, particularly within bacteriology, as the evolutionary history revealed by between-gene evolutionary processes (e.g. recombination, hybridization and horizontal gene transfer) often conflicts with that from within-gene processes (e.g. nucleotide substitutions and indels). Indeed, this issue resides at the heart of the 3 examples of complex relationships cited above—the more we learn about genomes, the less tree-like does evolutionary
history seem to be. An up-to-date coverage of genomics must address this issue of network-generating processes, and its absence from these volumes (other than recombination) is an obvious limitation.

In summing up, these 2 volumes are in one-way impressive: they summarize a large amount of pertinent information in an accessible form. As such, they will presumably take a valued place on the bookshelf of every genome evolutionist. In another way, however, the volumes accurately reflect the current problems with genomic studies, in which quantity takes precedence over quality. Instead of being a logical development of basic statistical analysis, bioinformatics seems to have abandoned the fundamental principles of high-quality data analysis, in which carefully planned and executed analyses are valued, and instead become a slave to data quantity. Presumably, this situation will eventually sort itself out, but at the moment, to use a time-worn metaphor, the cart is definitely leading the horse.

REFERENCES


David A. Morrison, Section for Parasitology (SWEP AR), Swedish University of Agricultural Sciences, 751 89 Uppsala, Sweden; Email: David.Morrison@slu.se