BOOK REVIEW

If you wish to order or require further information regarding the titles reviewed here, please write to or telephone the BMJ Bookshop, PO Box 295, London WC1H 9JR. Tel 020 7383 6244. Fax 020 7383 6662. Payment can be made by cheque in sterling drawn on a UK bank or by credit card (Mastercard, Visa, or American Express) stating card number, expiry date, and full name. (The price and availability are occasionally subject to revision by the Publishers.)

Comparative Genomics. Editors D Sankoff, J H Nadeau. Dordrecht: Kluwer Academic Publishers, 2000. ISBN 0-7923-6583-6. (J Med Genet 2001;38:807)

So we now have the complete (well almost) sequence of the human genome, to add to those of a fruit fly, a worm, a plant, yeast, and countless bacteria. What now? No doubt there will be more complete sequences. The mouse genome should be along shortly and zebrafish should be soon after that. But what will all this sequence tell us and how can we use it to get what we want? At one level we can use sequence as we have always used sequence, to work out where genes are in the genome, where and when they are expressed, whether they might be involved in disease phenotypes, etc. Comparing sequences between different organisms, we can work out which bits of genes are more likely to be functional, which non-coding regions are likely to be functional, etc. We can also work out how organisms are related to one another. Complete genomes allow us to do the same with a larger sample size and in some cases more easily. The thesis of *Comparative Genomics*, however, is that we can also use the new genomic sequence to look at a new dimension of problems and to look at old problems in a new way.

Consider, for example, the old problem of understanding the relationships between the species. To do this using sequence data, we usually align our orthologous genes and try to derive some measure of the similarity between sequences. We then presume that two species with similar sequences are more closely related than species whose sequence is

very different from that of the former two. This method is good but isn't the last word. So, we might ask, does gene order contain extra information that might make for an alternative approach? We know in bacteria that gene order is forever changing. Surely then if we find two species that have similar gene orders (that is, just one or two rearrangements), these must be closely related. We could, in principle, do the same with gene content (that is, which genes a species has), as discussed by Gu. While the premise is simple, the practice is fraught with difficulties, such as how to construct a measure of distance for highly rearranged genomes, how to handle the differences between circular (for example, mitochondrial) and linear genomes, and how to make allowance for horizontal transfer. The analysis is most especially difficult without estimates of the relative rates and sizes of rearrangements (inversions, transpositions, duplications, etc), and is made still more difficult if these rates differ between lineages. Indeed, some go so far as to say that the enterprise is doomed. The contributors to the present volume seem rather more upbeat.

It is these sorts of worries that pervade the book and this in turn is reflected in its major emphasis on the development of computer algorithms and the necessary statistics with which to address the issues. The purely computational and statistical chapters are hard reading, often presented as series of proofs. However, as this book is one of a series on computational biology, that much might be expected. For readers of a more biological bent, the remaining chapters should prove more involving.

It is not only phylogenetics that is potentially aided by these data. There has, for example, been a long controversy over whether the human genome is the result of two rounds of tetraploidisation early in vertebrate evolution. It has been noted, for example, that there are numerous genes in humans which exist in four copies that exist in only one in fruit flies. Curiously, it now seems that such whole genome diploidisations have occurred where we were not really expecting them (in yeast and *Arabidopsis*), but the story is more ambiguous for vertebrates. Unfortunately the review chapter on the problem as applied to vertebrates is rather short and not especially penetrating. This is not atypical in that many of the 39 chapters fall short of being thorough going critical reviews.

It is, however, the newer questions that seem the most interesting, at least to this

reader. The big question is whether, when looking at genomes, we are looking at the product of natural selection or just chance. For example, are we looking at nothing more than a string of genes that could be in any order? When we see two genes linked, should we be asking why they are linked or is it all just pure accident? Should we ask why gene density is non-random, why certain genes are X or Y linked, or why some genes are replete with introns and others not?

Disappointingly, most of these questions are not addressed, but the gene order issue is touched on a few times. Early evidence from the bacterial work suggested that everything just gets jumbled, but closer analysis has shown some striking patterns. Most notably, the genes whose proteins physically interact tend to be physically linked and stay linked. In one of the better reviews, Andersson and Eriksson ask why this might be. Bork and colleagues argue that the pattern is so strong that linkage can be used to predict function. There are suggestions, reviewed by Trachtulec and Forejt, that in the mammalian genome there may well also be some gene clusters (for example, around *Notch4*) that have stayed together longer than would be expected were everything being randomly jumbled about. They make the original suggestion that these regions may be associated with areas of matrix attachment.

I should have liked to see more consideration of these sorts of issues. For example, an obvious problem concerns the differing amounts of "junk" DNA in different genomes. This issue does not receive any serious treatment while being one of the most striking differences between genomes. Furthermore, there have been phylogenetically controlled analyses of this and related issues but neither the comparative method nor the results are mentioned. For a book on comparative genomics not to discuss the advances in the methodology of comparative analysis seems a bizarre omission. Ultimately, then, on the broader issues, the book was rather disappointing and the helpful insights appear as rare nuggets. However, this perhaps reflects more the book's concern with a particular set of methods than with answers. If methods to address gene order problems are your concern, however, then this is as good a place to start as any.

LAURENCE D HURST