Heredity (2003) 91, 542–543

www.nature.com/hdy

BOOK REVIEW

Meet the parents: bioinformatics and computational biology

Frontiers in Computational Genomics

MY Galperin and EV Koonin
Caister Academic Press; 2003. pp. 346
Price £*90, hardback. ISBN 0-9542464-4-6*

Heredity (2003) **91,** 542–543. doi:10.1038/sj.hdy.6800366

Reviewed by CAM Semple

The term bioinformatics is most often used to describe the development of software to process, store and visualize sequence data. To some, it has become synonymous with the notion of computational biology, rather than a particular subdivision servicing the genomics community. This is understandable given the explosion of computational resources that accompanied the flood of genomic sequence data from the 1990s until the present day, but computational biology was an established field two decades earlier. These earlier activities were concerned with good old-fashioned hypothesis testing but without recourse to wet lab work, and encompassed molecular evolution, population genetics, quantitative genetics and ecological genetics. Of course, these kinds of computational enquiry have continued and as we enter the 'postgenomic' era, with demand subsiding for yet more genomic annotation browsers, computational biology as investigative science can only be set to grow. The hope is that much of the activity in bioinformatics can be redirected to form something approaching a theoretical branch of molecular biology that constitutes a predictive science in its own right (see Claverie, 2000). This timely book explores recent, pioneering activities in computational biology, emphasizing studies of genome evolution, where wet lab work is difficult or impossible.

The first five chapters review recent developments in perennial fields, such as gene prediction, alignment methods and the prediction of protein structure and function. In each case, the crosstalk with the study of molecular evolution is made evident. In the opening chapter, Guigo and Wiehe reprise their work comparing the accuracy of gene prediction algorithms. This is as good a review of the field as any other and discusses the recent progress made by using comparative genomic data. Identifying gene boundaries precisely and predicting the mRNA sequence they encode correctly in metazoan genomes remains a major challenge. Underlining the distance, we have to go to reach a full set of reliable gene predictions, let alone a predictive theory of how the structures of genes relate to alternative splicing and regulation. Chapter 2 sees Godzik engagingly review the development of sensitive alignment algorithms from the viewpoint of someone with a depth of knowledge about the underlying theory, as well as the practical problems in this area. He ends with a discussion of the use of structural data in remote homology detection, which is where Eisenhaber *et al* begin in Chapter 3. Here, the focus is on threading techniques; a particular set of methods for remote homology detection that can use a variety of structural data to predict the structures of uncharacterized protein sequences. Although popular several years ago, these methods fell out of favour to some extent when they failed to prove their worth in the large-scale functional annotation of protein sequences necessitated by genome sequencing projects. The in-depth statistical analysis of Eisenhaber *et al* suggests fundamental reasons for their failure. In Chapter 4, Kopp *et al* discuss comparative modelling of protein structure, where a known protein structure is used as a template to estimate the structure of an uncharacterized sequence. Given the ongoing structural genomics initiatives, which aim to determine structures for at least one example of each protein fold family, such comparative approaches are certain to become the methods of choice for structure prediction. Integrating such widespread structural annotation with functional data will make *in silico* simulation of cell and molecular biology a much more realistic proposition. Chapter 5 is a review of sequence-based methods of protein domain and motif identification, and is written by two distinguished proponents of the field: Ponting and Bateman. The emphasis is on generating testable hypotheses about protein function that can be followed up in the wet lab.

Chapters 6–8 are concerned with what might be thought of as the next stage of functional annotation; the elucidation of interacting networks of proteins. Here again the idea is to use evolutionary conservation to reveal function, and the research advances our knowledge of molecular evolution as much as it contributes to functional annotation. In Chapter 6, Huynen and Snel describe methods to predict the participation of proteins in a common pathway based on evolutionary conserved, spatial relationships between the genes that encode them. For a particular pathway (the assembly of iron–sulphur clusters in Proteobacteria and mitochondria), they show the power of comparative evolutionary analysis to shed light on functional relationships within the pathway. Events such as gene duplication, loss and fusion that are observed for the same genes in different species are used to infer subsets or modules of interacting proteins within the pathway. This theme is continued by Federova *et al* in Chapter 7, who use similar investigations of 'genomic context' to predict interactions within the two most common, multicomponent complexes in bacteria, the ABC-type transporters and two-component signal transduction proteins. This approach is taken furthest in Chapter 8 by Gelfand and Laikova, who show some success in reconstructing the evolution of regulatory networks in bacterial genomes.

Chapter 9 takes a diversion into the RNA world and ends up in the wet lab, as Huttenhofer and Brosius review the field of experimental 'RNomics' (a term so laboured and cumbersome it makes 'transcriptomics' seem positively welcoming). The emphasis is on the expanding universe of small nonmessenger RNA (snmRNA) molecules, how to detect them and how they are related. As they point out, the impressive computational progress in this field largely depends on the detection of similarity, which sets rigid limits to the

discovery of novel classes of snmRNAs. The feedback between the computational analysis and experimental work make this field a model of collaborative computational science. The final two chapters are concerned with developments in the new field of evolutionary genomics, within which Eugene Koonin, the senior author of these chapters and this book as a whole, is an acknowledged giant. In Chapter 10, 'Genome-Scale Phylogenetic Trees' Koonin and co-workers assess the impact of multiple genome sequences on the feasibility of a 'tree of life', a single phylogenetic tree depicting the history of life. In spite of ubiquitous horizontal gene transfer and differential gene loss they show the way to a revised tree, which preserves a three-domain view of life. In Chapter 11, the focus shifts to the proteome, where the authors develop their ideas on the birth, death and innovation model (BDIM) for the evolution of proteome composition. This model postulates three elementary processes governing proteome evolution: domain birth (duplication with divergence), death (inactivation or deletion) and innovation (emergence of novel domains). Appropriately, this final chapter is the one that fully realizes the promise of the book, to show the route from bioinformatics as stamp collecting to a theoretical branch of molecular biology. They start from the familiar, narrow phenomenological treatment of domain families; what domains are present and how they are related. However, they go on to assess the characteristics of the BDIM that best fit this data, testing various mathematical models of the relationships between the three elementary processes. This modelling reveals several new facets of genome evolution, most importantly it shows that within the BDIM framework there is an equilibrium state that is approached rapidly from any initial starting point. In this equilibrium state, the number of protein families in each size class remains constant, with a distribution dependent on the particular parameters of the BDIM. Comparisons to real data suggest that the domain diversities encoded by genomes are close to a steady state, and indicate that evolving lineages go through long periods of relative stasis punctuated by short periods of dramatic change where genomic complexity increases or decreases. The authors note the parallel with the punctuated equilibrium concepts of Gould and Eldredge (Gould, 2002) and implicitly suggest a link between evolution at the molecular and macroevolution as observed in the paleontological record. This provides a fitting end to the book, demonstrating the exciting prospects for modelling genome evolution using the amassed sequence data to test theoretical predictions.

References

Claverie JM (2000). From bioinformatics to computational biology. *Genome Res* **10**: 1277–1279.

Gould SJ (2002). *The Structure of Evolutionary Theory.* Harvard University Press: Cambridge, MA.

CAM Semple
MRC Human Genetics Unit,
Crewe Road, Edinburgh EH4 2XU, UK
E-mail: colins@hgu.mrc.ac.uk