Hui-Huang Hsu, Advanced Data Mining Technologies in Bioinformatics, Editors: Idea Group Publishing, pp 239

With the increasingly vital and central role that bioinformatics is playing today in all areas of biomedical research, resulting in dozens of new areas of application each year, the book “Advanced Data Mining Technologies in Bioinformatics”, by Hui-Huang Hsu, could have been just what was missing from the bookshelves of computational and experimental biologists alike.

On the other hand, I have to confess that the word “advanced” in the title of any technical book aimed at an entire field—say, for instance. Advanced Mathematical techniques in Economics—inevitably causes me a certain degree of malaise. The fundamental question is whether one can in fact cover a topic at the advanced level without sacrificing the generality of the discourse or without making some marginal or questionable choices in the selection of the specific topics to be covered at an advanced level. Thus, when I started reading the new manuscript, I was not sure whether I would find a well-organized structure with the potential to help further my understanding of the domain or rather a random collection of scattered and likely difficult to understand topics. As a note to future authors interested in publishing books on advanced topics, note that my anxiety would have been dramatically alleviated if the book had been restricted to a more specific subfield of bioinformatics, say the analysis of microarray expression profile time series or the inference of protein interaction networks.

Unfortunately, these somewhat irrational fears turned out to be well funded as the book by Dr. Hsu attempts to cover way too much ground, in a seemingly uncoordinated fashion, often giving more space to obscure rather than truly advanced techniques. The majority of these problems, to be fair, are due to the fact that the manuscript is organized as a loose collection of research articles, covering a variety of topics (some of which have little to do with data mining) without any kind of skeletal structure to hold the various parts together. While such a structure is justified when covering a monographic subject, it should not be used when addressing an entire field of research. One of the problems of such an organization is that the book fails to identify its target audience in either the mathematically inclined experimental biologist, with individual chapters addressing either audience at will. Bioinformaticians will no doubt be concerned by the haphazard structure of the book, ad hoc selection of topics and techniques, and highly heterogeneous treatment of the various topics. The experimental biologists are likely to stop reading on page 11, after learning that “a gene is a possible protein-coding region” and that “Viruses from animals cause a lot of human disease [such as] Foot and Mouth disease”. A typical example of a non-coding gene is given by micro RNAs and of course human transmission of Foot and Mouth disease is exceedingly rare, with only one reported case in the medical literature (Lancet 1967).

While organization and structure are important, the true limitation of this book is that the proposed compilation of papers fails to deliver a solid foundation of the state of the art in methods for data mining in bioinformatics research. Rather, it proposes a variety of ad hoc and at times obscure techniques that are not in the mainstream of data analysis, such as Chapters III (Combinatorial Fusion Analysis) and VIII (Pattern Differentiation and Formulations), or for covering topics that have almost nothing to do with data mining, such as Chapters V (Proteomics with Mass Spectrometry) and VII (Algorithmic Aspects of Protein Threading). One cannot help but wonder why Chapter III, on evidence fusion, would completely ignore Bayesian evidence integration techniques, a staple of modern bioinformatics, in favor of approaches such as voting or linear combination, which have been shown to have inferior performance. Similarly, why were none of the most recent techniques for gene expression analysis covered in Chapter X in favor of a variant of Principal Component Analysis for feature selection?

At the level of the individual chapter, there is significant variability in the quality of the topic coverage, and in the complexity of analytical techniques. Some chapters, such as Chapter V, VI, XV, and XVI are reasonably well written but lack any analytical insight and are probably more appropriate for the experimentalist that would like to get a list of available applications, respectively, for proteomics, phylogenetics, text mining, and Genome Annotation. On the other hand, they are of limited value to the bioinformatician who would like to achieve an in depth understanding of these problem and of the available analytical approaches. Others, such as Chapter VIII, XI, XIII, and XIV are more qualitative than analytical in nature and have little to offer to our “advanced” understanding of
the field. Only, a few are appropriate for such a book. Among these are chapter IX, which introduces “Parameterless Clustering Techniques for Gene Expression Analysis” (basically a standard Parameter Optimization approach) and Chapter XII, which introduces a reasonable, albeit relatively simple Bayesian Evidence Integration approach.