Fundamentals of Data Mining in Genomics and Proteomics

Edited by Werner Dubitzky, Martin Granzow and Daniel P. Berrar

More than ever before, research and development in genomics and proteomics depends on the analysis and interpretation of large amounts of data generated by high-throughput techniques. With the advance of computational systems biology, this situation will become even more manifest as scientists will generate truly large-scale data sets by simulating biological systems and conducting synthetic experiments. To optimally exploit such data, life scientists need to understand the fundamental concepts and properties of the fast-growing arsenal of analytical techniques and methods from statistics and data mining. Typically, the relevant literature and products present these techniques in a form which is either very simplistic or highly mathematical, favoring formal rigor over conceptual clarity and practical relevance. ***Fundamentals of Data Mining in Genomics and Proteomics*** addresses these shortcomings by adopting an approach which focuses on fundamental concepts and practical applications.

The book presents key analytical techniques used to analyze genomic and proteomic data by detailing their underlying principles, merits and limitations. An important goal of this text is to provide a highly intuitive and conceptual (as opposed to intricate mathematical) account of the discussed methodologies. This treatment will enable readers with interest in analysis of genomic and proteomic data to quickly learn and appreciate the essential properties of relevant data mining methodologies without recourse to advanced mathematics. To complement the conceptual discussions, the book draws upon the lessons learned from applying the presented techniques to concrete analysis problems in genomics and proteomics. The caveats and pitfalls of the discussed methods are highlighted by addressing questions such as: What can go wrong? Under which circumstances can a particular method be applied and when should it not be used? What alternative methods exist? Extensive references to related material and resources are provided to assist readers in identifying and exploring additional information.

The structure of this text mirrors the typical stages involved in deploying a data mining solution, spanning from data pre-processing to knowledge discovery to result post-processing. It is hoped that this will equip researchers and practitioners with a useful and practical framework to tackle their own data mining problems in genomics and proteomics. In contrast to some texts on machine learning and biological data analysis, a deliberate effort has been made to incorporate important statistical notions. By doing so the book is following demands for a more statistical data mining approach to analyzing high-throughput data. Finally, by highlighting limitations and open issues, ***Fundamentals of Data Mining in Genomics and Proteomics*** is intended to instigate critical thinking and avenues for new research in the field.

ISBN 0-387-47508-7

springer.com