PREFACE

Computational biology is nowadays one of the cornerstones in biological and medical data analysis and has a long and proud history originating in the 1960s from the fields of biophysics and protein biochemistry, notably the modeling of enzymatic reactions and other kinetic parameters. With the advent of improved and easier to access computing systems came the possibility of exploring biological systems to a much greater depth, especially linked to large-scale analytics platforms of biological samples, such as whole-genome sequencing tools, arrays, mass spectrometry, and many more. Such a considerable volume of data procured in a fast-paced technology-dependent manner required new ways to handle, manage, and analyze the information through improved data analytics streams, which was accomplished by borrowing and applying know-how from other sciences, such as mathematics, statistics, and computer sciences to biology, medicine, and disease analysis. This led to a vast expansion of data repositories and available computational tools feeding into reference databases and constantly improving our understanding of complex biological mechanisms. Ultimately, our ability to handle vast amounts of complex data enables us to integrate the various data streams into a contextualized system through systems approaches, network analysis, and modeling methodologies. Although it is evident that many gaps in our understanding of how any given biological system works still remain, more powerful systems, platforms, and procedures have started to emerge, such as automated decision machines, artificial intelligence, pattern matching approaches, and integrated and integrative data handling protocols, which will help us to continue uncovering new insights.

This book is aimed at both novices and specialists in the field of computational biology and brings together a selection of approaches at the cutting edge of technology and shows both how data and analytics procedures aid us in expanding our understanding of biology and how aberrant or modulated processes can lead to diseases. The first section (chapters 1–5) provides a more general overview. Chapter 1 gives an introduction to image-based systems biology of multicellular spheroids for experimentalists and theoreticians. Here, mathematical models of spheroids are used to investigate cellular interactions since tissues, cells, and even smaller components can be abstracted to spheroids to give a three-dimensional representation of the structural organization within a system. Chapter 2 discusses integrative biology approaches applied to human diseases, in particular, to multifactorial and complex interactions encountered in studying aberrant etiology. Concepts and analytics techniques are introduced for single-layer omics methods as well as procedures to integrate multi-omics data, leading to meaningful and relevant biological insights. Following this is the approach of using machine learning or deep learning in omics data analysis and precision medicine, as described in Chapter 3: deep learning allows us to identify complex patterns and create predictive models from omics data, as well as medical image analysis. Chapter 4 addresses the use of computational biology and bioinformatics in biological sequence analysis, where not only sequence alignments are an important step but also feature detection and selection are of significance. Supervised and

unsupervised learning, neural networks, and hidden Markov models are discussed, as well as deep sequencing or next-generation sequencing data analysis procedures using artificial intelligence and machine learning methods. Statistical procedures to analyze multi-omics data are presented in Chapter 5, using multivariate statistical methods for high-dimensional multiset omics data analysis. Application of canonical correlation analysis, redundancy analysis, and penalized versions are commonly used in omics dataflows, and this chapter gives an overview of how these methods came to match the statistical challenges that come with high-dimensional multiset omics data analysis.

A more specific overview of approaches in computational biology is given in chapters 6–9. Statistical methods for RNA sequencing data analysis are presented in Chapter 6. It covers the statistical models, model assumptions, and challenges encountered in RNA sequencing data analysis, including differential analysis, clustering approaches, and pathway analysis. Here, data analytics packages and embedded statistical methods and how they perform using real-world data are described. Chapter 7 addresses computational epigenomics, ranging from fundamental research to disease prediction and risk assessment. The epigenome encompasses several chemical properties of DNA and DNA-associated proteins that are tissue-specific, distinctive for a disease state, and sensitive to environmental conditions. Mining of genomic data sets and their associated epigenomic features, as well as the computational approaches used to assess statistical significance in comparative analyses, are discussed in this chapter. Chapter 8 discusses computational approaches in proteomics, where an overview of proteomic approaches, biological sample considerations, and data acquisition methods is given. Additionally, data processing software solutions for the various steps and further functional analyses of biological data are presented, which enable the comparison of various data sets as a summation of individual experiments, to cross-compare sample types and other metadata. Chapter 9 reviews cheminformatics and computational approaches in metabolomics using data mining methods and bioinformatics tools, including machine learning approaches. In this chapter, the main technical procedures used in metabolomics data acquisition, data processing, and pipelines, and the ways in which metabolomics data can aid in elucidating aberrant pathways and metabolic dysfunctions in disease, are discussed.

The last two chapters cover more specialized topics. Chapter 10 discusses the nature of feature selection in high-dimensional data using entropy information through statistical inference concepts of entropy in microarray data clustering in order to reduce the multi-dimensionality inherent in the source data to allow data summarization and the specific selection of gene sets associated with modulated conditions such as those found in diseases. The last review, Chapter 11, addresses structural pattern mining approaches applied to cryo-electron tomography using template-based and template-free procedures, where the observation of cellular organelles and macromolecular complexes at nanometer resolution with native conformations requires supervised deep learning-based pattern mining approaches in order to identify and reconstruct biological structures on the cellular as well as molecular level.

A full comprehensive summary of work carried out in the field of computational biology would span many volumes as this discipline is now deeply embedded in practically all large-scale, multi-subject, or integrative data analytics investigations.

Not only is this a very active field in terms of applications but also in the development of novel algorithms, resources, and pipelines. Condensing a torrent of data into contextualized, coherent, and understandable information to explain biology, disease, or to be used in fields such as personalized medicine is no longer possible without the aid of computational tools. Over time, it is expected that new and exciting developments in computational biology will allow us to gain an as yet unprecedented ability to make sense out of seemingly random data in a timely and precise manner. We believe the readers would enjoy the work presented in this book and will be both enlightened and encouraged to further the understanding of the biological world and how we work as a complex assembly of cells and an organism as a whole.

Holger Husi, Dr sc nat
Division of Biomedical Science
University of the Highlands and Islands, UK
October 2019
Doi: http://dx.doi.org/10.15586/computationalbiology.2019.pr