Preface

The new edition of this book is rather different from the first edition, though the general organization may seem quite similar. A new, small part, focused on the Big Data issue, has been added to the three parts already present in the first edition (Databases, Computational Techniques, and Prediction Methods). And the contents of the old parts have been substantially modified.

The book philosophy was maintained. Since the theoretical foundations of the biological sciences are extremely feeble, any discovery must be strictly empirical and cannot overtake the horizon of the observations. The central importance of empirical information is mirrored in the fact that experimental observations are being produced ceaselessly, in a musical accelerando, and Biology is becoming more and more a "data-driven" scientific field. The European Bioinformatics Institute, part of the European Molecular Biology Laboratory, is one of the largest biology-data repositories with its 20 petabytes of data—20,000 terabytes hard disks, like those that are commonly installed in our personal computers—and the development of innovative procedures for data storage and distribution became compelling [1, 2].

However, it must be remembered that "data" is not enough. For example, the promises of full human genome sequencing with regard to medical and biotechnological applications have been realized not even nearly to the expectations. Most importantly, more than half of the human genes still remain without any or with grossly insufficient functional characterization, the understanding of noncoding RNA functions is enigmatic and, most likely, three quarters of molecular pathways and assemblies in human are still open for discovery [3, 4]. In other words, with no appropriate scientific questions, data remain inert and discoveries are impossible. Without the observations made during the voyage on the Beagle, Darwin would have never written *On the Origin of the Species.* Similarly, rules of heredity were discovered by the friar Gregor Mendel and not by his sacristan. In other words, good science is made by good questions.

Databases and data mining tools are nevertheless indispensable in the era of data abundance and excess, which contrasts the not-so-ancient era when the problem was the access to the scarce data. In this book, the reader can find a description of several important databases: First, the genomic databases and their accession tools at the National Center for Biotechnology Information (1); then the archives of macromolecular three-dimensional structures (2). A chapter is focused on databases of protein–protein interactions (3) and another on thermodynamics information on protein and mutant stability (4). A further chapter is devoted to the "Kbdock" protein domain structure database and its associated web site for exploring and comparing protein domain–domain interactions and domain–peptide interactions (5). Structural data are archived also in PDB_REDO databank, which provides re-refined and partially rebuilt crystallographic structure models for PDB entries (6). This addresses a crucial point in databases—the quality of the data [4]—which is considered also in the next chapter, focused on tools and problems in building high-quality subsets of the Protein Data Bank (7). The last chapter is devoted to large-scale homology-based annotations (8).

The second part of the book, dedicated to data mining tools, hosts two chapters focused on data quality check and improvement. One focuses the attention on the identification and correction of erroneous sequences (9) and the other describes tools that allow one to improve pseudo-atomic models from Cryo-Electron Microscopy experiments (10). Then, a chapter describes tools in the ever-green motif of the substitution matrices (11). The problem of reproducibility of biochemical data is then addressed in Chapter 12 and tools to align RNA sequences are described in Chapter 13.

New developments in the computational treatment of protein conformational disorder are then summarized in Chapter 14, while interesting procedures for kinase family/subfamily classifications are described in Chapter 15. Then, new techniques to identify latent regular structures in DNA sequence (16) and new tools to predict protein crystallizability (17) are described. Chapter 18 is then focused on new ways to analyze sequence alignments, Chapter 19 describes tools of data mining based on ontologies, and Chapter 20 summarizes techniques of functional annotations based on metabolomics data. Then, a chapter is devoted to bacterial genomics data analyses (21) and another to prediction of pathophysiological effects of mutations (22). Chapter 23 is focused on drug–target interaction predictions, Chapter 24 deals with predictions of protein residue contacts, and the last Chapter (25) of this part describes the recipe for protein sequence-based function prediction and its implementation in the latest version of the ANNOTATOR software suite.

Two chapters are then grouped in the final part of the book, focused on the analyses of Big Data. Chapter 26 deals with metagenomes analyses and Chapter 27 describes resources and data mining tools in plant genomics and proteomics.

Pavia, Italy *Oliviero Carugo*
Vienna, Austria
Singapore, Singapore *Frank Eisenhaber*

References

1. Marx V (2013) The big challenges of Big Data. Nature 498: 255–260
2. Stein LD, Knoppers BM, Campbell P, Getz G, Korbel JO (2015) Create a cloud commons. Nature 523: 149–151
3. Eisenhaber F (2012) A decade after the first full human genome sequencing: when will we understand our own genome? J. Bioinform Comput Biol 10:1271001
4. Kuznetsov V, Lee HK, Maurer-Stroh S, Molnar MJ, Pongor S, Eisenhaber B, Eisenhaber F (2013) How bioinformatics influences health informatics: usage of biomolecular sequences, expression profiles and automated microscopic image analyses for clinical needs and public health. Health Inform Sci Syst 1: 2

Springer