Preface

Most life science researchers will agree that biology is not a truly theoretical branch of science. The hype around computational biology and bioinformatics beginning in the nineties of the 20th century was to be short lived (1, 2). When almost no value of practical importance such as the optimal dose of a drug or the three-dimensional structure of an orphan protein can be computed from fundamental principles, it is still more straightforward to determine them experimentally. Thus, experiments and observations do generate the overwhelming part of insights into biology and medicine. The extrapolation depth and the prediction power of the theoretical argument in life sciences still have a long way to go.

Yet, two trends have qualitatively changed the way how biological research is done today. The number of researchers has dramatically grown and they, armed with the same protocols, have produced lots of similarly structured data. Finally, high-throughput technologies such as DNA sequencing or array-based expression profiling have been around for just a decade. Nevertheless, with their high level of uniform data generation, they reach the threshold of totally describing a living organism at the biomolecular level for the first time in human history. Whereas getting exact data about living systems and the sophistication of experimental procedures have primarily absorbed the minds of researchers previously, the weight increasingly shifts to the problem of interpreting accumulated data in terms of biological function and biomolecular mechanisms. It is possible now that biological discoveries are the result of computational work, for example, in the area of biomolecular sequence analysis and gene function prediction (2, 3).

Electronically readable biomolecular databases are at the heart of this development. Biological systems consist of a giant number of biomacromolecules, both nucleic acids and proteins together with other compounds, organized in complexes pathways, subcellular structures such as organelles, cells, and the like that is interpreted in a hierarchical manner. Obviously, much remains unknown and not understood. Nevertheless, electronic databases organize the existing body of knowledge and experimental results about the building blocks, their relationships, and the corresponding experimental evidence in a form that enables the retrieval, visualization, comparison, and other sophisticated analyses. The significance of many of the pieces of information might not be understood when they enter databases; yet, they do not get lost and remain stored for the future.

Importantly, databases allow analyses of the data in a continuous workflow detached from any further experimentation itself. In a formal, mathematical framework, researchers can now develop theoretical approaches that may lead to new insights at a meta-analytic level. Indeed, results from many independently planned and executed experiments become coherently accessible with electronic databases. Together, they

can provide an insight that might not be possible from the individual pieces of information in isolation. It is also interesting to see this work in a human perspective: in the framework of such meta-analyses, people of various backgrounds who have never met essentially cooperate for the sake of scientific discoveries via database entries. From the technical viewpoint, because the data are astronomically numerous and the algorithms for their analysis are complex, the computer is the natural tool to help researchers in their task; yet, it is just a tool and not the center of the intellectual concept. The ideas and approaches selected by researchers driven by the goal to achieve biologically relevant discoveries remain the most important factor. Due to the need of computer-assisted data analysis, electronic availability of databases, the possibility of their download for local processing, the uniform structure of all database entries as well as the accuracy of all pieces of information including that for the level of experimental evidence are of utmost importance. To allow curiosity-driven research for as many as possible researchers and to enable the serendipity of discovery, the full public availability of the databases is critical.

Nucleic acid and protein sequence and structure databases were the first biological data collections in this context; the emergence of the sequence homology concept and the successes of gene function prediction are scientific outcomes of working with these data (3). To emphasize, they would be impossible without prior existence of the sequence databases. Thus, biological data mining is going to become the core of biological and biomedical research work in the future, and every member of the community is well advised to keep himself informed about the sources of information and the techniques used for "mining" new insights out of databases. This book is thought as a support for the reader in this endeavor.

The variety of biological databases reflects the complexity of and the hierarchical interpretation we use for the living world as well as the different techniques that are used to study them (4). The first section of the book is dedicated to describing concepts and structures of important groups of databases for biomolecular mechanism research. There are databases for sequences of genomes, nucleic acids such as RNAs and proteins, and biomacromolecular structures. With regard to proteins, databases collect instances of sequence architectural elements, thermodynamic properties, enzymes, complexes, and pathway information. There are many more specialized databases that are beyond the scope of this book; the reader is advised to consult the annual January database supplement of the journal "*Nucleic Acids Research*" for more detail (5).

The second section of this book focuses on formal methods for analyzing biomolecular data. Obviously, biological data are very heterogeneous and there are specific methodologies for the analysis of each type of data. The chapters of this book provide information about approaches that are of general relevance. Most of all, these are methods for comparison (measuring similarity of items and their classification) as well as concepts and tools for automated learning. In all cases, the approaches are described with the view of biological database mining.

The third section provides reviews on concepts for analyzing biomolecular sequence data in context with other experimental results that can be mapped onto genomes. The

topics range from gene structure detection in genomes and analyses of transcript sequences over aspects of protein sequence studies such as conformational disorder, 2D, 3D, and 4D structure prediction, protein crystallizability, recognition of post-translational modification sites or subcellular translocation signals to integrated protein function prediction.

It should be noted that the biological and biomedical scientific literature is the largest and possibly most important source of information. We do not analyze the issue here in this book since there is a lot in the flow. Whereas sources such as PUBMED or the Chemical Abstracts currently provide bibliographic information and abstracts, the trend is towards full-text availability. With the help of the open access movement, this goal might be practically achieved in a medium term. The processing of abstracts and full articles for mining biological facts is an area of actively ongoing research and exciting developments can be expected here.

Creating and maintaining a biological database requires considerable expertise and generates an immense work load. Especially maintaining and updating are expensive. Although future success of research in the life sciences depends on the completeness and quality of the data in databases and of software tools for their usage, this issue does not receive sufficient recognition within the community as well as from the funding agencies. Unfortunately, the many academic groups feel unable to continue the maintenance of databases and software tools because funding might cover only the initial development phase but not the continued maintenance. An exit into commercial development is not a true remedy; typically, the access to the database becomes hidden by a system of fees and its download for local processing is excluded. Likewise, it appears important to assess before the creation of the database whether it will be useful for the scientific community and whether the effort necessary for maintenance is commensurate with the potential benefit for biological discovery (6). For example, maintaining programs that update databases automatically is a vastly more efficient way than cases where all entries need to be curated manually in an individual manner.

We hope that this book is of value for students and researchers in the life sciences who wish to get a condensed introduction to the world of biological databases and their applications. Thanks go to all authors of the chapters who have invested considerable time for preparing their reviews. The support of the Austrian GENAU BIN programs (2003–2009) for the editors of this book is gratefully acknowledged.

Oliviero Carugo
Frank Eisenhaber

References

1. Ouzounis, C.A. (2000) Two or three myths about bioinformatics. *Bioinformatics* 17, 853–854
2. Eisenhaber, F. (2006) Bioinformatics: Mystery, astrology or service technology. Preface for "Discovering Biomolecular Mechanisms with Computational Biology", Eisenhaber, F. (Ed.), 1st edition, pp. pp.1–10. Georgetown, New York: Landes Biosciences, Springer

3. Eisenhaber, F. (2006) Prediction of protein function: Two basic concepts and one practical recipe. In Eisenhaber, F. (Ed.), "Discovering Biomolecular Mechanisms with Computational Biology", 1st edition, pp. 39–54. Georgetown, New York: Landes Biosciences, Springer
4. Carugo, O., Pongor, S. (2002) The evolution of structural databases. *Trends Biotech.* 20, 498–501
5. Galperin, M.Y., Cochrane, G.R. (2009) Nucleic acids research annual database issue and the NAR online molecular biology database collection in 2009. *Nucleic Acids Res.* 37, D1–D4
6. Wren, J.D., Bateman, A. (2008) Databases, data tombs and dust in the wind. *Bioinformatics* 24, 2127–2128

http://www.springer.com/978-1-60327-240-7

Data Mining Techniques for the Life Sciences
Carugo, O.; Eisenhaber, F. (Eds.)
2010, XII, 408 p. 89 illus., Hardcover
ISBN: 978-1-60327-240-7
A product of Humana Press