

Preface

Many problems arising in biological, chemical, and medical research, which could not be solved in the past due to their dimension and complexity, are nowadays tackled by means of automatic elaboration. Powerful computers are indeed used intensively for solving many problems having biological origin, thus creating the emerging field of science called “bioinformatics.” However, the success of such approaches depends not only on brute computational strength of those computers, but also, and often critically, on the mathematical quality of the models and of the algorithms underlying those solution procedures.

Solving a problem may be seen as converting information, in such a way that the solution of the problem (information in output) is extracted from its description (information in input), possibly passing through a number of intermediate states. By adopting this view, information handled when dealing with many of the above-mentioned problems becomes, at some stage, a sequence. Nature often encodes relevant information into sequences. Therefore, a central role in bioinformatics is played by sequence analysis problems or by the related problems of analyzing the effects or the behavior of some sequence.

The present volume offers a detailed overview of some of the most interesting mathematical approaches to sequence analysis and other sequence-related problems. Special emphasis is devoted to problems concerning the most relevant biopolymers (proteins and genetic sequences), but the exposition is not limited to them. A considerable effort has been made to render the volume comprehensible to researchers coming from either of the two hemispheres of bioinformatics: mathematics and computer science on one side, and biology, chemistry, and medicine on the other.

Rather than an exhaustive coverage of the topic, which would be clearly impossible to do in just one book, the volume is intended as a snapshot of the latest research development and of the potentialities that operations research and machine learning techniques bring in this interdisciplinary field of research. Moreover, the volume aims at bridging the two mentioned halves of bioinformatics that are still quite disjoint, promoting a cross-fertilization hopefully fostering future research in the field.

Primary selection criterion for the chapters has been scientific quality and importance. Additional selection criteria have been: (1) considering only approaches having a nontrivial mathematical basis; and (2) providing up to date contents not already largely available in other books published on similar subjects.

Organization of the Volume

Due to the wide heterogeneity of the matter, from the point of view of both problems considered and techniques presented, it may be useful to the reader tracing the following short sketch of the volume organization.

The first part of the volume deals with problems originating from the study of protein sequences. Proteins and peptides are polymers made from units called amino acids, and a basic problem is the determination of their amino acid sequence when that is unknown. This is sometimes called analysis of the primary structure. In Chap. 1, Bruni deals with this problem, with a focus on peptides, since proteins are essentially polypeptide chains, and describes exact and complete approaches based on propositional logic.

To be able to perform their biological functions, proteins fold into specific spatial conformations. Another relevant problem is the determination of such structures, known as the problem of protein structure analysis or prediction. In particular, the disposition of highly regular substructures in the protein sequence, such as helices, sheets, and strands, is called the secondary structure, while the three-dimensional structure of a single protein molecule, and the spatial arrangement of the above-mentioned elements of the secondary structure, is called the tertiary structure.

In Chap. 2, Di Lena et al. describe approaches to protein structure analysis based on decomposition, with specific attention to the secondary structure prediction and the protein contact map prediction by means of machine learning techniques. In Chap. 3, Patrizi et al. tackle again the problem of secondary structure prediction, performing a classification by means of nonlinear binary optimization techniques, with the aim of detecting isoform proteins considered as markers in oncology. Similarly, in Chap. 4, Biba et al. describe approaches to the protein folding prediction by modeling the sequence by means of Markov logic networks, that is, networks obtained by introducing probability in first-order logic.

The volume then gradually moves to problems originating from the study of genetic sequences. Deoxyribonucleic acid, or DNA, is a long polymer made from repeating units called nucleotides. It contains the genetic instructions used in the development and functioning of all known living organisms. In Chap. 5, Ceci et al. deal with the problem of discovering motifs, that are sequence patterns frequently appearing in DNA, RNA, or proteins, and therefore probably having specific biological functions. They are discovered by mining association rules in the three-dimensional space.

In Chap. 6, Mosca and Milanesi consider the problem of studying intermolecular interactions among DNA, RNA, and proteins obtained by means of sequence analysis techniques. When viewing those interactions at a system level, the dynamics of biochemical pathways can be simulated, and therefore better understood, by means of mathematical models.

In Chap. 7, Graça et al. deal with the problem of determining haplotype information, that is, genetic information inherited from ancestors, from genotype information, that is, all the genetic constitution of an individual, using approaches based on propositional logic. On related themes, in Chap. 8, Catanzaro describes the

problem of calculating phylogenies, that is, graphs representing the evolutionary relationships among species. Several optimization models for estimating them from molecular data such as DNA and RNA under different paradigms are explained and discussed.

In Chap. 9, Salvi et al. tackle the problem of performing studies of human genome by means of data mining techniques, known as genome-wide association studies, for a stratified population. This means that the individuals of the population are not uniform but carry different genetic backgrounds, and this often produces false association results. The effects of different statistical techniques are considered to devise an efficient strategy for overcoming this problem.

The last part of the volume considers problems originating from the study of polymers not having biological origin. Polymerization reactions can be divided into: (i) addition polymerization, producing the so-called addition polymers (also classified as chain-growth polymers, with some exceptions), which grow one monomer at a time, and (ii) condensation polymerization, producing the so-called condensation polymers (also classified as step-growth polymers), which grow eliminating small molecules during the synthesis. In Chap. 10, Montaudo deals with the problem of predicting the sequence distribution of addition polymers; while in Chap. 11, Montaudo discusses the same problem for condensation polymers, using in both a variety of mathematical techniques.

Rome, Italy
March 2010

Renato Bruni

Mathematical Approaches to Polymer Sequence
Analysis and Related Problems

Bruni, R. (Ed.)

2011, X, 248 p., Hardcover

ISBN: 978-1-4419-6799-2