Pradipta Maji
Sushmita Paul

Scalable Pattern Recognition Algorithms

Applications in Computational Biology and Bioinformatics

Springer

Scalable Pattern Recognition Algorithms

Pradipta Maji · Sushmita Paul

Scalable Pattern Recognition Algorithms

Applications in Computational Biology and Bioinformatics

Springer

Pradipta Maji
Indian Statistical Institute
Kolkata, West Bengal
India

Sushmita Paul
Indian Statistical Institute
Kolkata, West Bengal
India

ISBN 978-3-319-05629-6 ISBN 978-3-319-05630-2 (eBook)
DOI 10.1007/978-3-319-05630-2
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014933668

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To my daughter

Pradipta Maji

To my parents

Sushmita Paul

Foreword

It is my great pleasure to welcome a new book on “Scalable Pattern Recognition Algorithms: Applications in Computational Biology and Bioinformatics” by Prof. Pradipta Maji and Dr. Sushmita Paul.

This book is unique in its character. Most of the methods presented in it are based on profound research results obtained by the authors. These results are closely related to the main research directions in bioinformatics. The existing conventional/traditional approaches and techniques are also presented, wherever necessary. The effectiveness of algorithms that are proposed by the authors is thoroughly discussed along with both quantitative and qualitative comparisons with other existing methods in this area. These results are derived through experiments on real-life data sets. In general, the presented algorithms display excellent performance. One of the important aspects of the methods proposed by the authors is their ability to scale well with the inflow data. It shall be mentioned that the authors provide in each chapter the directions for future research in the corresponding area.

The main aim of bioinformatics is the development and application of computational methods in pursuit of biological discoveries. Among the hot topics in this field are: sequence alignment and analysis, gene finding, genome annotation, protein structure alignment and prediction, classification of proteins, clustering and dimensionality reduction of gene expression data, protein–protein docking or interactions, and modeling of evolution. From a more general view, the aim is to discover unifying principles of biology using tools of automated knowledge discovery. Hence, knowledge discovery methods that rely on pattern recognition, machine learning, and data mining are widely used for analysis of biological data, in particular for classification, clustering, and feature selection.

The book is structured according to the major phases of a pattern recognition process (clustering, classification, and feature selection) with a balanced mixture of theory, algorithms, and applications. Special emphasis is given to applications in computational biology and bioinformatics.

The reader will find in the book a unified framework describing applications of soft computing, statistical, and machine learning techniques in construction of efficient data models. Soft computing methods allow us to achieve high quality solutions for many real-life applications. The characteristic features of these methods are tractability, robustness, low-cost solution, and close resemblance with humanlike decision making. They make it possible to use imprecision, uncertainty, approximate reasoning, and partial truth in searching for solutions. The main research directions in soft computing are related to fuzzy sets, neurocomputing, genetic algorithms, probabilistic reasoning, and rough sets. By integration or combination of the different soft computing methods, one may improve the performance of these methods.

The authors of the book present several newly developed methods and algorithms that combine statistical and soft computing approaches, including: (i) neural network tree (NNTree) used for identification of splice-junction and protein coding region in DNA sequences; (ii) a new approach for selecting miRNAs from microarray expression data integrating the merit of rough set-based feature selection algorithm and theory of $B.632+$ bootstrap error rate; (iii) a robust thresholding technique for segmentation of brain MR images based on the fuzzy thresholding technique; (iv) an efficient method for selecting set of bio-basis strings for the new kernel function, integrating the Fisher ratio and a novel concept of degree of resemblance; (v) a rough set-based feature selection algorithm for selecting sets of effective molecular descriptors from a given quantitative structure activity relationship (QSAR) data set.

Clustering is one of the important analytic tools in bioinformatics. There are several new clustering methods presented in the book. They achieve very good results on various biomedical data sets. That includes, in particular: (i) a method based on Pearson's correlation coefficient that selects initial cluster centers, thus enabling the algorithm to converge to optimal or nearly optimal solution and helping to discover co-expressed gene clusters; (ii) a method based on Dunn's cluster validity index that identifies optimal parameter values during initialization and execution of the clustering algorithm; (iii) a supervised gene clustering algorithm based on the similarity between genes measured with use of the new quantitative measure, whereby redundancy among the attributes is eliminated; (iv) a novel possibilistic biclustering algorithm for finding highly overlapping biclusters having larger volume and mean squared residue lower than a predefined threshold.

The reader will also find several other interesting methods that may be applied in bioinformatics, such as: (i) a computational method for identification of disease-related genes, judiciously integrating the information of gene expression profiles, and the shortest path analysis of protein–protein interaction networks; (ii) a method based on *f*-information measures used in evaluation criteria for gene selection problem.

This book will be useful for graduate students, researchers, and practitioners in computer science, electrical engineering, system science, medical science, bioinformatics, and information technology. In particular, researchers and practitioners in industry and R&D laboratories working in the fields of system design, pattern recognition, machine learning, computational biology and bioinformatics, data mining, soft computing, computational intelligence, and image analysis may benefit from it.

The authors and editors deserve the highest appreciation for their outstanding work.

Warsaw, Poland, December 2013 Andrzej Skowron

Preface

Recent advancement and wide use of high-throughput technologies for biological research are producing enormous size of biological data distributed worldwide. With the rapid increase in size of biological data banks, understanding the biological data has become critical. Such an understanding could lead us to the elucidation of the secrets of life or ways to prevent certain currently non-curable diseases. Although laboratory experiment is the most effective method for investigating the biological data, it is financially expensive and labor intensive. A deluge of such information coming in the form of genomes, protein sequences, and microarray expression data has led to the absolute need for effective and efficient computational tools to store, analyze, and interpret these multifaceted data.

Bioinformatics is the conceptualizing biology in terms of molecules and applying informatics techniques to understand and organize the information associated with the molecules, on a large scale. It involves the development and advancement of algorithms using techniques including pattern recognition, machine learning, applied mathematics, statistics, informatics, and biology to solve biological problems usually on the molecular level. Major research efforts in this field include sequence alignment and analysis, gene finding, genome annotation, protein structure alignment and prediction, classification of proteins, clustering and dimensionality reduction of microarray expression data, protein–protein docking or interactions, modeling of evolution, and so forth. In other words, bioinformatics can be described as the development and application of computational methods to make biological discoveries. The ultimate attempt of this field is to develop new insights into the science of life as well as creating a global perspective, from which the unifying principles of biology can be derived. As classification, clustering, and feature selection are needed in this field, pattern recognition tools and machine learning techniques have been widely used for analysis of biological data as they provide useful tools for knowledge discovery in this field.

Pattern recognition is the scientific discipline whose goal is the classification of objects into a number of categories or classes. It is the subject of researching object description and classification method. It is also a collection of mathematical, statistical, heuristic, and inductive techniques of the fundamental role in executing the tasks like human beings on computers. In a general setting, the process of pattern recognition is visualized as a sequence of a few steps: data acquisition; data preprocessing; feature selection; and classification or clustering. In the first step, data are gathered via a set of sensors depending on the environment within which the objects are to be classified. After data acquisition phase, some preprocessing tasks such as noise reduction, filtering, encoding, and enhancement are applied on the collected data for extracting pattern vectors. Afterward, a feature space is constituted to reduce the space dimensionality. However, in a broader perspective this stage significantly influences the entire recognition process. Finally, the classifier is constructed, or in other words, a transformation relationship is established between features and classes.

Pattern recognition, by its nature, admits many approaches, sometimes complementary, sometimes competing, to provide the appropriate solution for a given problem. For any pattern recognition system, one needs to achieve robustness with respect to random noise and failure of components and to obtain output in real time. It is also desirable for the system to be adaptive to the changes in the environment. Moreover, a system can be made artificially intelligent if it is able to emulate some aspects of the human reasoning system. Soft computing and machine learning approaches to pattern recognition are attempts to achieve these goals. Artificial neural network, genetic algorithms, fuzzy sets, and rough sets are used as the tools in these approaches. The challenge is, therefore, to devise powerful pattern recognition methodologies by symbiotically combining these tools for analyzing biological data in more efficient ways. The systems should have the capability of flexible information processing to deal with real-life ambiguous situations and to achieve tractability, robustness, and low-cost solutions.

Various scalable pattern recognition algorithms using soft computing and machine learning approaches, and their real-life applications, including those in computational biology and bioinformatics, have been reported during the last 5–7 years. These are available in different journals, conference proceedings, and edited volumes. This scattered information causes inconvenience to readers, students, and researchers. The current volume is aimed at providing a treatise in a unified framework describing how soft computing and machine learning techniques can be judiciously formulated and used in building efficient pattern recognition models. Based on the existing as well as new results, the book is structured according to the major phases of a pattern recognition system (classification, feature selection, and clustering) with a balanced mixture of theory, algorithm, and applications. Special emphasis is given to applications in computational biology and bioinformatics.

The book consists of 11 chapters. Chapter 1 provides an introduction to pattern recognition and bioinformatics, along with different research issues and challenges related to high-dimensional real-life biological data sets. The significance of pattern recognition and machine learning techniques in computational biology and bioinformatics is also presented in Chap. 1. Chapter 2 presents the design of a hybrid learning model, termed as neural network tree (NNTree), for identification of splice-junction and protein coding region in DNA sequences. It incorporates the advantages of both decision tree and neural network. An NNTree is a decision tree, where each non-terminal node contains a neural network. The versatility of this method is illustrated through its application in splice-junction and gene identification problems. Extensive experimental results establish that the NNTree produces more accurate classifier than that previously obtained for a range of different sequence lengths, thereby indicating a cost-effective alternative in splice-junction and protein coding region identification problem.

The prediction of protein functional sites is an important issue in protein function studies and drug design. In order to apply the powerful kernel-based pattern recognition algorithms such as support vector machine to predict functional sites in proteins, amino acids need encoding prior to input. In this regard, a new string kernel function, termed as the modified bio-basis function, is presented in Chap. 3. It maps a nonnumerical sequence space to a numerical feature space using a bio-basis string as its support. The concept of zone of influence of bio-basis string is introduced in the new kernel function to take into account the influence of each bio-basis string in nonnumerical sequence space. An efficient method is described to select a set of bio-basis strings for the new kernel function, integrating the Fisher ratio and the concept of degree of resemblance. The integration enables the method to select a reduced set of relevant and nonredundant bio-basis strings. Some quantitative indices are described for evaluating the quality of selected bio-basis strings. The effectiveness of the new string kernel function and bio-basis string selection method, along with a comparison with existing bio-basis function and related bio-basis string selection methods, is demonstrated on different protein data sets using the new quantitative indices and support vector machine.

Quantitative structure activity relationship (QSAR) is one of the important disciplines of computer-aided drug design that deals with the predictive modeling of properties of a molecule. In general, each QSAR data set is small in size with a large number of features or descriptors. Among the large amount of descriptors present in the QSAR data set, only a small fraction of them is effective for performing the predictive modeling task. Chapter 4 presents a rough set-based feature selection algorithm to select a set of effective molecular descriptors from a given QSAR data set. The new algorithm selects the set of molecular descriptors by maximizing both relevance and significance of the descriptors. The performance of the new algorithm is studied using the R^2 statistic of support vector regression method. The effectiveness of the new algorithm, along with a comparison with existing algorithms, is demonstrated on several QSAR data sets.

Microarray technology is one of the important biotechnological means that allows to record the expression levels of thousands of genes simultaneously within a number of different samples. An important application of microarray gene expression data in functional genomics is to classify samples according to their gene expression profiles. Among the large amount of genes present in microarray gene expression data, only a small fraction of them is effective for performing a certain diagnostic test. In this regard, mutual information has been shown to be successful for selecting a set of relevant and nonredundant genes from microarray data. However, information theory offers many more measures such as the f-information measures that may be suitable for selection of genes from microarray gene expression data.

Chapter 5 presents different *f*-information measures as the evaluation criteria for gene selection problem. The performance of different *f*-information measures is compared with that of mutual information based on the predictive accuracy of naive Bayes classifier, *k*-nearest neighbor rule, and support vector machine. An important finding is that some *f*-information measures are shown to be effective for selecting relevant and nonredundant genes from microarray data. The effectiveness of different *f*-information measures, along with a comparison with mutual information, is demonstrated on several cancer data sets.

One of the most important and challenging problems in functional genomics is how to select the disease genes. In Chap. 6, a computational method is reported to identify disease genes, judiciously integrating the information of gene expression profiles and shortest path analysis of protein–protein interaction networks. While the gene expression profiles have been used to select differentially expressed genes as disease genes using mutual information-based maximum relevance-maximum significance framework, the functional protein association network has been used to study the mechanism of diseases. Extensive experimental study on colorectal cancer establishes the fact that the genes identified by the integrated method have more colorectal cancer genes than the genes identified from the gene expression profiles alone. All these results indicate that the integrated method is quite promising and may become a useful tool for identifying disease genes.

The microRNAs or miRNAs regulate expression of a gene or protein. It has been observed that they play an important role in various cellular processes and thus help in carrying out normal functioning of a cell. However, dysregulation of miRNAs is found to be a major cause of a disease. Various studies have also shown the role of miRNAs in cancer and utility of miRNAs for the diagnosis of cancer. In this regard, Chap. 7 presents a new approach for selecting miRNAs from microarray expression data. It integrates the merit of rough set-based feature selection algorithm reported in Chap. 4 and theory of $B.632+$ bootstrap error rate. The effectiveness of the new approach, along with a comparison with other algorithms, is demonstrated on several miRNA data sets.

Clustering is one of the important analyses in functional genomics that discovers groups of co-expressed genes from microarray data. In Chap. 8, different partitive clustering algorithms such as hard *c*-means, fuzzy *c*-means, rough-fuzzy *c*-means, and self-organizing maps are presented to discover co-expressed gene clusters. One of the major issues of the partitive clustering-based microarray data analysis is how to select initial prototypes of different clusters. To overcome this limitation, a method is reported based on Pearson's correlation coefficient to select initial cluster centers. It enables the algorithm to converge to an optimum or near optimum solutions and helps to discover co-expressed gene clusters. In addition, a method is described to identify optimum values of different parameters of the initialization method and the clustering algorithm. The effectiveness of different algorithms is demonstrated on several yeast gene expression time-series data sets using different cluster validity indices and gene ontology-based analysis.

In functional genomics, an important application of microarray data is to classify samples according to their gene expression profiles such as to classify cancer versus normal samples or to classify different types or subtypes of cancer. Hence, one of the major tasks with the gene expression data is to find groups of co-regulated genes whose collective expression is strongly associated with the sample categories or response variables. In this regard, a supervised gene clustering algorithm is presented in Chap. 9 to find groups of genes. It directly incorporates the information about sample categories into the gene clustering process. A new quantitative measure, based on mutual information, is reported that incorporates the information about sample categories to measure the similarity between attributes. The supervised gene clustering algorithm is based on measuring the similarity between genes using the new quantitative measure. The performance of the new algorithm is compared with that of existing supervised and unsupervised gene clustering and gene selection algorithms based on the class separability index and the predictive accuracy of naive Bayes classifier, *k*-nearest neighbor rule, and support vector machine on several cancer and arthritis microarray data sets. The biological significance of the generated clusters is interpreted using the gene ontology.

The biclustering method is another important tool for analyzing gene expression data. It focuses on finding a subset of genes and a subset of experimental conditions that together exhibit coherent behavior. However, most of the existing biclustering algorithms find exclusive biclusters, which is inappropriate in the context of biology. Since biological processes are not independent of each other, many genes may participate in multiple different processes. Hence, nonexclusive biclustering algorithms are required for finding overlapping biclusters. In Chap. 10, a novel possibilistic biclustering algorithm is presented to find highly overlapping biclusters of larger volume with mean squared residue lower than a predefined threshold. It judiciously incorporates the concept of possibilistic clustering algorithm into biclustering framework. The integration enables efficient selection of highly overlapping coherent biclusters with mean squared residue lower than a given threshold. The detailed formulation of the new possibilistic biclustering algorithm, along with a mathematical analysis on the convergence property, is presented. Some quantitative indices are reported for evaluating the quality of generated biclusters. The effectiveness of the algorithm, along with a comparison with other algorithms, is demonstrated on yeast gene expression data set.

Finally, Chap. 11 reports a robust thresholding technique for segmentation of brain MR images. It is based on the fuzzy thresholding techniques. Its aim is to threshold the gray level histogram of brain MR images by splitting the image histogram into multiple crisp subsets. The histogram of the given image is thresholded according to the similarity between gray levels. The similarity is assessed through a second-order fuzzy measure such as fuzzy correlation, fuzzy entropy, and index of fuzziness. To calculate the second-order fuzzy measure, a weighted co-occurrence matrix is presented, which extracts the local information more accurately. Two quantitative indices are reported to determine the multiple thresholds of the given histogram. The effectiveness of the algorithm, along with a comparison with standard thresholding techniques, is demonstrated on a set of brain MR images.

The relevant existing conventional/traditional approaches or techniques are also included wherever necessary. Directions for future research in the concerned topic are provided in each chapter. Most of the materials presented in the book are from our published works. For the convenience of readers, a comprehensive bibliography on the subject is also appended in each chapter. It might have happened that some works in the related areas have been omitted due to oversight or ignorance.

The book, which is unique in its character, will be useful to graduate students and researchers in computer science, electrical engineering, system science, medical science, bioinformatics, and information technology both as a textbook and as a reference book for some parts of the curriculum. The researchers and practitioners in industry and R&D laboratories working in the fields of system design, pattern recognition, machine learning, computational biology and bioinformatics, data mining, soft computing, computational intelligence, and image analysis will also be benefited.

Finally, the authors take this opportunity to thank Mr. Wayne Wheeler and Mr. Simon Rees of Springer-Verlag, London, for their initiative and encouragement. The authors also gratefully acknowledge the support provided by Dr. Chandra Das of Netaji Subhash Engineering College, Kolkata, India and the members of Biomedical Imaging and Bioinformatics Lab, Indian Statistical Institute, Kolkata, India for preparation of a few chapters of the manuscript. The book has been written when one of the authors, Dr. S. Paul, held a CSIR Fellowship of the Government of India. This work is partially supported by the Indian National Science Academy, New Delhi (grant no. SP/YSP/68/2012).

Kolkata, India, January 2014

Pradipta Maji
Sushmita Paul

Contents

Part I Classification

Part II Feature Selection

Chapter 1
Introduction to Pattern Recognition and Bioinformatics

1.1 Introduction

With the gaining of knowledge in different branches of biology such as molecular biology, structural biology, and biochemistry, and the advancement of technologies lead to the generation of biological data at a phenomenal rate [286]. The enormous quantity and variety of information are being produced from the data of the myriad of projects that study gene expression, determine the protein structures encoded by the genes, and detail how these products interact with one another. This deluge of biological information has, in turn, led to an absolute need for computerized databases to store, organize, and index the data, and for specialized tools to view and analyze the data. Hence, computers have become indispensable to biological research. Such an approach is ideal due to the ease with which computers can handle large quantities of data and probe the complex dynamics observed in nature.

Bioinformatics is a multidisciplinary research area that conceptualizes biology in terms of molecules and applies information techniques to understand and organize the information associated with these molecules on a large scale. It involves the development and advancement of algorithms using techniques including pattern recognition, machine learning, applied mathematics, statistics, informatics, and biology to analyze the complete collection of DNA (the genome), RNA (the transcriptome), and protein (the proteome) of an organism [275]. Major research efforts in this field include sequence alignment and analysis, gene finding, genome annotation, protein structure alignment and prediction, classification of proteins, clustering and dimensionality reduction of microarray expression data, protein–protein docking or interactions, modeling of evolution, and so forth. In other words, bioinformatics can be described as the development and application of computational methods to make biological discoveries. The ultimate attempt of this field is to develop new insights into the science of life as well as creating a global perspective, from which the unifying principles of biology can be derived [20, 22, 209, 302, 377, 391].

Pattern recognition is the scientific discipline whose goal is the classification of objects into a number of categories or classes. It is the subject of researching

P. Maji and S. Paul, *Scalable Pattern Recognition Algorithms*,
DOI: 10.1007/978-3-319-05630-2_1,

object description and classification method. It is also a collection of mathematical, statistical, heuristic, and inductive techniques of fundamental role in executing the tasks like human being on computers [209, 260, 263]. As classification, clustering, and feature selection are needed in bioinformatics, pattern recognition and machine learning techniques have been widely used for analysis of biological data as they provide useful tools for knowledge discovery in this field. The massive biological databases are generally characterized by the numeric as well as textual, symbolic, and pictorial data. They may contain redundancy, errors, and imprecision. The pattern recognition is aimed at discovering natural structures within such massive and often heterogeneous biological data. It is visualized as being capable of knowledge discovery using generalizations and magnifications of existing and new algorithms. Therefore, pattern recognition plays a significant role in bioinformatics [20, 22, 209, 302, 377, 391]. It deals with the process of identifying valid, novel, potentially useful, and ultimately understandable patterns in voluminous, possibly heterogeneous biological data sets.

One of the main problems in biological data analysis is uncertainty. Some of the sources of this uncertainty include imprecision in computations and vagueness in class definition. Pattern recognition, by its nature, admits many approaches, sometimes complementary, sometimes competing, to provide the appropriate solution of a given problem. An efficient pattern recognition system for bioinformatics tasks should possess several characteristics such as online adaptation to cope with the changes in the environment, handling nonlinear class separability to tackle real-life problems, handling of overlapping classes or clusters for discriminating almost similar but different objects, real-time processing for making a decision in a reasonable time, generation of soft and hard decisions to make the system flexible, verification and validation mechanisms for evaluating its performance, and minimizing the number of parameters in the system that have to be tuned for reducing the cost and complexity. The property to emulate some aspects of the human processing system can be helpful for making a system artificially intelligent.

Soft computing and machine learning approaches to pattern recognition are attempts to achieve these goals. Artificial neural network, genetic algorithms, information theory, fuzzy sets, and rough sets are used as the tools in these approaches. The challenge is, therefore, to devise powerful pattern recognition methodologies by symbiotically combining these tools for analyzing biological data in more efficient ways. The systems should have the capability of flexible information processing to deal with real-life ambiguous situations and to achieve tractability, robustness, and low-cost solutions. Various scalable pattern recognition algorithms using soft computing and machine learning approaches have been developed to successfully address different problems of computational biology and bioinformatics [33, 56, 68, 83, 89, 98, 107, 131, 198, 210, 211, 318, 324, 350, 357, 363–365, 375, 376, 380].

The objective of this book is to provide some results of investigations, both theoretical and experimental, addressing the relevance of information theory, artificial neural networks, fuzzy sets, and rough sets to bioinformatics with real-life applications. Various methodologies are presented based on information theoretic measures, artificial neural networks, fuzzy sets, and rough sets for classification, feature selection, and clustering. The emphasis of these methodologies is given on (a) handling biological data sets which are large, both in size and dimension, and involve classes that are overlapping, intractable, and/or having nonlinear boundaries, (b) demonstrating the significance of pattern recognition and machine learning for dealing with the biological knowledge discovery aspect, and (c) demonstrating their success in certain tasks of bioinformatics and medical imaging as examples. Before describing the scope of the book, a brief overview of molecular biology and pattern recognition is provided.

The structure of the rest of this chapter is as follows: Section 1.2 briefly presents a description of the basic concept of molecular biology. In Sect. 1.3, several bioinformatics problems are reported, which are important to retrieve useful biological information from large data sets using pattern recognition and machine learning techniques. In Sect. 1.4, the pattern recognition aspect is elaborated, discussing its components, tasks involved, and approaches, along with the role of soft computing in bioinformatics and computational biology. Finally, Sect. 1.5 discusses the scope and organization of the book.

1.2 Basics of Molecular Biology

The molecular biology deals with the formation, structure, and function of macromolecules essential to life, such as carbohydrates, nucleic acids, and proteins, including their roles in cell replication and the transmission of genetic information [190]. This field overlaps with other areas of biology and chemistry, particularly genetics and biochemistry. This section presents the basic concepts of nucleic acids and proteins.

1.2.1 Nucleic Acids

The weakly acidic substance present inside a nuclei is known as nucleic acids. They are large biological molecules essential for all known forms of life. They include deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) [190].

1.2.1.1 DNA

It contains the instructions needed by the cell to carry out its functions [190]. DNA consists of two long interwoven strands that form the famous double helix. Each strand is built from a small set of constituent molecules called nucleotides. The first two parts of the nucleotides are used to form the ribbon-like backbone of the DNA strand, and are identical in all nucleotides. These two parts are a phosphate group and a sugar called deoxyribose. The third part of the nucleotide is the base. There are four different bases, which define the four different nucleotides, namely, thymine

(**T**), cytosine (**C**), adenine (**A**), and guanine (**G**). The base pair complementarity makes a DNA molecule double stranded. If specific bases of one strand are aligned with specific bases on the other strand, the aligned bases can hybridize via hydrogen bonds, weak attractive forces between hydrogen and either nitrogen or oxygen. The specific complementary pairs are **A** with **T** and **G** with **C**. Two hydrogen bonds occur between **A** and **T**, whereas three bonds are formed between **C** and **G**. This makes **C**–**G** bonds stronger than **A**–**T** bonds.

DNA is the genetic material, used in development and functioning of all known living organisms and many viruses. It contains informations that are required to construct other important components of a cell like protein and RNA molecules. This biological information of DNA is decoded with the help of ribosomes, which links amino acids in an order specified by messenger RNA (mRNA), using transfer RNA molecules to carry amino acids and to read the mRNA three nucleotides at a time. The genetic code is highly similar among all organisms, and can be expressed in a simple table with 64 entries. These 64 codons code for 20 different amino acids. The code defines how sequences of these nucleotide triplets, called codons, specify which amino acid will be added next during protein synthesis. Amino acids play central roles both as building blocks of proteins and as intermediates in metabolism. The DNA sequences that code for protein are known as genes, other part of DNA is known as junk DNA. Much of this DNA has no known biological function. However, many types of it do have known biological functions, including the transcriptional and translational regulation of protein coding sequences. A brief description of important components and processes of DNA is as follows [190]:

- *Gene* is a molecular unit of heredity of a living organism. Living beings depend on genes, as they specify all proteins and functional RNA chains. Genes hold the information to build and maintain an organism's cells and pass genetic traits to offspring. All organisms have many genes corresponding to various biological traits, some of which are immediately visible, such as eye color or number of limbs, and some of which are not, such as blood type, increased risk for specific diseases, or the thousands of basic biochemical processes that comprise life.
- *Gene expression* is the process by which information from a gene is used in the synthesis of a functional gene product. These products are often proteins, but in nonprotein coding genes such as ribosomal RNA genes or transfer RNA genes, the product is a functional RNA. The process of gene expression is used by all known life—eukaryotes (including multicellular organisms), prokaryotes (bacteria and archaea), and viruses—to generate the macromolecular machinery for life.
- *Transcription* is the process of making an RNA copy of a gene sequence. In a eukaryotic cell, this copy, called mRNA molecule, leaves the cell nucleus and enters the cytoplasm, where it directs the synthesis of the protein, which it encodes. However, in a prokaryotic cell there is no nucleus, so the transcription as well as translation take place in cytoplasm.
- *Translation* is the process of translating the sequence of a mRNA molecule to a sequence of amino acids during protein synthesis. The genetic code describes the relationship between the sequence of base pairs in a gene and the corresponding

amino acid sequence that it encodes. In the cell cytoplasm, the ribosome reads the sequence of the mRNA in groups of three bases to assemble the protein.

1.2.1.2 RNA

The mRNA and other types of RNAs are single-stranded nucleic acids made up of ribose sugar, phosphate group, and nucleobases (**G**, **A**, uracil (**U**), **C**). The genetic information stored in DNA is transferred into RNA through transcription by DNA polymerase, and the information is decoded when RNA is translated into proteins. The proteins largely constitute the machinery that makes life live. They carry out all structural, catalytic, and regulatory functions. Hence, RNAs mostly play the passive role of a messenger. RNAs can be divided into two classes, namely, coding RNA and noncoding RNA.

The RNAs that code for proteins are known as coding RNA. The transcribed coding RNAs, that is, mRNAs are further translated into proteins. The mRNA serves as a template for protein synthesis. It is transcribed from a gene and then translated by ribosomes in order to manufacture a protein. Hence, it is known as coding RNA. The sequence of a strand of mRNA is based on the sequence of a complementary strand of DNA. The RNAs those do not translated into proteins are known as noncoding RNAs. The noncoding RNAs have been found to carry out very diverse functions, from mRNA splicing and RNA modification to translational regulation. MicroRNA (miRNA) is one type of noncoding RNAs. The miRNAs are small noncoding RNAs of length around 22 nucleotides, present in animal and plant cell. They regulate the expression of mRNAs posttranscriptionally, resulting in translational repression and gene silencing. Hence, miRNAs are related to diverse cellular processes and regarded as important components of the gene regulatory network [275].

1.2.2 Proteins

Proteins are organic compounds made of amino acids arranged in a linear chain and folded into a globular or fibrous form [185]. The amino acids in a polymer are joined together by the peptide bonds between the carboxyl and amino groups of adjacent amino acid residues. The sequence of amino acids in a protein is defined by the sequence of a gene, which is encoded in the genetic code. Amino acids can be divided into two groups, namely, essential amino acids and nonessential amino acids. The liver, and to a much lesser extent the kidneys, can convert amino acids used by cells in protein biosynthesis into glucose by a process known as gluconeogenesis. The essential amino acids, which must be obtained from external sources such as food, are leucine, isoleucine, valine, lysine, threonine, tryptophan, methionine, phenylalanine, and histidine. On the other hand, nonessential amino acids are synthesized in our body from other amino acids. The nonessential amino acids are arginine, alanine, asparagine, aspartic acid, cysteine, glutamine, glutamic acid, glycine, proline, serine, and tyrosine. In the form of skin, hair, callus, cartilage, muscles, tendons, and ligaments, proteins hold together, protect, and provide structure to the body of a multicelled organism. In the form of enzymes, hormones, antibodies, and globulins, they catalyze, regulate, and protect the body chemistry. In the form of hemoglobin, myoglobin, and various lipoproteins, they effect the transport of oxygen and other substances within an organism.

1.3 Bioinformatics Tasks for Biological Data

This section presents the major biological problems and associated tasks involved in computational biology and bioinformatics.

1.3.1 Alignment and Comparison of DNA, RNA, and Protein Sequences

An alignment is a mutual placement of two or more sequences which exhibit where the sequences are similar, and where they differ. These include alignment and prediction of DNA, RNA, protein sequences, and fragment assembly of DNA. An optimal alignment is the one that exhibits the most correspondences and the fewest differences. There are mainly two types of alignment methods, namely, global alignment and local alignment. Global alignment [239] maximizes the number of matches between the sequences along the entire length of the sequence, while local alignment [325] gives a highest scoring to local match between two sequences. Global alignment includes all the characters in both sequences from one end to the other, and is excellent for sequences that are known to be very similar. If the sequences being compared are not similar over their entire lengths, but have short stretches within them that have high levels of similarity, a global alignment may miss the alignment of these important regions, and local alignment is then used to find these internal regions of high similarity.

Pairwise comparison and alignment of protein or nucleic acid sequences is the foundation upon which most other bioinformatics tools are built. Dynamic programming is an algorithm that allows for efficient and complete comparison of two or more biological sequences, and the technique is known as the Smith–Waterman algorithm [325]. It refers to a programmatic technique or algorithm which, when implemented correctly, effectively makes all possible pairwise comparisons between the characters (nucleotide or amino acid residues) in two biological sequences. Spaces may need to be inserted within the sequences for alignment. Consecutive space is defined as a gap. The final result is a mathematical, but not necessarily biological, optimal alignment of the two sequences. A similarity score is also generated to describe how similar the two sequences are, given the specific parameters used. A few of the many popular alignment techniques are BLAST [7], FASTA [272], and PSI-BLAST [8].

A multiple alignment [242] arranges a set of sequences in a manner that positions homologous sequences in a common column. There are different conventions regarding the scoring of a multiple alignment. In one approach, the scores of all the induced pairwise alignments contained in a multiple alignment are simply added. For a linear gap penalty, this amounts to scoring each column of the alignment by the sum of pair scores in this column [308]. Although it would be biologically meaningful, the distinctions between global, local, and other forms of alignment are rarely made in a multiple alignment. A full set of optimal pairwise alignments among a given set of sequences will generally overdetermine the multiple alignment. If one wishes to assemble a multiple alignment from pairwise alignments, one has to avoid closing loops, that is, one can put together pairwise alignments as long as no new pairwise alignment is included to a set of sequences which is already part of the multiple alignment.

1.3.2 Identification of Genes and Functional Sites from DNA Sequences

Gene finding is concerned with identifying stretches of sequence, usually genomic DNA, that are biologically functional. This especially includes identification of protein coding genes, but may also include identification of other functional elements such as noncoding RNA genes and regulatory regions. Since in human body the protein coding regions account for only a few percent of the total genomic sequence, identifying protein coding genes within large regions of uncharacterized DNA is a difficult task. In bacterial DNA, each protein is encoded by a contiguous fragment called an open reading frame, beginning with a start codon and ending with a stop codon. In eukaryotes, especially in vertebrates, the coding region is split into several fragments called exons, and the intervening fragments are called introns. So, finding eukaryotic protein coding genes in uncharacterized DNA sequences is essentially predicting exon–intron structures. Different works related to identification of protein coding genes are discussed in [99, 101, 102, 348].

Another important problem in bioinformatics is the identification of several functional sites in genomic DNA such as splice sites or junctions, start and stop codons, branch points, promoters and terminators of transcription, polyadenylation sites, topoisomerase II binding sites, topoisomerase I cleavage sites, and various transcription factor-binding sites. Such local sites are called signals, and the methods for detecting them are called signal sensors. Genomic DNA signals can be contrasted with extended and variable length regions such as exons and introns, which are recognized by different methods called content sensors. Identification of splice sites, introns, exons, start and stop codons, and branch points constitutes the major subtask in gene prediction and is of key importance in determining the exact structure of genes in genomic sequences.

In order to study gene regulation and have a better interpretation of microarray expression data, promoter prediction, and transcription factor-binding site's (TFBS) discovery have become important. A cell mechanism recognizes the beginning of a gene or gene cluster with the help of a promoter and is necessary for the initiation of transcription. The promoter is a region before each gene in the DNA that serves as an indication to the cellular mechanism that a gene is ahead. There exist a number of approaches that find differences between sets of known promoter and nonpromoter sequences [171, 189]. Due to the lack of robust protein coding signatures, current promoter predictions are much less reliable than protein coding region predictions. Once regulatory regions, such as promoters, are obtained, finding the TFBS motifs within these regions may proceed either by enumeration or by alignment to find the enriched motifs. Recognition of regulatory sites in DNA fragments has become particularly popular because of the increasing number of completely sequenced genomes and mass application of DNA chips. Experimental analyses have identified fewer than 10 % of the potential promoter regions, assuming that there are at least 30,000 promoters in the human genome, one for each gene.

1.3.3 Prediction of Protein Functional Sites

The prediction of functional sites in proteins is another important problem in bioinformatics. It is an important issue in protein function studies and hence, drug design. The problem of functional sites prediction deals with the subsequences; each subsequence is obtained through moving a fixed length sliding window residue by residue. The residues within a scan form a subsequence. If there is a match between a subsequence and a consensus pattern of a specific function, a functional site is then identified within the subsequence or the subsequence is labeled as functional, otherwise nonfunctional. To analyze protein sequences, BLAST [7], FASTA [272], PSI-BLAST [8], suffix-tree based algorithms [4], regular expression matching representations [337], and finite state machines [304, 305] are a few of the many pattern recognition algorithms that use characters or strings as their primitive type.

However, it has been found that the relation between functional sites and consensus patterns may not be always simple and the development and the use of more complicated and hence, more powerful pattern recognition algorithms is a necessity. The artificial neural networks trained with backpropagation [55, 236, 280], Kohonen's self-organizing map [13], feedforward and recurrent neural networks [19, 20], biobasis function neural networks [38, 338, 376, 378–380], and support vector machine [56, 226, 375] have been widely used to predict different functional sites in proteins such as protease cleavage sites of HIV (human immunodeficiency virus) and Hepatitis C Virus, linkage sites of glycoprotein, enzyme active sites, post-translational phosphorylation sites, immunological domains, Trypsin cleavage sites, protein–protein interaction sites, and so forth.

1.3.4 DNA and RNA Structure Prediction

DNA structure plays an important role in a variety of biological processes. Different dinucleotide and trinucleotide scales have been described to capture various aspects of DNA structure including base stacking energy, propeller twist angle, protein deformability, bendability, and position preference [19]. Three dimensional DNA structure and its organization into chromatin fibers are essential for its functions, and are applied in protein binding sites, gene regulation, and triplet repeat expansion diseases.

An RNA molecule is considered as a string of n characters $R = r_1 r_2 \cdots r_n$ such that $r_i \in \{\mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{U}\}$. Typically, n is in the hundreds, but could also be in thousands. The secondary structure of the RNA molecule is a collection S of a set of stems and each stem consisting of a set of consecutive base pairs $(r_i r_j)$ (for example, **GU**, **GC**, **AU**). Here, $1 \leq i \leq j \leq n$ and $(r_i$ and $r_j)$ are connected through hydrogen bonds. If $(r_i, r_j) \in S$, in principle, we should require that r_i be a complement to r_j and that $j - i > t$, for a certain threshold t as it is known that an RNA molecule does not fold too sharply on itself.

Attempts to automatically predict the RNA secondary structure can be divided in essentially two general approaches. The first involves the overall free energy minimization by adding contributions from each base pair, bulged base, loop, and other elements [1]. The second type of approach [360] is more empirical and it involves searching for the combination of nonexclusive helices with a maximum number of base pairings, satisfying the condition of a tree-like structure for the biomolecule. Within the latter, methods using dynamic programming are the most common [360, 395]. The methods for simulating the folding pathway of an RNA molecule [312, 313, 366] and locating significant intermediate states are important for the prediction of RNA structure [29, 127, 311] and its associated function.

1.3.5 Protein Structure Prediction and Classification

Identical protein sequences result in identical 3D structures. So, it follows that similar sequences may result in similar structures, and this is usually the case. However, identical 3D structures do not necessarily indicate identical sequences as there is a distinction between homology and similarity. There are a few examples of proteins in the databases that have nearly identical 3D structures, and are therefore homologous, but do not exhibit significant or detectable sequence similarity. Pairwise comparisons do not readily show positions that are conserved among a whole set of sequences and tend to miss subtle similarities that become visible when observed simultaneously among many sequences. Hence, one wants to simultaneously compare several sequences. Structural genomics is the prediction of the 3D structure of a protein from the primary amino acid sequence [21, 60, 70, 73, 112, 128, 150, 166, 175, 219, 220, 245, 268, 280, 287, 294, 295, 297, 329]. This is one of the most challenging tasks in bioinformatics as a protein's function is a consequence of its structure.

There are five levels of protein structure. While the primary structure is the sequence of amino acids that compose the protein, the secondary structure of a protein is the spatial arrangement of the atoms constituting the main protein backbone. The supersecondary structure or motif is the local folding pattern built up from particular secondary structures. On the other hand, tertiary structure is formed by packing secondary structural elements linked by loops and turns into one or several compact globular units called domains, that is, the folding of the entire protein chain. A final protein may contain several protein subunits arranged in a quaternary structure.

Protein sequences almost always fold into the same structure in the same environment. Hydrophobic interaction, hydrogen bonding, electrostatic, and other van der Waals type interactions also contribute to determine the structure of the protein. Many efforts are underway to predict the structure of a protein, given its primary sequence. A typical computation of protein folding would require computing all the spatial coordinates of atoms in a protein molecule, starting with an initial configuration and working up to a final minimum-energy folding configuration [31, 74, 174, 176, 273, 284, 303, 349]. Sequence similarity methods can predict the secondary and tertiary structures based on homology to known proteins. Secondary structure prediction methods include the methods proposed by Chou and Fasmann [70], and Garnier et al. [112]. Artificial neural networks [280, 287] and nearest neighbor methods [294, 295] are also used for this purpose. Tertiary structure prediction methods [349] are based on energy minimization, molecular dynamics, and stochastic searches of conformational space.

1.3.6 Molecular Design and Molecular Docking

When two molecules are in close proximity, it can be energetically favorable for them to bind together tightly. The molecular docking problem is the prediction of energy and physical configuration of binding between two molecules. A typical application is in drug design, in which one might dock a small molecule that is a described drug to an enzyme one wishes to target. For example, HIV protease is an enzyme in the AIDS virus that is essential to its replication. The chemical action of the protease takes place at a localized active site on its surface. HIV protease inhibitor drugs are small molecules that bind to the active site in HIV protease and stay there, so that the normal functioning of the enzyme is prevented. Docking software allows us to evaluate a drug design by predicting whether it will be successful in binding tightly to the active site in the enzyme. Based on the success of docking, and the resulting docked configuration, designers can refine the drug molecule [63, 188, 232, 374].

On the other hand, quantitative structure–activity relationship deals with establishing a mathematical correlation between calculated properties of molecules and their experimentally determined biological activity. These relationships may further