

PREFACE

Over the past few decades, major advances in the field of molecular biology, coupled with advances in genomic technologies, have led to an explosive growth in the biological information generated by the scientific community. This deluge of genomic information has, in turn, led to a requirement for computerized databases to store, organize and to index the data, and there is pressure for specialized tools to view and analyze the data. Bioinformatics, a field devoted to the interpretation and analysis of biological data using computational techniques, has evolved in response to this need. It is an interdisciplinary field involving biology, computer science, mathematics and statistics to analyze biological sequence data, genome content & arrangement, and to predict the function and structure of macromolecules. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be derived.

Soft computing is a consortium of methodologies that work synergistically and provides, in one form or another, flexible information processing capabilities for handling real life ambiguous situations. Its aim, unlike conventional (hard) computing, is to exploit the tolerance for imprecision, uncertainty, approximate reasoning and partial truth in order to achieve tractability, robustness, low solution cost, and close resemblance with human like decision making. At this juncture, fuzzy sets (FS), artificial neural networks (ANN), evolutionary algorithms (EAs) (including genetic algorithms (GAs), genetic programming (GP), evolutionary strategies (ES)), support vector machines (SVM), wavelets, rough sets (RS), simulated annealing (SA), particle swarm optimization (PSO), memetic algorithms (MA), ant colony optimization (ACO), tabu search (TS) and chaos theory are the major components of soft computing.

The different tasks involved in the analysis of biological data include sequence alignment, genomics, proteomics, DNA and protein structure pre-

diction, gene/promoter identification, phylogenetic analysis, analysis of gene expression data, metabolic pathways, gene regulatory networks, protein folding, docking, and molecule and drug design. Data analysis tools used earlier in bioinformatics were mainly based on statistical techniques like regression and estimation. Role of soft computing in bioinformatics gained significance with the need of handling large, complex, inherently uncertain, data sets in biology in a robust and computationally efficient manner.

Much of the biological data are inherently uncertain and noisy, thus making fuzzy sets a natural framework for analyzing them. The learning capability of neural nets both in the supervised and unsupervised domains can be utilized effectively when extracting patterns from large datasets. This is particularly true in data-rich environments as in the case of biological data. Most of the bioinformatic tasks involve search and optimization of different criteria (like energy, alignment score, overlap strength), while requiring robust, fast and close approximate solutions. Evolutionary and other search algorithms like TS, SA, ACO, PSO etc. provide powerful searching methods to explore huge and multi-modal solution spaces. Moreover, since many of the problems involve multiple conflicting objectives, application of multi-objective optimization algorithms like multi-objective genetic algorithms appears to be natural and appropriate.

This book is an attempt to bring together research articles by eminent scientists and active practitioners reporting recent advances in integrating soft computing techniques, either individually or in an hybridized manner, for analyzing biological data in order to extract more and more meaningful information and insights from them. Biological data to be considered for analysis include sequence data, structure data, and microarray data. These data types are typically complex in nature, and require advanced methods to deal with them. Characteristics of the methods and algorithms reported here include the use of domain-specific knowledge for reducing the search space, dealing with uncertainty, partial truth and imprecision, efficient linear and/or sub-linear scalability, incremental approaches to knowledge discovery, and increased level and intelligence of interactivity with human experts and decision makers.

The book has three parts. The first part provides an overview of the areas of bioinformatics and soft computing. The second part deals with applica-

tions of different soft computing techniques for sequence and structure analysis. The last part deals with different studies involving gene expression data.

In Chapter 1, Tang and Kim provide an overview of bioinformatics with special reference to the task of mining massive amount of data from high throughput genomic experiments. They discuss about the classical tasks in bioinformatics and their recent developments. These include sequence alignment, genome sequencing and fragment assembly, gene annotation, RNA folding, motif finding and protein structure prediction. Among the emerging topics resulting from new genome technologies, they discuss about comparative genomics, pathway reconstruction, microarray analysis, proteomics and protein-protein interaction.

In Chapter 2, Konar and Das present a lucid overview of the soft computing paradigm. They discuss in detail about the scope of soft computing to overcome the limitations of traditional artificial intelligence techniques. Some of the major soft computing tools, such as fuzzy logic, neural networks, evolutionary algorithms and probabilistic reasoning, are introduced. Merits of hybridizing these techniques are mentioned along with a discussion on some popular hybridizations namely, neuro-fuzzy, neuro-genetic, fuzzy-genetic etc. Finally some emerging areas of soft computing like artificial life, particle swarm optimization, artificial immune system, rough sets and granular computing, chaos theory and ant colony systems are discussed.

Gallardo, Cotta and Fernández consider the problem of inferring a phylogenetic tree given genomic, proteomic, or even morphological data in Chapter 3. Although the classical approaches for solving this problem are inherently limited, given the computational hardness of this problem, they can be useful when used in combination with other meta heuristic search techniques. Such a model, hybridizing memetic algorithms and branch-and-bound techniques, is described in this chapter.

In Chapter 4, Wang and Wu tackle the problem of RNA classification using support vector machines. First, a review of the recent advances in this field is presented. Thereafter, they present a new kernel that takes advantage of both global and local structural information in RNAs and uses the information to classify RNAs. A part of the kernel is based on recurring substrings from RNA molecules while the other part is based on counting

bi-grams in RNA molecules. Experimental results using nine families of non-coding RNA sequences taken from the Rfam database demonstrate the good performance of the new kernel and show that it outperforms existing kernels when applied to classifying non-coding RNA sequences.

Wavelet transform is a powerful signal processing tool that can analyze the data in multiple resolutions. Recently there has been a growing interest in using wavelet transforms in the analysis of biological sequences and biology related signals. In Chapter 5, Krishnan and Li review the applications of wavelet transforms for motif search, sequence comparison, protein secondary structure prediction, detection of transmembrane proteins and hydrophobic cores, mass spectrometry and disease related applications.

Moving from sequence information to surface information, a system for surface motif extraction from protein molecular surface data, called SUMOMO, is proposed by Shrestha and Ohkawa in Chapter 6. Since surface motifs cannot be assigned predetermined shapes and sizes, to extract surface motifs of different sizes, a given set of protein molecular surfaces is divided into several small surfaces collectively called unit surfaces. A filtering process is then applied based on the fact that active sites from proteins having a particular function have similar shape and physical properties. Subsequently, negative instances of active sites are used to further reduce the number of possible active site candidates.

In Chapter 7, Pollastri, Baú and Vullo present a simple and effective scalable architecture called Distill for *ab initio* prediction of protein C_α traces based on predicted structural features. It uses a set of state-of-the-art predictors of protein features based on machine learning techniques and trained on large, non-redundant subsets of the PDB, and a simple and fast 3D reconstruction algorithm guided by a pseudo-energy defined according to these predicted features. The reconstruction algorithm employs simulated annealing in its search phase. Results show that Distill can generate topologically correct predictions for a significant fraction of short proteins with 150 or fewer residues.

In Chapter 8, Bandyopadhyay, Santra, Maulik and Muehlenbein deal with the problem of using evolutionary computation techniques for designing small ligands that can bind to the active site of a target protein, there by inhibiting its function. The proposed method uses a variable string

length genetic algorithm to encode a tree-shaped ligand constructed using functional groups from a given library. The size of the tree is kept variable. Results on four proteins demonstrate the superiority of the method as compared with some earlier attempts in this direction.

The following five chapters deal with applications of soft computing to different problems related to microarray data analysis. In Chapter 9, Noman and Iba tackle the task of reconstructing genetic network from expression profile by using an improved evolutionary algorithm. The method is tested on simulated data, and is also used to analyze microarray data for predicting the interaction among the genes in SOS DNA repair network in *Escherichia coli*.

In Chapter 10, Deb, Reddy and Chaudhuri model the task of classifying gene expression data by identifying a relevant subset of the genes as one of multi-objective optimization. The minimizing criteria are the classifier size and the number of misclassified instances in training and test samples. A multi-objective evolutionary algorithm (EAs) is applied as the underlying optimization technique. The standard weighted voting method is used to design a unified procedure for handling two and multi-class problems. The use of multi-objective EAs here is unique in finding multiple high-performing classifiers in a single simulation run. The designed classifier is used to classify three two-class cancer data sets, Leukemia, Lymphoma, and Colon. Using the multi-objective genetic algorithm NSGA-II for this task, the authors report much higher accuracies on several data sets as compared to previous studies.

A similar study is reported in Chapter 11, where Gupta, Jayaraman and Kulkarni employ ant colony optimization for performing feature selection for classification of microarray data. Support vector machine is used as the underlying classification method. Results again demonstrate the effectiveness of the application of soft computing techniques to this problem.

It is observed in Chapter 12 that since microarray data can be noisy and incomplete, selected features with feature selection methods can be incomplete. Moreover, no one classification algorithm can be perfect for several data sets. To solve this problem, Cho and Park propose an ensemble of three methods for classifying gene expression data. The first method uses negatively correlated gene subsets and combines their results with Bayesian approach. The second one uses combinatorial ensemble approach based

on elementary single classifiers, and the last one searches the optimal pair of feature-classifier ensemble with genetic algorithm. Results are demonstrated on lymphoma and colon data sets.

Finally, Chapter 13 deals with the task of clustering microarray data using a fuzzy partitioning method. Here, Mukhopadhyay, Maulik and Bandyopadhyay use a novel multi-objective clustering algorithm for this purpose. The clustering problem is posed as one of optimization of two different fuzzy cluster validity indices, namely, the Xie-Beni index and the FCM-index J_m . NSGA-II is used as the underlying multi-objective optimization strategy. Comparison with a single objective version of the problem, and the widely used K-means, K-medoids, fuzzy C-means and hierarchical clustering methods demonstrate the superiority of the proposed method.

In summary, the chapters on the applications of soft computing techniques for analyzing biological data provide a representative selection of the available methods and their evaluation in real domains. While the field is rapidly evolving with the availability of new data and new tools, these chapters clearly indicate the importance and potential benefit of synergetically combining the potentials of classical and soft computing methods for facing the newer challenges in biological data mining. The book will be useful to graduate students and researchers in computer science, bioinformatics, computational and molecular biology, electrical engineering, system science, and information technology both as text and reference book for some parts of the curriculum. The researchers and practitioners in industry and R & D laboratories will also be benefited.

We take this opportunity to thank all the authors for contributing chapters related to their current research work that provide the state of the art in advanced methods for analyzing biological data. We are grateful to Ms. Yubing Zhai of World Scientific Publishing Co. Pte. Ltd. for her initiative and constant support.

Sanghamitra Bandyopadhyay

Ujjwal Maulik

Jason T. L. Wang

August, 2006