

Preface

Computational Intelligence has been an active area of research from which a great potential has emerged over the past decade. There are numerous successful applications of Computational Intelligence in various subfields of biology, including bioinformatics, computational genomics, protein structure prediction, biochemical pathways, microarray data analysis and neuronal systems modeling and analysis. However, there are several open problems in biology which require advanced and efficient computational methodologies. The major issue is the voluminous data that constitute the above mentioned problems that can be reduced and analyzed with the help of automated tools. There is lots of information hidden inside this voluminous data so it is essential for a new generation of computational theories and tools to assist humans in extracting knowledge from big data. Unfortunately, researchers in the field of biology are often unaware of the abundance of computational techniques that they can be used to analyze and understand the data underlying their research interest/inquiries. On the other hand, computational intelligence or big data practitioners are often unfamiliar with the particular problems that their new, state-of-the-art algorithms can be applied to.

Computational intelligence is an established paradigm, where new theories with an adequate biological understanding have been evolving. Current experimental systems have many characteristics of biological computers and are being built to perform a variety of tasks that are difficult or impossible to do with conventional computers. Computational intelligence methods are now being applied to problems in bioinformatics and big data analytics. This area is highly heterogeneous with a combination of technologies such as neural networks, fuzzy systems, rough sets, evolutionary computing, swarm intelligence, probabilistic reasoning, multi agent systems etc. The recent trend is to integrate different components to take advantage of complimentary features and to develop a synergistic system. Bioinformatics involves the creation and advancement of algorithms using techniques including computational intelligence, applied mathematics, statistics, biochemistry and informatics to solve biological problems at molecular level.

The transformation of data into usable knowledge is by no means an easy task even for high-performance large-scale data processing, including exploiting parallelism of current and upcoming computer architectures for data mining. Moreover, this data may involve uncertainty in many different forms. Many different models, like fuzzy sets, rough sets, soft sets, neural networks, their generalizations and hybrid models obtained by combining two or more of these models have been found to be fruitful in representing data. These models are also quite useful for analysis. More often than not, big data are reduced to include only the important characteristics necessary from a particular study point of view or depending upon the application area. So, reduction techniques have been developed. Often the collected data have missing values. These values need to be generated or the tuples with missing values are eliminated from the data set before analysis. More importantly, these new challenges may compromise; sometimes even

deteriorate the performance, efficiency and scalability of the dedicated data intensive computing systems. The later approach sometimes leads to loss of information and hence is not preferred. This brings up many research issues in biomedical and healthcare research community in the form of capturing and accessing data effectively. In addition, fast processing while achieving high performance and high throughput, and storing it efficiently for future use is another issue.

Bioinformatics, computational intelligence, knowledge discovery in databases, cloud computing, parallel and distributed computing, network security, data mining, knowledge representation are the fields that have evolved into an importance and active area of research because of theoretical challenges associated with the problems of discovering intelligent solutions for voluminous data/ Big data.

The bioinformatics fields are in a stage of new breakthroughs in dealing with dramatically increased information in terms of number, size and complexity. A new term "Big Data" is used to refer to such large and often diverse, complex, longitudinal and distributed datasets. Big data is usually too large and complex to be processed using traditional data processing applications. It is not so much that the data cannot be processed using traditional tools, but it cannot be processed in reasonable time without newer and faster processing tools. There is and will continue to be an increasing amount of research in this area. As (Shah N. H., 2012) stated, "Data centric approaches that compute on massive amounts of data often referred "Big Data", to discover patterns and to make clinically relevant predictions will gain adoption. Major research efforts in the field include sequence alignment, gene finding, genome assembly, protein structure alignment, protein structure prediction, prediction of gene expression and protein-protein interactions, and the modeling of evolution. Hence, in other words, bioinformatics can be described as the application of computational methods to make biological discoveries (Baldi, P., Brunak, S., 1998). The ultimate goal of the 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 (Baldi, P., Brunak, S., 1998). There are at least 26 billion base pairs (bp) representing the various genomes available on the server of the National Center for Biotechnology Information (NCBI) (Ezziane, Z., 2006). Besides the human genome with about 3 billion bp, many other species have their complete genome available there. Cohen (2004) explained the needs of biologists to utilize and help interpret the vast amounts of data that are constantly being gathered in genomic research. He also pointed out the basic concepts in molecular cell biology, and outlined the nature of the existing data, and illustrated the algorithms needed to understand cell behavior.

Bioinformatics and computational biology are concerned with the use of computation to understand biological phenomena and to acquire and exploit biological data, increasingly large-scale data. Methods from bioinformatics and computational biology are increasingly used to augment or leverage traditional laboratory and observation-based biology. This transformation from a data-poor to a data-rich field began with DNA sequence data, but is now occurring in many other areas of biology (Cohen, 2004). Computational intelligence methods are now being applied to problems in molecular biology and bioinformatics (Baldi, P., Brunak, S., 1998). To name a few, (Tasoulis et al., 2008) present an application of neural networks, evolutionary algorithms, and clustering algorithms to DNA microarray experimental data analysis; Liang and Kelemen (2008) propose a time lagged recurrent neural network with trajectory learning for identifying and classifying gene functional patterns from the heterogeneous nonlinear time series microarray experiments. In a nutshell, which becomes quite apparent in light of the current research pursuits, the area is heterogeneous with a combination of such technologies as neural networks, fuzzy systems, evolutionary computation, swarm intelligence, and probabilistic reasoning. The recent

trend is to integrate different components to take advantage of complementary features and to develop a synergistic system. Hybrid architectures like neuro-fuzzy systems, evolutionary-fuzzy systems, evolutionary neural networks, evolutionary neuro-fuzzy systems, rough-neural, rough-fuzzy, etc. are widely applied for real world problem solving.

COMPUTATIONAL INTELLIGENCE TECHNIQUES

Progress in machine learning technology provided various advanced tools for predictions of complex biological problems. Among the many machine learning approaches, artificial neural networks, fuzzy sets, particle swarm optimization, genetic algorithms, and rough sets are the most recently applied computational intelligent tools in the structure prediction.

Artificial Neural Networks: Artificial neural networks have been developed as generalizations of mathematical models of biological nervous systems. In a simplified mathematical model of the neuron, synapses are represented by connection weights that modulate the effect of the associated input signals, and the nonlinear characteristic exhibited by neurons is represented by a transfer function. The neuron impulse is computed as the weighted sum of the input signals, transformed by the transfer function (Bishop, C.M., 1995). The learning capability of an artificial neuron is achieved by adjusting the weights in accordance to the chosen learning algorithm. The basic architecture consists of three types of neuron layers: input, hidden, and output layers. In feed-forward networks the signal flow is from input to output units strictly in a feed-forward direction. The data processing can extend over multiple (layers of) units, but no feedback connections are present, that is, connections extending from outputs of units to inputs of units in the same layer or previous layers. Recurrent networks contain feedback connections. Contrary to feed-forward networks, the dynamical properties of such networks are important.

There are several other neural network architectures (Elman network, adaptive resonance theory maps, competitive networks etc.) depending on the properties and requirement of the application. The learning of the network may be classified into three distinct types. These are supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, an input vector is presented at the input layer together with a set of desired responses, one for each node, at the output layer. A forward pass is done and the errors or discrepancies, between the desired and actual response for each node in the output layer, are found. These are then used to determine weight changes in the network according to the prevailing learning rule. The best-known examples of this technique occur in the back-propagation algorithm, the delta rule, and perceptron rule. In unsupervised learning (or self-organization) an output unit is trained to respond to clusters of patterns within the input. In this paradigm the system is supposed to discover statistically salient features of the input population. Reinforcement learning is learning what to do—how to map situations to actions—so as to maximize a numerical reward signal. These two characteristics, trial-and-error search and delayed reward are the two most important distinguishing features of reinforcement learning.

Rough Sets: Rough set theory is a methodology fairly new to the medical domain capable of dealing with uncertainty in data. It is used to discover data dependencies, evaluate the importance of attributes, discover the patterns of data, reduce redundant objects and attributes, seek the minimum subset of attributes, recognize and classify objects. Moreover, it is being used for extraction of rules from databases. Rough sets have proven useful for representation of vague regions in spatial data. One advantage of rough sets is the creation of readable if-then rules. Such rules have a potential to reveal new patterns

in the data material. Furthermore, they also collectively function as a classifier for unseen data. Unlike other computational intelligence techniques, rough set analysis requires no external parameters and uses only the information presented in the given data.

In rough sets theory, the data is collected in a table, called decision table. Rows of the decision table correspond to objects, and columns correspond to attributes. In the data set, we assume that class labels to indicate the class to which each example belongs are given. We call the class label the decision attribute and the rest of the attributes the condition attributes. Rough sets theory defines three regions based on the equivalent classes induced by the attribute values Lower approximation, upper approximation, and the boundary. Lower approximation contains all the objects which are classified surely based on the data collected, and upper approximation contains all the objects which can be classified probably, while the boundary is the difference between the upper approximation and the lower approximation (Pawlak, Z., Grzymala-Busse, J., Slowinski, R., Ziarko, W., 1995). An indiscernibility relation partitions the set of cases or objects into a number of equivalence classes. An equivalence class of a particular object is simply the collection of objects that are indiscernible to the object in question (Pawlak, Z., 1982, 1991).

Fuzzy Logic (FL) and Fuzzy Sets (FS): (Zadeh, L.A., 1965) introduced the concept of fuzzy logic to present vagueness in linguistics, and further implement and express human knowledge and inference capability in a natural way. Fuzzy logic starts with the concept of a fuzzy set. An FS set is a set without a crisp, clearly defined boundary. It can contain elements with only a partial degree of membership. A Membership Function (MF) is a curve that defines how each point in the input space is mapped to a membership value (or degree of membership) between 0 and 1. The input space is sometimes referred to as the universe of discourse. Let X be the universe of discourse and x be a generic element of X . A classical set A is defined as a collection of elements or objects $x \in X$, such that each x can either belong to or not belong to the set A , $A \subseteq X$. By defining a characteristic function (or membership function) on each element x in X , a classical set A can be represented by a set of ordered pairs $(x, 0)$ or $(x, 1)$, where 1 indicates membership and 0 non-membership. Unlike conventional set mentioned above, fuzzy set expresses the degree to which an element belongs to a set. Hence the characteristic function of a fuzzy set is allowed to have value between 0 and 1, denoting the degree of membership of an element in a given set.

Evolutionary Algorithms: Evolutionary Algorithms are adaptive methods, which may be used to solve search and optimization problems, based on the genetic processes of biological organisms. Over many generations, natural populations evolve according to the principles of natural selection and "survival of the fittest," first clearly stated by Charles Darwin in *The Origin of Species*. By mimicking this process, evolutionary algorithms are able to 'evolve' solutions to real world problems, if they have been suitably encoded. Usually grouped under the term Evolutionary Algorithms (EA) or Evolutionary Computation (EC), we find the domains of genetic algorithms, evolution strategies, evolutionary programming, genetic programming, and learning classifier systems (Goldberg, D.E., 1989). They all share a common conceptual base of simulating the evolution of individual structures via processes of selection, mutation, and reproduction. The processes depend on the perceived performance of the individual structures as defined by the environment.

Particle Swarm Optimization: Swarm intelligence (Kennedy, J. et al., 2001) is a collective behavior of intelligent agents in decentralized systems. Although there is typically no centralized control dictating the behavior of the agents, local interactions among them often cause a global pattern to emerge. Most of the basic ideas are derived from real swarms in the nature including ant colonies, bird flocking, honeybees, bacteria and microorganisms, etc. Ant Colony Optimization (ACO), have already been applied successfully to solve several engineering optimization problems. Swarm models are population based

and the population is initialized with a set of potential solutions. These individuals are then manipulated (optimized) over much iteration using several heuristics inspired from the social behavior of insects in an effort to find the optimal solution. The concept of particle swarms, although initially introduced for simulating human social behaviors, has become very popular these days as an efficient search and optimization technique.

The objective of this volume is to disseminate inspiring research results and exemplary computational tools of data mining approaches to cross disciplinary researchers and practitioners from bioinformatics, data mining disciplines, the life sciences and healthcare domains. The book will cover fundamentals of data mining techniques designed to tackle data analysis challenges mentioned above, applications of computational intelligent tools to find optimal generalized solutions for biological problems, and demonstrate with real applications how bio-inspired algorithms can enable biologists, engineers and healthcare scientists to make insightful observations and invaluable discoveries from their data.

ORGANIZATION OF THE BOOK

This book is organized into 4 major parts, covering the following topics: Data Mining and Pattern Discovery (Section 1), Computational Intelligence in Bioinformatics (Section 2), Nature-inspired computing for analysis of DNA and Protein Microarray Data (Section 3), and Bio-Inspired Algorithms and Engineering Applications (Section 4).

Section 1: Data Mining and Pattern Discovery

Chapter 1, entitled “Unleashing Artificial Intelligence onto Big Data - A Review”, discusses mining of vast amounts of data in the order of petabytes or even exabytes using artificial intelligent tools, deriving meaningful results. Application of artificial intelligence for finding patterns in enormous quantities of data and confluence of Big Data and Artificial Intelligence allows companies to automate and improve complex descriptive, predictive and prescriptive analytical tasks. This review article focuses on applications of artificial intelligence to Big Data, its Limitations and issues. Chapter 2, entitled “Hybrid Ensemble Learning Methods for Classification of Microarray Data”, addresses the major problems of discrimination of potential genes from microarray datasets using hybrid ensemble framework. In this chapter, a hybrid RotBagg ensemble framework has been proposed to address the problem specified above. This technique is an integration of Rotation Forest and Bagging ensemble which in turn preserves the basic characteristics of ensemble architecture i.e., diversity and accuracy. Three different feature selection techniques are employed to select subsets of genes to improve the effectiveness and generalization of the RotBagg ensemble. Chapter 3, entitled “Recent trends in spatial data mining and its challenges”, discusses spatial data mining to analyze the spatial characteristics of vast amount of data used in analytics for decision making. Understanding of spatial characteristics of data will enhance the accuracy of decision. The spatial database contains numerous amounts of spatial and non-spatial data of different forms. Interpretation and analysis of vast data is far beyond human ability. This chapter briefs about the characteristics of spatial data mining and the methods of spatial data mining in recent years. Chapter 4, entitled “Knowledge Representation: A Semantic Network Approach”, introduces semantic network approach as one of the potential knowledge representation (KR) technique which is one of the hottest areas of research in data mining, AI and big data analysis. Knowledge represented in an effective

way, helps in easy traversal, search, reasoning and inference. Limitations of semantic network are also described. To overcome these limitations a modified version is proposed in this chapter that is named as Semantic Fuzzy Networks (SFN) in which the relationship between two concepts was more realistically provided by fuzzy membership function.

Section 2: Computational Intelligence in Bioinformatics

In Chapter 5, entitled “Development of Novel Multi-Objective based Model for Protein Structural Class Prediction”, concept of Chou’s pseudo amino acid feature representation method has been used to represent protein samples. Protein folding has played a vital role in rational drug design, pharmacology and many other applications. On the basis of a statistical framework a low complexity functional link artificial neural network and a complex novel hybrid model using radial basis function neural network and multi-objective algorithm based classifier are introduced to predict protein structural class. Chapter 6, entitled “Rough Fuzzy Set Theory and Neighborhood Approximation Based Modeling for Spatial Epidemiology”, introduces modern epidemiological studies to understand individual and social level inferences and their role in the transmission and distribution of disease instances. The geographic relevance in epidemiology has been analysed in concurrence with these inferences. The substantial amount of data involved in an epidemiological study is usually very large and intuitively involves missing values and uncertainty. Rough Set Theory (RST) has been used in medical informatics for ‘outcome prediction’ and ‘feature selection’. This chapter proposes the use of rough sets in conjunction with techniques like Fuzzy sets and Granular (Neighborhood Approximation) computing for the classic problem of data representation, dimensionality reduction, generation and harvest of minimal rules. Chapter 7, entitled “Applying CI in Biology through PSO”, addresses the challenges faced by conventional computing methods in dealing with real world problems by natural systems. This chapter will discuss some of the problems existing in computational biology, their contemporary solution methods followed by the use of PSO to address those problems. Along with this several applications of PSO, some relevant fields are discussed having some future research directions on this field. Chapter 8, entitled “Application of Rough Set Based Models in Medical Diagnosis”, discusses several applications of rough set theory in diagnosing medical data. An efficient Rough set theory model to capture uncertainty in data and the processing of data using rough set techniques is discussed. Rule generation is an inherent component in rough set analysis. So, medical systems which have uncertainty inherent can be handled in a better way using rough sets and its variants. In Chapter 9, entitled “Application of Uncertainty Models in Bioinformatics”, information related to the researches enhanced using uncertainty models in life sciences and biomedical Informatics is discussed. The main emphasis of this paper is to present the general ideas for the time line of different uncertainty models to handle uncertain information and their applications in the various fields of biology.

Section 3: Nature-Inspired Computing for Analysis of DNA and Protein Microarray Data

In Chapter 10, entitled “Computational Methods for Prediction of Protein-Protein Interactions: PPI Prediction Methods”, various computational approaches for predicting protein-protein interactions are discussed. Methods like 1) Genomic information based methods 2) Structure based methods 3) Network topology based methods: 4) Literature and data mining based methods 5) Machine learning methods

are employed in this chapter to predict the PPI. For ease of use and convenience, a summary of various databases and software for PPI prediction has been provided. Chapter 11, entitled "Analysis of Microarray Data using Artificial Intelligence Based Techniques", discusses artificial intelligence based techniques for the analysis of microarray gene expression data. Artificial neural networks, fuzzy logic, genetic algorithms, and many other nature-inspired algorithms are described for this purpose. In this chapter, we present artificial intelligence based techniques for the analysis of microarray gene expression data. Further, challenges in the field and future work direction have also been suggested. Chapter 12, entitled "Extraction of Protein Sequence Motif Information using Bio-Inspired Computing" introduces PSO K-Means clustering and bi-clustering approaches to extract the motif information. The Motif is extracted based on the structure homogeneity of the protein sequence. In this work, the clusters and bi-clusters are compared based on homogeneity and motif information extracted. Chapter 13, entitled "Study of basic concepts on the development of protein microarray - gene expression profiling: Protein microarray", introduces detail description on the fabrication, types, data analysis, and application of protein microarray technology towards gene expression profiling. The physical and biological activity of any organisms is mainly depended on the genetic information which stored in DNA. A process at which a gene gives rise to a phenotype is called as gene expression. Protein microarray mainly used to identify the interactions and activities of proteins with other molecules, and to determine their function for a system at normal state and stressed state. Chapter 14, entitled "Personalized Medicine in the Era of Genomics", introduces a model that aims at customizing healthcare and tailoring medicine according to an individual's genetic makeup. It classifies individuals that differ in their susceptibility to a particular disease or response to a particular treatment into subpopulations based on individual's unique genetic and clinical information along with environmental factors. This chapter aims at acquainting readers with the underlying concepts and components of personalized medicine supplemented with some disease-based case studies, discussing challenges and recent advancements in the implementation of the model of personalized medicine.

Section 4: Bio-Inspired Algorithms and Engineering Applications

In Chapter 15, entitled "Evolutionary Computing Approaches to System Identification", discusses an important class of engineering problem called system identification which is an essential requirement for obtaining models of system of concern that would be necessary for controlling, analyzing the systems. The chapter explain what are different evolutionary computing techniques used in the past and the state-of-the-art technologies on evolutionary computation. Then, some case studies have been included how the system identification of a number of complex systems effectively achieved by employing these evolutionary computing techniques. Chapter 16, entitled "BFO Optimized automatic load frequency control of a multi area power system", presents an analysis on operation of Automatic Load Frequency Control (ALFC) by developing models in SIMULINK which helps us to understand the principle behind ALFC including the challenges. The three area system is being taken into account considering several important parameters of ALFC like integral controller gains (K_i), governor speed regulation parameters (R_i), and frequency bias parameters (B_i), which are being optimized by using Bacteria Foraging Optimization Algorithm (BFOA). This will help the power industries for easier and cheaper realization of the governor. The performance of BFOA is also investigated through the convergence characteristics which reveal that the Bacteria Foraging Algorithm is quite faster in optimization such that there is reduction in the computational burden and also minimal use of computer resource utilization. Chapter

17, entitled "Solution of some Differential Equation in Fuzzy Environment by Extension Principle method and its application in Biomathematics", solves two mathematical biological models in fuzzy environment. One species prey predator model is considered with fuzzy initial data. The solution procedures of the fuzzy differential equation are taken as extension principle method. Chapter 18, entitled "Application of Computational Intelligence Techniques in Wireless Sensor Networks: The State of the Art", discusses application of computational intelligence techniques in wireless sensor networks on the coverage problem in general and area coverage in particular. After providing different types of coverage encountered in WSN, we present a possible classification of coverage algorithms. Then we dwell on area coverage which is widely studied due to its importance.

REFERENCES

- Altman, R. B., Valencia, A., Miyano, S., & Ranganathan, S. (2001). Challenges for intelligent systems in biology. *IEEE Intelligent Systems*, 16(6), 14–20. doi:10.1109/5254.972065
- Baldi, P., & Brunak, S. (1998). *Bioinformatics: The Machine Learning Approach*. Cambridge: MIT Press.
- Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.
- Cohen, J. (2004). Bioinformatics: An introduction for computer scientists. *ACM Computing Surveys*, 36(2), 122–158. doi:10.1145/1031120.1031122
- Ezziane, Z. (2006). Applications of artificial intelligence in bioinformatics: A review. *Expert Systems with Applications*, 30(1), 2–10. doi:10.1016/j.eswa.2005.09.042
- Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading: Addison-Wesley Publishing.
- Gusfield, D. (2004). Introduction to the IEEE/ACM transactions on Computational Biology and Bioinformatics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1), 2–3. doi:10.1109/TCBB.2004.9
- Kennedy, J., Eberhart, R., & Shi, Y. (2001). *Swarm Intelligence*. San Francisco: Morgan Kaufmann Academic Press.
- Liang, Y., & Kelemen, A. (2008). Time course gene expression classification with time lagged recurrent neural network. *Computational Intelligence in Bioinformatics. Studies in Computational Intelligence*, 94, 149–163. doi:10.1007/978-3-540-76803-6_5
- Mitra, S., & Hayashi, Y. (2006). Bioinformatics with soft computing. *IEEE Transactions on Systems, Man and Cybernetics. Part C, Applications and Reviews*, 36(5), 616–635. doi:10.1109/TSMCC.2006.879384
- Pawlak, Z. (1982). Rough sets. *Intl. J. Comp. Inform. Science*, 11(5), 341–356. doi:10.1007/BF01001956
- Pawlak, Z. (1991). *Rough Sets – Theoretical Aspects of Reasoning About Data*. Dordrecht: Kluwer.
- Pawlak, Z., Grzymala-Busse, J., Slowinski, R., & Ziarko, W. (1995). Rough sets. *Communications of the ACM*, 38(11), 88–95. doi:10.1145/219717.219791

Shah, N. H. (2012). Translational bioinformatics embraces big data. *Yearbook of Medical Informatics*, 7(1), 130–134.

Tasoulis, D. K., Plagianakos, V. P., & Vrahatis, M. N. (2008). Computational intelligence algorithms and DNA microarrays. *Computational Intelligence in Bioinformatics. Studies in Computational Intelligence*, 94, 1–31. doi:10.1007/978-3-540-76803-6_1

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8(3), 338–353. doi:10.1016/S0019-9958(65)90241-X