COMPUTATIONAL INTELLIGENCE AND BIOINFORMATICS

Mei Liu
Department of Computer Science, New Jersey Institute of Technology, USA

Xue-wen Chen
Department of Computer Science, Wayne State University, USA

Keywords: Computational Intelligence, Bioinformatics, Gene Expression, Multiple Sequence Alignment, Protein-protein Interaction, Protein Structure

Contents

Summary

In this chapter, we present a brief overview of bioinformatics and computational intelligence (CI) methods including artificial neural networks, genetic algorithms, and fuzzy systems. A number of representative applications of CI methods in bioinformatics are discussed, including CI methods for gene expression analysis, for multiple sequence alignment, for protein-protein interaction prediction, and for protein secondary structure prediction.

1. Introduction

The exponential growth of biological data, especially with the advent of new high-throughput technologies, has transformed the field biology into a data rich discipline. The sheer amount of biological data has created tremendous challenges for data analysis and knowledge discovery where we are faced with the complication of “data rich but information poor”. It is more important than ever to interpret the biological data efficiently and rapidly (Reichhardt, 1999). This has led to a new area known as bioinformatics in which computational algorithms are being developed to understand the biological data. Computational intelligence (CI) methodologies (e.g., artificial neural networks (ANNs), fuzzy systems, and evolutionary algorithms) are being extensively applied to solve biological problems (Fogel, et al., 2008; Seiffert, et al., 2005). In this chapter, we will briefly review CI methods (Section 2) and bioinformatics (Section 3), followed by example applications of CI methods for bioinformatics problems (Section 4) including gene expression analysis, multiple sequence alignment, protein-protein interaction prediction, and secondary structure prediction.

2. Computational Intelligence: An Overview

Computational intelligence (CI) is a branch of computer science that aims to solve complex problems that are either difficult to formulate or NP-hard. CI is often perceived as a consortium of computational methodologies that embraces neural networks, fuzzy logic and evolutionary approaches such as genetic algorithms.

2.1. Artificial neural networks (ANNs)

ANNs are biologically inspired computational models composed of many simple processing elements called artificial neurons that mimic the properties of biological neurons. ANN algorithms learn from a collective behavior of these artificial neurons and adapt to input data by altering its structure based on external or internal information that flows through the network. In an ANN, the neurons are interconnected by weighted connections or synapses and these weights contain the network knowledge. Each neuron performs limited operations and works in parallel with other neurons to solve problems quickly. A typical ANN consists of three types of layers: input, hidden, and output layers (Figure 1). The input layer is used to encode instances presented to the network for processing. The processing elements or artificial neurons in the input layer are called input nodes, which may represent an attribute or feature value of the input instance. Consequently, the number of input neurons is equal to the number of features plus one (a bias term). In the hidden layer, neurons add up the weighted input of each node from the input layer and then pass the sum to a non-linear function known as an activation or transfer function. Some of the basic and widely adopted transfer functions include radial basis function and sigmoidal function. Lastly, the output layer contains output units, which combine weighted outputs from hidden neurons and assign values to the input instance.

The behavior of a neural network largely depends on the interactions of its neurons or network architecture. There are different types of ANN for solving specific problems; for example, feed-forward neural networks, Kohonen self-organizing maps, and recurrent neural networks (RNNs). Feed-forward neural network is a common architecture where the signal flows from input to output units through multiple layers in only one direction. The most popular feed-forward networks include perceptrons, multi-layer perceptrons and radial basis networks. RNNs, on the other hand, allow feedback where connections between neurons form a directed cycle, which makes it to exhibit dynamic temporal behavior. RNNs can be useful in applications like un-segmented connected handwriting recognition (Graves, et al., 2009). There are several other ANN architectures such as Elman network, adaptive resonance theory maps, competitive networks, and etc. Researchers should decide on which ANN architecture to use based on properties and requirements of their applications.

ANNs are typically trained with training data. An ANN is characterized by the network architecture (e.g., number of layers, number of hidden neurons) and the associated parameters (e.g., connection weights). There are various methods in assigning weights to the connections. One option is to set the weights explicitly with a priori knowledge, and the other option is to learn the weights from training patterns. Three distinct learning paradigms exist: supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, training examples consisting of input vectors are analyzed along with their desired output values where a forward pass is performed and the errors between the desired and inferred outputs are calculated. The errors can in turn be used to determine weight changes of the connections according to learning rules. This technique is best illustrated by the back propagation algorithm. In contrast, unsupervised learning algorithms attempt to find hidden structures from unlabelled data where an output unit is trained on clusters of patterns within the input data. Kohonen self-organizing map is the best example for ANNs trained using unsupervised learning. Lastly, reinforcement learning is to learn what actions to take by trying them so as to maximize the cumulative reward. For an extensive review of the different ANN architectures and learning algorithms, readers may refer to (Bishop, 1995).

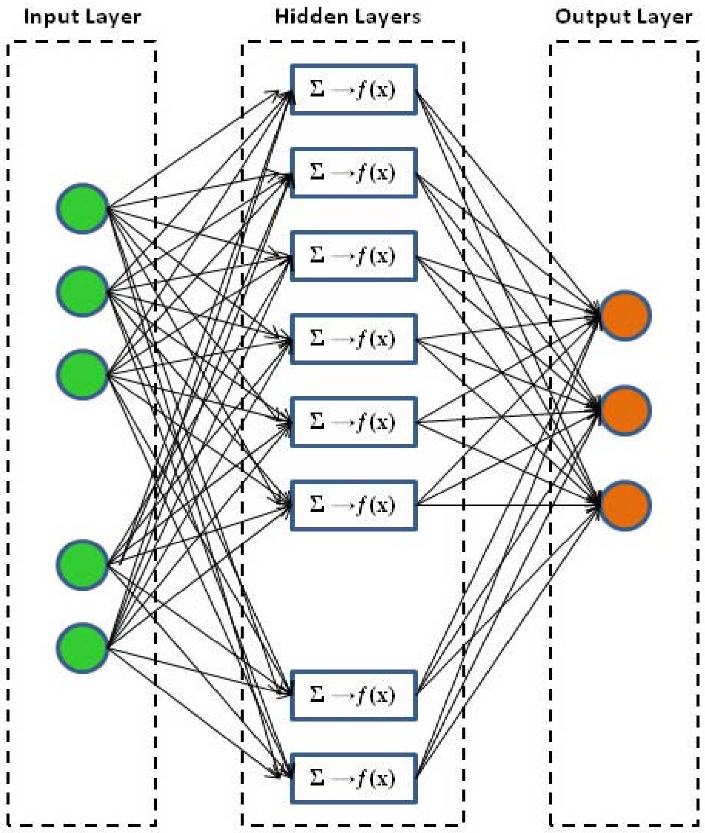

Figure 1. A typical artificial neural network (ANN) structure is composed of an input layer, hidden layer, and output layer of neurons. One or more hidden layers are needed for the non-linear transformation of the input nodes to the output nodes. On the other hand, the resulting model is linear if the input nodes are directly connected to the output nodes (no hidden layer). The neurons between layers are interconnected by weighted connections or synapses. For network optimization, not only these weights can be optimized, but also the entire topology of the ANN can be adjusted; for example, number of layers, number of nodes per layer, number of connections, and etc.

2.2. Fuzzy Logic

Fuzzy logic was first introduced by Zadeh (1965) to express vagueness in human knowledge. In contrast to the traditional binary logic theory where variables have either true or false values, fuzzy logic deals with reasoning that is approximate rather than exact, which allows a variable to have different degrees of truth that ranges between 0 and 1. It is similar to human reasoning and is especially useful in situations where clear-cut decision boundaries are not possible. For example, in case of population height, if the average height is 180cm, binary logic would determine a person of 179cm as medium height and 181cm as tall. However, in fuzzy logic, each variable such as small, medium, and tall represents a range of values that may overlap with each other. In other words, the highest values of the set 'small' can overlap with the lowest values of the set 'medium'.

Fuzzy logic can tolerate incomplete data and provide approximate solutions to problems in which other methods have difficulties. It works well with many pattern recognition problems where the classes are not precisely defined. For example, in bioinformatics, a gene's membership to a gene cluster cannot be accurately defined, definitely not by an arbitrary threshold of expression as in classical approaches. For additional information on fuzzy systems and their applications to bioinformatics, readers should refer to Bezdek and Castelaz (1977), Cox (1994, Dong, et al (2006), Keller and Tahani (1992), Mordeson, et al (2000), Szczepanniak, et al (2000), Torres and Nieto (2006) and Zimmermann (2001).

2.3. Evolutionary Computation

Evolutionary computation involves iterative processes such as growth or development of a population to solve search and optimization problems. It is based on the Darwinian principles of evolution that natural populations evolve according to natural selection and "survival of the fittest". Evolutionary computing techniques mostly entail evolutionary algorithm (EA) and swarm intelligence (SI).

Under EAs, there are genetic algorithms (GA) (Bremmerman, 1962; Holland, 1975; MIchalewicz, 1996), evolutionary programming (EP) (Fogel, et al., 1966), and evolution strategy (ES) (Rechenberg, 1973). According to Fogel (2008), EP and ES can be perceived as abstractions of the Darwinian evolution at the phenotypic level, but GA should be perceived as abstractions of evolution at the genotypic level. Nevertheless, they all share a common evolutionary mechanism: reproduction, mutation, recombination, and selection.

For instance, GA seeks optimal solution to a complex problem in a parallel fashion by using techniques that mimic the natural evolutionary process such as selection, insertion, deletion, and crossover (Goldberg, 1975; Holland, 1975; Mitra and Hayashi, 2006). The underlying idea is that the fittest candidate solutions in a population of solutions should survive and can evolve over time toward better solutions. GA usually starts with a randomly generated population of candidate solutions (called individuals). Then the fitness of every individual is assessed in the current generation and a number of individuals would be stochastically chosen based on their fitness level to either directly survive to the next generation or be modified (recombined or mutated) to produce new offspring. This evolutionary process continues until either a maximum number of generations have been generated or a satisfactory fitness level has been reached.

SI is a relatively new sub-field of evolutionary computing where the expression was coined by Beni and Wang (1989) in the context of cellular robotics. Since then, it has attracted much attention of researchers in bioinformatics related areas. SI was motivated by the collective and versatile behavior of living creatures (Bonabeau, et al., 2001; Engelbrecht, 2005) in groups such as swarms of bees, flocks of birds, colonies of ants, etc.. A prototypical example is ant colonies where the behavior of a single ant is often too simple but collectively an ant colony can effectively discover and attain food as well as adapt to rapidly changing surroundings. SI is composed of a population of simple agents (decentralized self-organized systems that are capable of executing certain operations) who can interact locally with one another and their environment to develop an intelligent global behavior in the pursuit of certain goals.

Some of the most popular SI algorithms include Ant Colony Optimization (ACO) and Particle Swarm Optimization (PSO). In ACO, ants migrate through the solution space guided by trails left by other ants in the population (Kelemen, et al., 2008). The concept of PSO was initially introduced to simulate human social behavior where the population of solutions is abstracted as a swarm of interacting particles in which each particle moves around according to its local best known position and best positions found by other particles (Kennedy, 1999).

3. Bioinformatics: An Overview

Over the past two decades, information technology has transformed biological science with the emergence of new research fields like bioinformatics. Bioinformatics is unquestionably a interdisciplinary field that involves the study of computational methods to analyze various types of biological data such as nucleotide (DNA/RNA) sequences and protein sequence, structure, function, pathways, and genetic interactions. The primary goal of bioinformatics is to discover new knowledge in biological processes through the development of efficient computational approaches. The rapid developments in genomic and molecular research technologies and information technologies have produced a tremendous amount of information. Bioinformatics supports a broad spectrum of research activities include mapping and analyzing DNA and protein sequences, comparing different sequences by aligning them, and creating 3D models of protein structures.

Definitions of the basic terms in bioinformatics are given below.

DNA - Deoxyribonucleic Acid is a double-stranded, helical molecule comprising a sequence of four bases called nucleotides – A (adenine), G (guanine), C (cytosine), and T (thymine) – in each strand. In a DNA double helix, each type of nucleotide on one strand normally interacts with just one type of nucleotide on the other strand, which is called complementary base pairing. Thus, A in one strand only bonds to T in the other, and G only bonds to C.

RNA – Ribonucleic Acid is a single-stranded molecule, like DNA, comprises of four nucleotides – A, G, C, and U (Uracil). RNA is produced from copying one of the two strands of a DNA molecule.

Codon is a length of three nucleotides in DNA that is translated by the cell as an amino acid in the protein. Among the 64 possible codons, 61 are usually read as one of the 20 amino acids, and the remaining 3 are read as stop codons indicating the end of a protein.

Protein is a molecule comprising a long chain of amino acids which is specified by the sequence of codons in a gene. In general, there are 20 standard amino acids. The chain of amino acids typically folds into a three-dimensional structure unique to each protein that facilitates biological activity.

Gene is a sequence of DNA that specifies a unit of biological function, usually the amino acid sequence of a protein.

4. Computational Intelligence in Bioinformatics

Bioinformatics aims to increase the understanding of biological systems through the development or application of efficient and intelligent algorithms. CI methods, such as ANN and GA, have been widely used for modeling knowledge in biological systems such as gene-expression analyses (Friedman, et al., 2000), multiple sequence alignment (Gondro and Kinghorn, 2007; Notredame and Higgins, 1996), protein interaction network inference (Chen and Liu, 2006; Lin, et al., 2009), protein structure prediction (Kuang, et al., 2004; Zhang, et al., 2005; Zimmermann and Hansmann, 2006), and many others. Following sections provide brief overview of the example applications of CI methods in bioinformatics.

4.1. Gene Expression Analysis

Gene expression is the process of using coded information in genes to synthesize proteins or functional RNAs (e.g., ribosomal RNA and transfer RNA) in a cell. The mere evidence of a gene being turned on or activated gives rise to an organism’s phenotype. The DNA microarray technology is widely used to measure expression levels of tens of thousands of genes simultaneously (Quackenbush, 2001). Gene expression values from microarray experiments are often represented as heat maps for visualization (Figure 2). It is crucial to partition the gene expression dataset into groups in order to understand the functional relationships between groups of genes; for example, to discover patterns in gene expression data for tumor and normal colon tissues (Alon, et al., 1999).

There has been a number of CI algorithms applied to cluster gene expression data such as hierarchical clustering (Eisen, et al., 1998; Wen, et al., 1998), principal component analysis (PCA) (Raychaudhuri, et al., 2000; Yeung and Ruzzo, 2001), GA (Li, et al., 2001), and ANN (Herrero, et al., 2001; Tamayo, et al., 1999; Toronen, et al., 1999). Herrero et al (2001) applied the Self-Organizing Tree Algorithm (SOTA) using an unsupervised neural network to analyze gene expression data. The SOTA algorithm (Dopazo and Carazo, 1997) is a neural network with a binary tree topology where each terminal node represents a cluster. It combines the advantages of both hierarchical clustering and Self-Organizing Map (SOM). Futschik and Kasabov (2002) used Fuzzy C-Means (FCM) clustering to achieve a robust analysis of gene expression time-series and addressed issues of parameter selection and cluster validity.

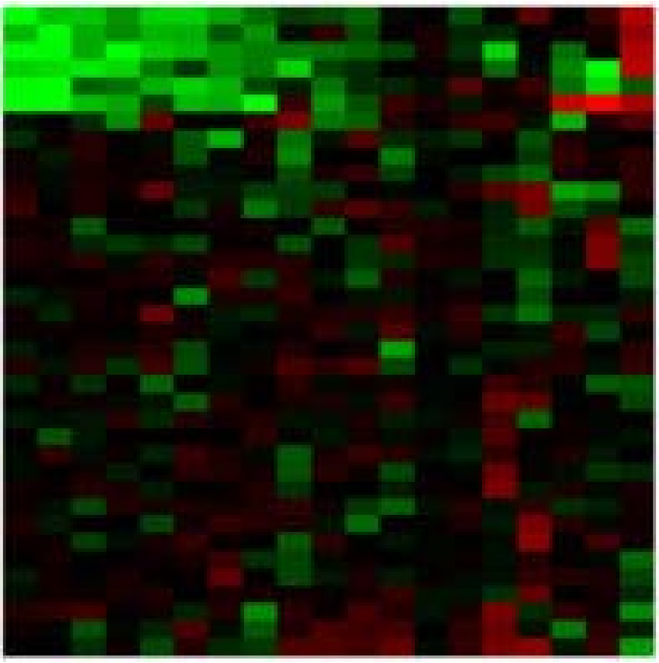

Figure 2. Example heat map for visualizing gene expression values from microarray experiments.

Novel CI algorithms have also been developed for clustering gene expression data. Yuhui et al (2002) proposed the Associative Clustering Neural Network (ACNN) approach to identify inherent clusters by evaluating association between any two gene samples through the interactions of all gene samples. The ACNN method was shown to yield more robust performance than the methods using direct distances for similarities. Xiao et al (2003) introduced a new hybrid clustering approach by combining the PSO and SOM. In their approach, PSO was used to evolve the weights for a SOM and the SOM with an added conscience factor (i.e., assigning each output neuron a bias) was used to cluster the dataset. Okada et al (2005) proposed a novel algorithm to determine biologically interpretable cluster boundaries by referring to functional annotations stored in genome databases. The proposed algorithm can generate set of clusters that are independent of each with respect to their gene function distributions.

4.2. Multiple Sequence Alignment

Multiple sequence alignment (MSA) refers to the process of aligning three or more primary biological sequences such as protein, DNA, and RNA to identify sequence conservations that may be a consequence of functional, structural, or evolutionary relationships between the sequences. For instance, given a family of N sequences $S = \left(S_1,\ S_2,\ \ldots,\ S_N\right)$, it is necessary to perform MSA to find common patterns of the family that may reveal shared evolutionary origins.

Bibliography

Alon, U., et al. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc Natl Acad Sci* U S A, **96**, 6745-6750. [This is a report on the application of a two-way clustering method for analyzing gene expression patterns of tumor and normal colon tissues]

Aloy, P. and Russell, R.B. (2003) InterPreTS: protein interaction prediction through tertiary structure, *Bioinformatics*, Oxford, England, **19**, 161-162. [The paper describes a web interface of a 3D structure based protein-protein interaction prediction method called InterPreTS (Interaction Prediction through Tertiary Structure)]

Apic, G., Gough, J. and Teichmann, S.A. (2001) An insight into domain combinations, *Bioinformatics*, Oxford, England, **17** Suppl 1, S83-89. [A survey of the set of protein domain family combinations present in the archaeal, bacterial, and eukaryote genomes]

Aytuna, A.S., Gursoy, A. and Keskin, O. (2005) Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces, *Bioinformatics*, Oxford, England, **21**, 2850-2855. [The paper presents a novel algorithm for protein-protein interaction prediction by employing a bottom-up approach combining structure and sequence conservation in protein interfaces]

Bader, J.S., et al. (2004) Gaining confidence in high-throughput protein interaction networks, *Nature biotechnology*, **22**, 78-85. [Presents a logistic regression approach using statistical and topological descriptors to predict biological relevance of protein-protein interactions obtained from high-throughput screens of yeast]

Ben-Hur, A. and Noble, W.S. (2005) Kernel methods for predicting protein-protein interactions, *Bioinformatics*, Oxford, England, **21** Suppl 1, i38-46. [Presents a kernel method for predicting protein-protein interactions using a combination of data sources such as protein sequences, Gene Ontology annotations, local network properties, and homologous interactions in other species]

Beni, G. and Wang, U. (1989) Swarm intelligence in cellular robotic systems. NATO Advanced Workshop on Robots and Biological Systems. Tuscany, Italy. [Presents the concept of Swarm Intelligence in relation to cellular robotic systems]

Bernstein, F.C., et al. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures, *J Mol Biol*, **112**, 535-542. [Details on the database of macromolecular structures called Protein Data Bank]

Bezdek, J.C. and Castelaz, P.F. (1977) Prototype Classification and Feature Selection with Fuzzy Sets, *IEEE Transactions on Systems Man and Cybernetics*, 7, 87-92. [Presents a fuzzy ISODATA algorithm to address feature selection and fuzzy classification problems]

Bishop, C.M. (1995) *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford. [This book provides a comprehensive description of the feed-forward neural networks from the perspective of statistical pattern recognition]

Bock, J.R. and Gough, D.A. (2001) Predicting protein--protein interactions from primary structure, *Bioinformatics*, Oxford, England, 17, 455-460. [This paper presents a Support Vector Machine method for predicting protein-protein interactions using protein sequence and associated physiochemical properties]

Bonabeau, E., Dorigo, M. and Theraulaz, G. (2001) Swarm intelligence: From natural to artificial systems, *Journal of Artificial Societies and Social Simulation*, **4**. [Introduces the concepts of ant optimization and simulation of social insects and outlines future research directions of biologically inspired algorithms]

Bremmerman, H.J. (1962) *Optimization through evolution and recombination. Self-Organizing Systems*. Spartan Press, Washington DC. [This book chapter discusses the optimization problem through evolutionary process]

Chen, S.-M., Lin, C.-H. and Chen, S.-J. (2005) Multiple DNA sequence alignment based on genetic algorithms and divide-and-conquer techniques, *International Journal of Applied Science and*

*Engineering***,** **3**, 89-100. [This paper presents a new method for multiple DNA sequence alignment using genetic simulated annealing techniques]

Chen, X.W. and Liu, M. (2005) Prediction of protein-protein interactions using random decision forest framework, *Bioinformatics***,** **21**, 4394-4400. [This paper introduces a protein domain-based random forest of decision trees framework to predict protein-protein interactions]

Chen, X.W. and Liu, M. (2006) Domain-based predictive models for protein-protein interaction prediction, Eurasip *Journal on Applied Signal Processing*. [This paper presents a protein domain-based neural network approach for protein-protein interaction prediction]

Chen, Y., et al. (2006) Partitioned optimization algorithms for multiple sequence alignment. Proc. 20th Intl. Conf. on Advanced Information Networking and Applications. pp. 618-622. [This study proposes a partitioning approach for multiple sequence alignment by utilizing the locality structure]

Chothia, C., et al. (2003) Evolution of the protein repertoire, *Science* New York, N.Y, 300, 1701-1703. [Analyzes the origins of the protein formation process including gene duplication, recombination and divergence]

Clare, A., et al. (2006) Functional bioinformatics for Arabidopsis thaliana, *Bioinformatics*, 22, 1130-1136. [This paper attempts to predict functions of *Arabidopsis* genes using sequence, predicted secondary structure, predicted structural domain, InterPro patterns, sequence similarity profile and expression data]

Cox, E. (1994) The Fuzzy Systems Handbook: A Practitioner's Guide to Building, *Using, Maintaining Fuzzy Systems. AP Professional, San Diego,* CA. [A comprehensive introduction to fuzzy logic]

Cuff, J.A. and Barton, G.J. (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction, *Proteins*, 40, 502-511. [Presents a neural network method for protein secondary structure prediction using multiple sequence alignment profiles]

Deng, M., et al. (2002) Inferring domain-domain interactions from protein-protein interactions, *Genome research*, 12, 1540-1548. [Introduces a maximum-likelihood estimation (MLE) approach to infer protein domain interactions from the protein interaction network]

Dickerson, R.E., Timkovich, R. and Almassy, R.J. (1976) The cytochrome fold and the evolution of bacterial energy metabolism, *J Mol Biol*, 100, 473-491. [Detailed outline for the evolution of photosynthesis and respiration in bacteria]

Dong, X., et al. (2006) Bioinformatics and fuzzy logic. IEEE International Conference on Fuzzy Systems. Vancouver, Canada, pp. 817-824. [Presents two applications of the fuzzy set theory in bioinformatics. One is fuzzy measurement of ontological similarity. The other is the application of fuzzy k-nearest neighbor algorithm in protein secondary structure prediction]

Dopazo, J. and Carazo, J.M. (1997) Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree, *J Mol Evol*, 44, 226-233. [Proposes a new unsupervised, growing, self-organizing neural network for phylogenetic analysis of a large number of sequences]

Eisen, M.B., et al. (1998) Cluster analysis and display of genome-wide expression patterns, *Proc Natl Acad Sci* U S A, 95, 14863-14868. [Presents a genome-wide expression data clustering technique that uses standard statistical algorithms to arrange genes according to similarity in gene expression patterns]

Engelbrecht, A.P. (2005) *Fundamentals of Computational Swarm Intelligence*. Wiley. [A comprehensive introduction to the computational paradigm of Swarm Intelligence]

Espadaler, J., et al. (2005) Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships, *Bioinformatics,* , 21, 3360-3368. [Describes a protein-protein interaction prediction technique using both structural similarities among domains of known interacting proteins and conservation of pairs of sequence patches involved in the interfaces]

Eyrich, V.A. and Rost, B. (2003) META-PP: single interface to crucial prediction servers, *Nucleic Acids Res*, 31, 3308-3310. [A server called META-PP provides access to a selected set of high-quality servers in the areas of comparative modeling, threading/fold recognition, secondary structure prediction and more specialized fields like contact and function prediction]

Feng, D.F. and Doolittle, R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees, *J Mol Evol*, 25, 351-360. [Describes a progressive alignment method that utilizes the Needleman and Wunsch pairwise alignment algorithm iteratively to achieve alignment of multiple sequences and to construct an evolutionary tree]

Fogel, G.B. (2008) Computational intelligence approaches for pattern discovery in biological systems, *Brief Bioinform*, 9, 307-316. [A review that provides introduction to computational intelligence methods and their application to biological problems]

Fogel, G.B., GCorne, D.W. and Pan, Y. (2008) *Computational Intelligence in Bioinformatics*. Wiley-IEEE Press, Piscataway, NJ. [This book covers the most relevant and popular computational intelligence methods applied in bioinformatics]

Fogel, L.J., Owens, A. and Walsh, M.J. (1966) *Artificial Intelligence Through Simulated Evolution*. Wiley, New York, NY. [This book covers the current research and developments of the use of evolutionary programming to generate artificial intelligence]

Friedman, N., et al. (2000) Using Bayesian networks to analyze expression data, *J Comput Biol*, 7, 601-620. [Describes the use of Bayesian networks for the analysis of microarray expression data to recover gene interactions]

Futschik, M.E. and Kasabov, N.K. (2002) Fuzzy clustering of gene expression data. Proc. 2002 IEEE Intl. Conf. on Fuzzy Systems. pp. 414-419. [This presents an application of fuzzy c-means clustering method to achieve a robust analysis of gene expression time-series]

Gardy, J.L., et al. (2003) PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria, *Nucleic Acids Res*, 31, 3613-3617. [Present an updated version of the PSORT tool for bacterial protein analysis]

Ginalski, K., et al. (2003) 3D-Jury: a simple approach to improve protein structure predictions, *Bioinformatics*, 19, 1015-1018. [3D-Jury is a system to generate meta-predictions using variable sets of models obtained from diverse sources]

Goldberg, D.E. (1975) *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Reading, MA. [This book presents computer techniques, mathematical tools and research results for the application of genetic algorithms]

Gomez, S.M., Lo, S.H. and Rzhetsky, A. (2001) Probabilistic prediction of unknown metabolic and signal-transduction networks, *Genetics*, 159, 1291-1298. [Describes a statistical model to predict unknown molecular interactions within regulatory networks by representing proteins as collections of domains or motifs]

Gomez, S.M. and Rzhetsky, A. (2002) Towards the prediction of complete protein--protein interaction networks, Pacific Symposium on Biocomputing, 413-424. [Presents a statistical method for the prediction of protein-protein interactions within an organism by treating proteins as collections of conserved domains]

Gondro, C. and Kinghorn, B.P. (2007) A simple genetic algorithm for multiple sequence alignment, *Genet Mol Res*, 6, 964-982. [Application of genetic algorithm for multiple sequence alignment]

Graves, A., et al. (2009) A Novel Connectionist System for Unconstrained Handwriting Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31, 855-868. [This paper introduces an approach based on recurrent neural network to sequence labeling tasks]

Greenbaum, D., et al. (2003) Comparing protein abundance and mRNA expression levels on a genomic scale, Genome biology, 4, 117. [Compared attempts in correlating protein abundance with mRNA expression levels]

Grigoriev, A. (2003) On the number of protein-protein interactions in the yeast proteome, *Nucleic Acids Res*, 31, 4157-4161. [This paper estimates the average number of interacting partners per protein in the yeast proteome]

Guermeur, Y., et al. (1999) Improved performance in protein secondary structure prediction by inhomogeneous score combination, *Bioinformatics*, 15, 413-421. [Proposes an ensemble method for protein secondary structure prediction]

Guimaraes, K.S., et al. (2006) Predicting domain-domain interactions using a parsimony approach, *Genome biology*, 7, R104. [Describes an approach to predict protein domain-domain interactions from a protein interaction network by applying a parsimony driven explanation of the network]

Guimaraes, K.S. and Przytycka, T.M. (2008) Interrogating domain-domain interactions with parsimony based approaches, *BMC bioinformatics*, 9, 171. [Introduces a Generalized Parsimonious Explanation method for the prediction of protein domain-domain interactions]

Herrero, J., Valencia, A. and Dopazo, J. (2001) A hierarchical unsupervised growing neural network for clustering gene expression patterns, *Bioinformatics*, 17, 126-136. [Presents the analysis of gene expression data using unsupervised neural network]

Holland, J.H. (1975) *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI. [This book covers the role of genetic algorithms in studies of complex adaptive systems]

Huang, C., et al. (2007) Predicting protein-protein interactions from protein domains using a set cover approach, *IEEE/ACM Trans Comput Biol Bioinform*, 4, 78-87. [Introduces Maximum Specificity Set cover method for the prediction of protein-protein interactions]

Hue, M., et al. (2010) Large-scale prediction of protein-protein interactions from structures, *BMC Bioinformatics*, 11, 144. [Presents a 3D structure based method for protein-protein interactions using support vector machine]

Huttenhower, C., et al. (2006) A scalable method for integration and functional analysis of multiple microarray datasets, *Bioinformatics*, Oxford, England, 22, 2890-2897. [Presents a scalable Bayesian framework to predict functional relationships from integrated microarray datasets]

Jaimovich, A., et al. (2006) Towards an integrated protein-protein interaction network: a relational Markov network approach, *J Comput Biol*, 13, 145-164. [This study presents a Markov network approach to integrate various protein attributes]

Jansen, R., et al. (2002) Integration of genomic datasets to predict protein complexes in yeast, *J Struct Funct Genomics*, 2, 71-81. [Explores the integration of genomic datasets to predict membership of individual genes in protein complexes]

Jansen, R., et al. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data, *Science* (New York, N.Y, 302, 449-453. [Describes a Bayesian network approach to predict protein interactions]

Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices, *J Mol Biol*, 292, 195-202. [Presents a two-stage neural network model to predict protein secondary structure based on the position specific scoring matrices generated by PSI-BLAST]

Jones, S. and Thornton, J.M. (1997a) Analysis of protein-protein interaction sites using surface patches, *J Mol Biol*, 272, 121-132. [Presents analysis of protein interactions sites using surface patches]

Jones, S. and Thornton, J.M. (1997b) Prediction of protein-protein interaction sites using patch analysis, *J Mol Biol*, 272, 133-143. [Describes a method to analyze a series of residue patches on the surface of protein structures and predicts the location of protein interaction sites using the information]

Kelemen, A., Abraham, A. and Chen, Y. (2008) *Computational Intelligence in Bioinformatics. Studies in Computational Intelligence*. Springer, Heidelberg. [Covers the application of computational intelligence in bioinformatics]

Keller, J.M. and Tahani, H. (1992) Implementation of conjunctive and disjunctive fuzzy logic rules with neural networks, *International Journal of Approximate Reasoning*, 6, 221-240. [Describes the use of fuzzy logic to model and manage uncertainty in neural networks]

Kelley, L.A., MacCallum, R.M. and Sternberg, M.J. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM, *J Mol Biol*, 299, 499-520. [Described a method to recognize remote protein sequence homologues by combining the multiple sequence profiles with the knowledge of protein structure]

Kendrew, J.C., et al. (1960) Structure of myoglobin: A three-dimensional Fourier synthesis at 2 A. resolution, *Nature*, 185, 422-427. [Article describing the structure of myoglobin]

Kennedy, J. (1999) Small worlds and mega-minds: Effects of neighborhood topology on particle swarm performance. 1999 Congress of Evolutionary Computation. pp. 1931-1938. [Describes the effects of neighborhood topologies on particle swarm optimization of four test functions]

Kim, W.K., Park, J. and Suh, J.K. (2002) Large scale statistical prediction of protein-protein interaction by potentially interacting domain (PID) pair, *Genome Inform*, 13, 42-50. [Introduces a statistical method for protein interaction prediction based on protein domains]

King, R.D., et al. (2000) Is it better to combine predictions?, *Protein Eng*, 13, 15-19. [Analysis of whether integrative approach is better by comparing the accuracy of individual protein secondary structure prediction methods against the accuracy obtained by combing predictions of the methods]

Kini, R.M. and Evans, H.J. (1996) Prediction of potential protein-protein interaction sites from amino acid sequence. Identification of a fibrin polymerization site, *FEBS letters*, 385, 81-86. [Presents a predictive method to identify protein interaction sites based on observations derived from the sequences]

Koonin, E.V., Wolf, Y.I. and Karev, G.P. (2002) The structure of the protein universe and genome evolution, *Nature*, 420, 218-223. [A comprehensive review on the structure of protein universe and genome evolution]

Kuang, R., Leslie, C.S. and Yang, A.S. (2004) Protein backbone angle prediction with machine learning approaches, *Bioinformatics*, 20, 1612-1621. [Describes support vector machine and neural network models to predict protein backbone conformational state]

Lee, I., et al. (2004) A probabilistic functional network of yeast genes, *Science* (New York, N.Y, 306, 1555-1558. [Presents a conceptual framework for integrating various functional genomic data]

Lee, S., et al. (2004) Exploring protein fold space by secondary structure prediction using data distribution method on Grid platform, *Bioinformatics*, 20, 3500-3507. [Presents a method based on a Grid platform to predict protein secondary structure]

Li, L., et al. (2001) Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method, *Bioinformatics*, 17, 1131-1142. [Describes an approach combining the genetic algorithm and the K-Nearest Neighbor method to identify genes that can jointly discriminate between normal vs tumor samples]

Lin, X., Liu, M. and Chen, X.W. (2009) Assessing reliability of protein-protein interactions by integrative analysis of data in model organisms, *BMC Bioinformatics*, 10 Suppl 4, S5. [This paper presents a Bayesian network-based integrative framework to assess the reliability of protein interactions using cross-species data]

Liu, J. and Rost, B. (2001) Comparing function and structure between entire proteomes, *Protein Sci*, 10, 1970-1979. [Application of a variety of simple bioinformatics tools to analyze 29 proteomes from eukaryotes, prokaryotes, and archaebacteria]

Liu, Y., Liu, N. and Zhao, H. (2005) Inferring protein-protein interactions through high-throughput interaction data from diverse organisms, *Bioinformatics*, Oxford, England, 21, 3279-3285. [Presents a likelihood approach to estimate protein domain-domain interaction probabilities by integrating large-scale protein interaction data from yeast, worm, and fruit fly organisms]

Lu, L., et al. (2003) Multimeric threading-based prediction of protein-protein interactions on a genomic scale: application to the Saccharomyces cerevisiae proteome, *Genome research*, 13, 1146-1154. [This describes a multimeric threading algorithm for the prediction of protein-protein interactions]

Lu, L., Lu, H. and Skolnick, J. (2002) MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading, *Proteins*, 49, 350-364. [This describes a multimeric threading algorithm for the prediction of protein-protein interactions]

Martin, S., Roe, D. and Faulon, J.L. (2005) Predicting protein-protein interactions using signature products, *Bioinformatics*, Oxford, England, 21, 218-226. [This article presents a support vector machine method that uses signature descriptor of interacting proteins for the prediction of protein-protein interactions]

Mewes, H.W., et al. (2006) MIPS: analysis and annotation of proteins from whole genomes in 2005, *Nucleic Acids Res*, 34, D169-172. [This article describes manually curated databases of genome information for several reference organisms]

MIchalewicz, Z. (1996) *Genetic Algorithms + Data Structures = Evolution Programs*. 3rd edn. Springer, Berlin, Germany. [This book presents the introduction of genetic algorithms and application of genetic algorithms in numerical optimization.]

Mitra, S. and Hayashi, Y. (2006) Bioinformatics with soft computing, *IEEE Transactions on Systems Man and Cybernetics* Part C-Applications and Reviews, 36, 616-635. [Surveys the role of different soft computing paradigms, like fuzzy sets, artificial neural network, evolutionary computation, and etc. in bioinformatics]

Montgomerie, S., et al. (2006) Improving the accuracy of protein secondary structure prediction using structural alignment, *BMC Bioinformatics*, 7, 301. [Proposes a method that performs structure-based sequence alignments as part of the protein secondary structure prediction process]

Mordeson, J.N., Malik, D.S. and Cheng, S.-C. (2000) *Fuzzy Mathematics in Medicine*, Physica. [This book presents a variety of types of fuzzy mathematics used in medical research]

Myers, C.L. and Troyanskaya, O.G. (2007) Context-sensitive data integration and prediction of biological networks, *Bioinformatics*, Oxford, England, 23, 2322-2330. [Proposes a Bayesian approach for context-sensitive integration and query-based recovery of biological process specific networks]

Nasser, S., et al. (2007) Multiple Sequence Alignment using Fuzzy Logic. *Proc. IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology*. pp. 304-311. [This article presents application of fuzzy logic for approximate matching of subsequences]

Notredame, C. and Higgins, D.G. (1996) SAGA: Sequence alignment by genetic algorithm, *Nucleic Acids Research*, **24,** 1515-1524. [Describes a software package SAGA for multiple sequence alignment using automatic schedule scheme to control the usage of 22 different operators by a genetic algorithm]

Nye, T.M., et al. (2005) Statistical analysis of domains in interacting protein pairs, *Bioinformatics*, Oxford, England, 21, 993-1001. [Presents a statistical approach to assign p-values to pairs of domain superfamilies to predict which domains come into contact in an interacting protein pair]

Okada, Y., et al. (2005) Knowledge-assisted recognition of cluster boundaries in gene expression data, *Artif Intell Med*, 35, 171-183. [This article proposes a clustering algorithm in which cluster boundaries are determined by referring to functional annotations stored in genome databases]

Ortiz, A.R., Strauss, C.E. and Olmea, O. (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison, *Protein Sci*, 11, 2606-2621. [Presents a sequence-independent structural alignment method that allows comparison of an experimental protein structure with an arbitrary low-resolution protein tertiary model]

Oyama, T., et al. (2002) Extraction of knowledge on protein-protein interaction by association rule discovery, *Bioinformatics*, Oxford, England, 18, 705-714. [Proposes to discover association rules related to protein-protein interactions]

Perutz, M.F., et al. (1960) Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis, *Nature*, 185, 416-422. [Details the 3D structure of hemoglobin using X-ray analysis]

Qi, Y., Klein-Seetharaman, J. and Bar-Joseph, Z. (2005) Random forest similarity for protein-protein interaction prediction from multiple sources, Pacific Symposium on Biocomputing, 531-542. [Presents a method to compute similarities for classifying pairs of proteins as interacting or not]

Quackenbush, J. (2001) Computational analysis of microarray data, *Nat Rev Genet*, 2, 418-427. [Provides a basic understanding of available computational tools for microarray data analysis]

Rasmussen, T.K. and Krink, T. (2003) Improved Hidden Markov Model training for multiple sequence alignment by a particle swarm optimization - evolutionary algorithm hybrid, *Biosystems*, 72, 5-17. [Reports the use of a hybrid algorithm combining particle swarm optimization with evolutionary algorithms to train Hidden Markov Models for multiple sequence alignment]

Raychaudhuri, S., Stuart, J.M. and Altman, R.B. (2000) Principal components analysis to summarize microarray experiments: Application to sporulation rime series. Proc. Pacific Symposium on Biocomputing. pp. 452-463. [Shows the application of principal component analysis to expression data allows summarization of the ways in which gene responses vary under different conditions]

Rechenberg, I. (1973) *Evolutionsstrategie: Optimerung technischer Systeme nach Prinzipien der biologischen Evolution.* Fromman-Holzboog, Stuttgart, Germany. [Describes evolutionary strategies for optimization]

Reichhardt, T. (1999) It's sink or swim as a tidal wave of data approaches, *Nature*, 399, 517-520. [Perspectives on data approaches for genomics]

Rhodes, D.R., et al. (2005) Probabilistic model of the human protein-protein interaction network, *Nature biotechnology*, 23, 951-959. [Presents a probabilistic analysis for integrating model organism protein interaction data, protein domain data, genome-wide gene expression data and functional annotation to predict protein interactions in humans]

Riley, M.L., et al. (2007) PEDANT genome database: 10 years online, *Nucleic Acids Res*, 35, D354-357. [Presents the PEDANT genome database that provides exhaustive annotation of 468 genomes by broad set of bioinformatics algorithms]

Riley, R., et al. (2005) Inferring protein domain interactions from databases of interacting proteins, *Genome biology*, 6, R89. [Describes a method called domain pair exclusion analysis (DPEA) for inferring protein domain interactions from databases of interacting proteins]

Rost, B. (1996) PHD: predicting one-dimensional protein structure by profile-based neural networks, *Methods Enzymol*, 266, 525-539. [Presents a method for predicting one-dimensional protein structure using profile-based neural networks]

Rost, B. (2001) Review: protein secondary structure prediction continues to rise, *J Struct Biol*, 134, 204-218. [This is a comprehensive review on methods for predicting protein secondary structure]

Rost, B. and Sander, C. (1993) Prediction of protein secondary structure at better than 70% accuracy, *J Mol Biol*, 232, 584-599. [Describes a two-layered feed-forward neural network model to predict the secondary structure of water-soluble proteins]

Rost, B. and Sander, C. (2000) *Third generation prediction of secondary structure. Protein Structure Prediction: Methods and Protocols*. Humana Press, Clifton, NJ. [This chapter provides a review of protein secondary-structure prediction methods]

Rost, B., Sander, C. and Schneider, R. (1994) PHD--an automatic mail server for protein secondary structure prediction, *Comput Appl Biosci*, 10, 53-60. [Presents an automatic mail server for protein secondary structure prediction]

Rost, B., Yachdav, G. and Liu, J. (2004) The PredictProtein server, *Nucleic Acids Res*, 32, W321-326. [Presents an internet service for sequence analysis and prediction of protein structure and function]

Seiffert, U., Jain, L.C. and Schweizer, P. (2005) *Bioinformatics Using Computational Intelligence Paradigms*. Springer Verlag, Heidelberg, Germany. [This book presents the latest results of bioinformatics and computational intelligence]

Selbig, J., Mevissen, T. and Lengauer, T. (1999) Decision tree-based formation of consensus protein secondary structure prediction, *Bioinformatics*, 15, 1039-1046. [Presents an approach to reveal subtle but systematic differences in the output of different secondary structure prediction methods]

Singh, R., Xu, J. and Berger, B. (2006) Struct2net: integrating structure into protein-protein interaction prediction, Pacific Symposium on Biocomputing, 403-414. [Presents a framework for predicting protein-protein interactions by integrating structure-based information with other functional annotations such as gene ontology, co-expression and co-location information]

Sprinzak, E. and Margalit, H. (2001) Correlated sequence-signatures as markers of protein-protein interaction, *J Mol Biol*, 311, 681-692. [Presents a protein interaction prediction method using the characteristic pairs of sequence-signatures learned from a database of experimentally determined interacting proteins]

Stothard, P., et al. (2005) BacMap: an interactive picture atlas of annotated bacterial genomes, *Nucleic Acids Res*, 33, D317-320. [Describes the BacMap interactive visual database containing fully labeled, zoomable and searchable chromosome maps from more than 170 bacterial species]

Szafron, D., et al. (2004) Proteome Analyst: custom predictions with explanations in a web-based tool for high-throughput proteome annotations, *Nucleic Acids Res*, 32, W365-371. [Presents a publicly available, high-throughput, web-based system for predicting various properties of each protein in an entire proteome]

Szczepanniak, P.S., Lisoba, P.J.G. and Kacprzyk, J. (2000) Fuzzy systems in Medicine, *Physica*. [Provides an introduction to the fundamental concepts of fuzziness with the recent advances in its application to medicine]

Szent-Gyorgyi, A.G. and Cohen, C. (1957) Role of proline in polypeptide chain configuration of proteins, *Science*, 126, 697-698. [Outlines the role of proline in polypeptide chain configuration of proteins]

Tamayo, P., et al. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation, *Proc Natl Acad Sci* U S A, 96, 2907-2912. [Describes the application of self-organizing maps for gene expression data analysis]

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res*, 22, 4673-4680. [Describes the CLUSTAL W method for multiple sequence alignment]

Tong, A.H., et al. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules, *Science* (New York, N.Y, 295, 321-324. [Presents a strategy to define protein interaction networks of peptide recognition modules by combining computational interaction predictions with experimental results]

Toronen, P., et al. (1999) Analysis of gene expression data using self-organizing maps, *FEBS Lett*, 451, 142-146. [Presents the application of self-organizing maps for gene expression data analysis]

Torres, A. and Nieto, J.J. (2006) Fuzzy logic in medicine and bioinformatics, *J Biomed Biotechnol*, 2006, 91908. [Provides a general view of the current applications of fuzzy logic in medicine and bioinformatics]

Van Domselaar, G.H., et al. (2005) BASys: a web server for automated bacterial genome annotation, *Nucleic Acids Res, 33*, W455-459. [Presents the BASys web server for automated annotation of bacterial genomic sequences]

Vogel, C., et al. (2004) Structure, function and evolution of multidomain proteins, *Curr Opin Struct Biol*, 14, 208-216. [Provides a comprehensive description of the structure, function, and evolution of proteins with multiple domains]

von Mering, C., et al. (2002) Comparative assessment of large-scale data sets of protein-protein interactions, *Nature,* 417, 399-403. [This is a comparison of different protein interaction prediction methods]

Wang, H., et al. (2007) InSite: a computational method for identifying protein-protein interaction binding sites on a proteome-wide scale, *Genome biology*, 8, R192. [Proposes a computational method that integrates high-throughput protein and sequence data to infer the specific binding regions of interacting protein pairs]

Wen, X., et al. (1998) Large-scale temporal gene expression mapping of central nervous system development, *Proc Natl Acad Sci* U S A, 95, 334-339. [Presents a systematic approach to measure multiple gene expression time series to produce a temporal map of developmental gene expression]

Wong, S.L., et al. (2004) Combining biological networks to predict genetic interactions, *Proc Natl Acad Sci* U S A, 101, 15682-15687. [Presents a method to predict synthetic sick or lethal (SSL) genetic interactions by combining biological networks]

Xiao, X., et al. (2003) Gene clustering using self-organizing maps and particle swarm optimization. Proc. 17th Intl. Symposium on Parallel and Distributed Processing. [Describes a hybrid clustering approach based on self-organizing maps and particle swarm optimization to cluster genes]

Yamanishi, Y., Vert, J.P. and Kanehisa, M. (2004) Protein network inference from multiple genomic data: a supervised approach, *Bioinformatics*, Oxford, England, 20 Suppl 1, i363-370. [Presents a supervised learning algorithm for protein network inference using multiple types of genomic data]

Ye, Y. and Godzik, A. (2004) Comparative analysis of protein domain organization, *Genome research*, 14, 343-353. [Presents a set of graph theory-based tools to compare protein domain organizations of different organisms]

Yeung, K.Y. and Ruzzo, W.L. (2001) Principal component analysis for clustering gene expression data, *Bioinformatics*, 17, 763-774. [This is a study of the effectiveness of principal components in capturing cluster structure from gene expression data]

Yuhui, Y., et al. (2002) Clustering gene data via associative clustering neural network. Proc. 9th Intl. Conf. on Information Processing. pp. 2228-2232. [Describes an approach to analyze gene expression data using Associative clustering Neural Network]

Zadeh, L.A. (1965) Fuzzy Sets, *Information and Control*, 8, 338-353. [Presents the fundamental concepts of fuzzy sets]

Zhang, G.-Z. and Huang, D.-S. (2004) Aligning multiple protein sequence by an improved genetic algorithm. Proc. IEEE Intl. Joint Conf. on Neural Networks. pp. 1179-1183. [Presents an improved genetic algorithm method for multiple sequence alignment]

Zhang, L.V., et al. (2004) Predicting co-complexed protein pairs using genomic and proteomic data integration, *BMC bioinformatics*, 5, 38. [Presents a probabilistic decision tree approach to predict co-complexed protein pairs by integrating high-throughput protein interaction datasets and other gene- and protein-pair characteristics]

Zhang, Q., Yoon, S. and Welsh, W.J. (2005) Improved method for predicting beta-turn using support vector machine, *Bioinformatics*, 21, 2370-2374. [Introduces a support vector machine algorithm to predict β-turns in proteins]

Zhong, W. and Sternberg, P.W. (2006) *Genome-wide prediction of C. elegans genetic interactions*, Science (New York, N.Y, 311, 1481-1484. [Presents a computational method for integrating interactome data, gene expression data, phenotype data, and functional annotation from yeast, worm, and fruit fly organisms to predict genome-wide genetic interactions in worm]

Zimmermann, H. (2001) *Fuzzy Set Theory and its Applications*. Kluwer Academic, Hingham, MA. [Provides an introduction to fuzzy set theory and its applications]

Zimmermann, O. and Hansmann, U.H. (2006) Support vector machines for prediction of dihedral angle regions, *Bioinformatics*, 22, 3009-3015. [Presents a multi-step support vector machine procedure to predict the dihedral angle state of residues from sequence]

Zvelebil, M.J., et al. (1987) Prediction of protein secondary structure and active sites using the alignment of homologous sequences, *J Mol Biol*, 195, 957-961. [Describes an approach to predict protein secondary structure and active sites using the alignment of homologous sequences

Biographical Sketches

M. Liu received her PhD degree in computer science from the University of Kansas in July, 2009 with a research focus in bioinformatics. Upon completing her doctoral study, she was awarded the National Library of Medicine postdoctoral training fellowship and completed the training in the Department of Biomedical Informatics at Vanderbilt University in June, 2012. She is currently an Assistant Professor in the Department of Computer Science at New Jersey Institute of Technology. Her research interest includes bioinformatics, biomedical informatics, data mining, and machine learning. She has published numerous papers in top rated journals such as PLoS ONE, Bioinformatics, and Journal of American Medical Informatics Association.

X. W. Chen is currently a Full Professor and Chair in the Department of Computer Science at Wayne State University. He is a senior IEEE member, the chair of IEEE Technical Committee on Bioinformatics. He is the editor-in-Chief of the *International Journal of Computational Intelligence in Bioinformatics and Systems Biology* and also serves in the editorial board in several other international journals. Dr. Chen received his PhD degree from Carnegie Mellon University, Pittsburgh, USA in 2001. He was a recipient of the NSF CAREER Award in 2007. He served as conference chair (and co-chairs) for several international conferences including ACM Conference on Information and Knowledge Management (CIKM), 2012; the IEEE International Conference on Bioinformatics and Biomedicine (BIBM) in 2009 and the International Conference in Machine Learning and Applications in 2008. He also served as a program committee member in numerous conferences such as KDD, CIKM, ICDM, and CEC. His research interest includes machine learning, data mining, bioinformatics, systems biology, and statistical modeling.