Natural Computing Methods in Bioinformatics: A Survey

Francesco Masulli [a,b,*] and Sushmita Mitra [c]

[a] *Department of Computer and Information Sciences, University of Genova, Via Dodecaneso 35, 16146 Genoa, ITALY.*

[b] *Center for Biotechnology, Temple University, 1900 N 12th Street Philadelphia PA 19122, USA.*

[c] *Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700 108, INDIA.*

Abstract

Often data analysis problems in Bioinformatics concern the fusion of multisensor outputs or the fusion of multi-source information, where one must integrate different kinds of biological data. Natural computing provides several possibilities in Bioinformatics, especially by presenting interesting nature-inspired methodologies for handling such complex problems. In this article we survey the role of natural computing in the domains of protein structure prediction, microarray data analysis and gene regulatory network generation. We utilize the learning ability of neural networks for adapting, uncertainty handling capacity of fuzzy sets and rough sets for modeling ambiguity, and the search potential of genetic algorithms for efficiently traversing large search spaces.

Key words: Natural Computing, Bioinformatics, protein structure prediction, microarray data analysis, biological networks modeling.

1 Introduction

The 20^{th} Century is frequently referred as the *Century of Biology*, given the huge developments of this scientific area that concluded that century with the great success of the Human Genome Project [1,2] producing the full human DNA sequencing.

* Corresponding author.
Email addresses: masulli@disi.unige.it (Francesco Masulli), sushmita@isical.ac.in (Sushmita Mitra).

Nowadays, we are at the beginning of the so called *Post-Genomics Era* characterized on one hand by the availability of immense amount of bioinformatics data (often in the public domain) and on the other hand by the need of new and efficient mathematical and algorithmic methods able to extract the information embedded in those data (for an introduction to Bioinformatics see, e.g., [3–5]), and as a matter of fact, the focus of the research in Bioinformatics is shifting from the development of efficient data storage methods to the extraction of useful information from data.

A vital issue concerns, in particular, the need of reliable algorithms capable of fusing the information embedded in heterogeneous biological data. For example, in protein structure prediction one should integrate with the available information on the sequence of residues some other biological information like hydrophobicity, evolutionary information, and solvent accessibility. Moreover, multisensor outputs in microarrays (where thousand of biological experiments are performed in parallel on a single slice of glass) may need to be combines for enhanced performance. This fusion can be at different levels of resolution, depending on the focus of interest of the user.

Natural Computing models, inspired in part by nature and natural systems, are a family of powerful data analysis methods able to transform available heterogeneous data into biological knowledge. They include Neural Networks mimicking the mechanisms of the nervous system [6], Fuzzy Systems based on an extension of traditional logic in order to represent uncertainty and qualitative reasoning [7], Machine Learning approaches [8], and general optimization techniques, such as Evolutionary Computation based on simulation of biological evolution [9,10], Swarm Intelligence based on simulation of social behavior of animals [11], Immunocomputing inspired by the biological immune system [12], and Simulated Annealing derived by Statistical Mechanics [13].

In recent years, many Natural Computing models have been successfully applied to the solution of complex problems related to signal processing, classification, clustering, feature selection, data visualization, data mining, and information fusion [14]. These provide efficient paradigms for fusion of different kinds of information. The aim is to synergistically merge the techniques so that they cooperate with each other in enhancing the overall performance.

The application of Natural Computing techniques encompasses several fields of Bioinformatics. An attempt has been made in this survey to compile some of the existing literature pertaining to the use of Natural Computing in Bioinformatics, concentrating on those fields where most of the research efforts have been concentrated in the last twenty years. Section 2 introduces the methodologies used for the prediction of Protein Structure. Section 3 presents a study of various natural computing tools used for the analysis of microarray data. Section 4 completes the survey with an overview of biological networks and

throws some light into their extraction. Finally, Section 5 concludes the article.

2 Protein Structure Prediction

A protein is a polymer constituted by a chain of amino acids (also called residues) linked by peptide bonds. Protein sequence can be represented by a string of alphabets, each of which belongs to 20 letters representing 20 amino acids. The biological function of a protein depends on its 3-dimensional structure (fold or tertiary structure).

The Protein Data Bank (PDB) [15] [1] contains at present time roughly 40,000 resolved 3D protein structures deposited while more than 2 million non-redundant protein sequences are known. This gap is due to the cost of the process of experimental determination of protein structures, either by X-ray crystallography or by nuclear magnetic resonance methods, that limits the growing of PDB to not more than 5000 new protein structures per year.

The folding of a protein corresponds to the minimization of its free energy and depends on its sequence and on external environment [16]. As the explicit minimization of protein potential functions from first principles is infeasible on today available computers, an extraordinary research effort has been carried out in the last 20 years in order to develop efficient methods able to predict the 3D structure of a protein starting from is amino acid sequence. A valuable international initiative stimulating this effort is the Critical Assessment of Techniques for Protein Structure Prediction CASP [2], a community-wide experiment taking place every two years since 1994, aimed to allow the research groups to assess the quality of their methods [17].

Prediction methods can be classified as [18]: (a) Template-based modeling (TBM), based on finding known structures (templates) related to the sequence to be modeled (target); (b) Free-modeling (FM), or *ab initio*, used when structural analogs do not exist in the PDB library or could not be successfully identified. At present, TBM methods are quite accurate, especially when the match with existing sequences is above 50%, while FM methods are less effective [18].

In 1998, Qian and Sejnowski [20] proposed the first application of neural networks to secondary protein structure prediction, making use of a multilayer perceptron and a binary sequence encoding method.

Subsequent studies proposed some variants of this approach [21–24,26]. For

[1] http://www.wwpdb.org/
[2] http://predictioncenter.org/

example, in [23] the input of the neural network has been augmented with the hydrophobicity of each residue, while [25] studied different encoding schemes and a modular architecture.

In all these works, protein sequences were analyzed using sliding windows of fixed-length segments. The goal of the neural networks was to correctly predict the secondary structure for the middle amino acid of the window, which coded for either a three-state (α helix, β sheet, and random coil) or a two-state (e.g. helix, non-helix). The three-state prediction was encoded using three or two output units of the neural network.

The obtained accuracy with those methods was not better than 65% for three-state prediction, i.e., not much better than a simple Bayesian statistical approach assuming independent probabilities of residues [24].

Some small improvements in accuracy were notified in [23] using protein tertiary structural class, while in [26] the neural network design for secondary structure prediction of globular proteins was extended to the prediction of membrane proteins, obtaining better results than those obtained with statistical methods.

The addition of evolutionary information in the form of multiple alignment profiles proposed in [27], substantially boosted the prediction accuracy, surpassing a 70% level of the average three-state accuracy. The multiple alignment profile contains the frequency of every possible amino acid in each position of a protein, as obtained from the multiple alignment in that position.

Some architectural enhancements where proposed in [28], introducing an adaptive encoding of amino acids, while [29] adopted position specific matrices for incorporating evolutionary information.

A Hidden Markov model method for finding remote homologs of protein sequences was proposed in [30]. The method begins with a single target sequence and iteratively builds a Hidden Markov model from the sequence and homologs found using the Hidden Markov model for database search.

Bi-directional recurrent neural networks, proposed in [31,33,32], can learn to make predictions of protein secondary structure based on variable ranges of dependencies. These architectures extend recurrent neural networks introducing non-causal bidirectional dynamics to capture both upstream and downstream information.

Hidden Markov models and bi-directional recurrent neural networks reached an accuracy between 75 and 79% in the three-state secondary structure prediction. Further improvements were obtained in [34], using a recursive and bi-directional neural network. The network takes as inputs the protein se-

quence, evolutionary information obtained from multiple sequence alignments, and predicted secondary structure and relative solvent accessibility (obtained in its turn using predictor based on an ensemble of three recursive neural networks trained on a non-redundant set of protein contact maps) [3].

Genetic algorithms have been applied to the determination of protein structure from sequence, using a full atom representation in [35]. A free energy function with point charge electrostatics and an area based solvation model is used.

A multi-objective evolutionary algorithm has been proposed in [36], as a search procedure for exploring the conformational space of the protein structure prediction problem for the minimization of two different interaction energies: local (bond atoms) and non-local (non-bond atoms).

A fuzzy sets-based adaptive neighborhood search optimization heuristic for protein structure prediction was proposed in [37]. A fuzzy generalization of contact map has been presented in [38].

A two-stage machine learning, information retrieval, approach to fold recognition has been studied in [39]. A set of similarity measures between query-template protein pairs is computed, including alignment scores, pairwise structural compatibility features, solvent accessibility, contact map and beta- strand pairings of the query protein against the tertiary structure of the template protein. Finally, these features were fed into support vector machines to predict the structural relevance of the query-template pairs.

In [40,41], multilayer perceptrons have been applied to protein tertiary structure prediction. A simple simulated annealing procedure to assemble native-like tertiary structures from fragments of unrelated protein structures with similar local sequences using Bayesian scoring functions is presented in [42].

The strong coupling between secondary and tertiary structure formation in protein folding was exploited to improve protein secondary structure prediction in [43]. The architecture of a neural network for secondary structure prediction that utilizes multiple sequence alignments was extended to accept low-resolution nonlocal tertiary structure information as an additional input.

3 Microarray Data Analysis

For the proper understanding of the function of genes and proteins, protein structure evaluation is essential. Each DNA array contains the measures of

[3] A contact map represents the distance between every two residues of a three-dimensional protein structure.

the level of expression of many genes. Gene expression data, coupled with various analysis methods, serves as an indispensable tool for the analysis of protein functions. Various distances and/or correlations can be computed from pairwise comparison of these patterns. Let $gene_j(e_{j1}, \ldots, e_{jn})$ denote the expression pattern for the jth gene for $i = 1, \ldots, n$ samples. The *Euclidean distance* between the jth and kth genes, computed as

$$d_{j,k} = \sqrt{\sum_i (e_{ji} - e_{ki})^2}, \tag{1}$$

is suitable when the objective is to cluster genes displaying similar levels of expression. Cluster validation can be done using either external and internal criterion analyses [65]. A quantitative data-driven framework has been developed [44] to evaluate different clustering algorithms, without using additional biological knowledge about the gene expression data. The *Pearson correlation coefficient* $-1 \leq r \leq 1$ measures the similarity in trend between two profiles (genes). The distance is given as

$$d_{j,k} = (1 - r) = 1 - \frac{\sum_i \{(e_{ji} - \hat{e}_j)(e_{ki} - \hat{e}_k)\}/n}{\sigma_{e_j} * \sigma_{e_k}}, \tag{2}$$

where $\hat{e}_j$ and σ_{e_j} indicate the mean and standard deviation, respectively, of all points of the jth profile.

Fuzzy c-means [45] is a well-known fuzzy partitive algorithm employed for clustering overlapping data. Use of fuzzy clustering enables genes to simultaneously belong to multiple groups, thereby revealing distinctive features of their function and regulation. Fuzzy c-means algorithm has been applied to cluster microarray data [46]. The value of the fuzzifier m is appropriately tuned for gene selection, based on resultant distribution of distances between genes. The selected genes exhibit tight association to the clusters.

Many proteins serve different functions depending on the demands of the organism, such that a corresponding set of genes is often coexpressed with multiple, distinct groups of genes under different conditions. This type of conditional coregulation of genes is modeled using a heuristically modified version of fuzzy c-means clustering [47], to identify overlapping partitions of genes based on the response of yeast cells to environmental changes.

Kohonen's SOM has been applied to the clustering of gene expression data [66–68]. It generates a robust and accurate clustering of large and noisy data, while providing effective visualization. SOMs require a selected node in the gene expression space (along with its neighbors) to be rotated in the direction of a selected gene expression profile (pattern). However, the predefinition of

a two-dimensional topology of nodes can often be a problem considering its *biological relevance*.

SOTA has also been applied to gene expression clustering [69]. As in SOMs the gene expression profiles are sequentially and iteratively presented at the terminal nodes, and the mapping of the node that is closest (along with its neighboring nodes) is appropriately updated. Upon convergence the node containing the most variable (measured in terms of distance) population of expression profiles is split into sister nodes, causing a growth of the binary tree. Unlike conventional hierarchical clustering, SOTA is linear in complexity to the number of profiles. The number of clusters need not be known in advance as in c-means clustering. The algorithm starts from the node having the most heterogeneous population of associated input gene profiles. A statistical procedure is followed for terminating the growing of the tree, thereby eliminating the need for an arbitrary choice of cutting level as in hierarchical models.

Classification of acute leukemia, having highly similar appearance in gene expression data, has been made by combining a pair of classifiers trained with mutually exclusive features [70]. Gene expression profiles were constructed from 72 patients having acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML), each constituting one sample of the DNA microarray [4] . Each pattern consists of 7129 gene expressions. A neural network combines the outputs of the multiple classifiers. Feature selection with nonoverlapping correlation (such as Pearson and Spearman correlation coefficients) encourages the classifier ensemble to learn different aspects of the training data in a wide solution space.

Fuzzy adaptive resonance theory (ART) network [48] has been employed for clustering the time series expression data related to the sporulation of budding yeast [49].

An evolving modular fuzzy neural network, involving dynamic structure growing (and shrinking), adaptive online learning and knowledge discovery in rule form, has been applied to the Leukemia and Colon cancer gene expression data [72]. Feature selection improves classification by reducing irrelevant attributes that do not change their expression between classes. The Pearson correlation coefficient is used to select genes that are highly correlated with the tissue classes. Rule generation provides physicians, on whom the final responsibility for any decision in the course of treatment rests, with a justification regarding how a classifier arrived at a judgement. Fuzzy logic rules, extracted from the trained network, handle the inherent noise in microarray data while offering the knowledge in a human-understandable linguistic form. These rules point to genes (or their combinations) that are strongly associated with specific types

[4] http://www.genome.wi.mit.edu/MPR

of cancer, and may be used for the development of new tests and treatment discoveries.

A dynamic fuzzy neural network, involving self-generation, parameter optimization and rulebase simplification, is used [71] for the classification of cancer data such as Lymphoma [5], small round blue cell tumor (SRBCT) [6], and liver cancer [7]. Initial feature selection is done in terms of t-tests. It is observed that a small number of important genes (5 out of 4026, 8 out of 2308, 24 out of 1648 features, in the three datasets respectively) succeed in attaining 100% classification.

The identification of gene subsets for classifying two-class disease samples has been modeled as a multiobjective evolutionary optimization problem, involving minimization of gene subset size to achieve reliable and accurate classification based on their expression levels. The Non-Dominated Sorting GA (NSGA-II) [73], a multiobjective GA, is used for the purpose. This employs elitist selection and an explicit diversity preserving mechanism, and emphasizes the non-dominated solutions. It has been shown that this algorithm can converge to the global Pareto front, while simultaneously maintaining the diversity of population.

Results are provided on three cancer samples, *viz.*, Leukemia, Lymphoma and Colon. An l-bit binary string, where l is the number of selected (filtered) genes in the disease samples, represents a solution. The major difficulties faced in solving the optimization problem include the availability of only a few samples as compared to the number of genes in each sample, and the resultant huge search space of solutions. Moreover many of the genes are redundant to the classification decision, and hence need to be eliminated. The three objectives simultaneously minimized are (i) the gene subset size, (ii) number of misclassifications in training, and (iii) number of misclassifications in test samples.

Some recent applications of GAs, in microarray, deal with biclustering. This aims at determining subsets of genes which are similarly expressed over an optimal subset of conditions (or samples), thereby better reflecting the biological reality. Existing greedy algorithms for biclustering often yield suboptimal solutions. GAs are employed [50], by integrating a greedy algorithm as a local search in order to improve the quality of biclustering. Optimization is done with respect to the conflicting goals of homogeneity and size. Results are provided on 2884 genes of yeast data, involving 17 conditions.

[5] http://llmpp.nih.gov/lymphoma/data/figure1/figure1.cdt
[6] http://research.nhgri.nih.gov/microarray/Supplement/
[7] http://genome-www.stanford.edu/hcc/

4 Biological Networks Modeling

Biological networks relate genes, gene products or their groups (like protein complexes or protein families) to each other in the form of a graph, where nodes and edges correspond to molecules and their existing inter-relationships respectively. The Kyoto Encyclopedia of Genes and Genomes (KEGG) database [51] provides a public standardized annotation of genes [8] .

Understanding of regulatory networks is crucial to the understanding of fundamental cellular processes involving growth, development, hormone secretion and cellular communication. Determination of transcriptional factors that control gene expression can offer further insight into the misregulated expressions common in many human diseases. All this is often crucial for application in drug development, medicine, nutrition, and other therapeutic activities.

A genetic regulatory network consists of a set of DNA, RNA, proteins and other molecules, and it describes regulatory mechanisms among them. Regulation of gene expression may occur at any stage of the cellular information flow from DNA, RNA to protein, like mRNA splicing, translational and post-translational control. Nevertheless, the one involving the initiation of transcription has been most widely studied in literature [52–54].

The gene regulatory network determines which subset of genes is expressed, upto what level, and in response to what conditions of the cellular environment. While the metabolic networks form the basis for the net accumulation of biomolecules in living organisms, the regulatory networks modulate their action – thereby leading to physiological and morphological changes. Time-series gene expression data measure mRNA abundance of genes over a sequence of time points, thereby enabling exploration of gene interactions over the entire genome.

Recurrent neural network has been used to model the dynamics of gene expression [55]. The significance of the regulatory effect of one gene product on the expression of other genes of the system is defined by a weight matrix. Multigenic regulation, involving positive and/or negative feedback, are considered. The process of gene expression is described by a single network, along with a pair of linked networks independently modeling the transcription and translation schemes.

Adaptive Double Self-Organizing Map (ADSOM) [56] provides a clustering strategy for identifying gene regulatory networks. It has a flexible topology and allows simultaneous visualization of clusters. DSOM combines features of SOM with two-dimensional position vectors, to provide a visualization tool for

[8] http://www.genome.ad.jp/kegg/

deciding on the required number of clusters. However, its free parameters are difficult to control to guarantee proper convergence. ADSOM updates these free parameters during training, and allows convergence of its position vectors to a fairly consistent number of clusters (provided its initial number of nodes is greater than the expected number of clusters). The effectiveness of ADSOM in identifying the number of clusters is proven by applying it to publicly available gene expression data from multiple biological systems such as yeast, human, and mouse.

Fuzzy rules of an activator-repressor model of gene interactions were used [57] to transform expression values into qualitative descriptors. The algorithm searches for regulatory triplets consisting of the activator, repressor and target genes. However the method is found to be computationally intensive and is limited to determining possible interactions between one positive and one negative regulator per gene. Clustering has been employed [58] as an interface to a fuzzy logic-based method, in order to improve the computational efficiency. Interactions between multiple genes were investigated [59] using a scalable linear variant of fuzzy logic.

Gene regulatory networks were inferred from microarray data [60], using GAs for interactive reverse engineering [9] . The chromosome of the GA corresponds to the floating point weight matrix between the gene time-steps [61]. The average of squared error, over all time-steps, is minimized as the fitness function. The cardinality of the connectivity is also simultaneously minimized. However the combinatorial complexity is expected to be unmanageable in real-world problems, involving a large number of genes [62].

A simple and novel correlation-based approach has been employed to automatically extract gene interaction networks from biclusters in microarray data [63]. The *Pearson correlation coefficient* $-1 \leq r \leq 1$, which measures the linear similarity in trend between two profiles (genes), was used. The distance is given as

$$d_{j,k} = (1 - r) = 1 - \frac{\sum_i \{(e_{ji} - \hat{e}_j)(e_{ki} - \hat{e}_k)\}/N}{\sigma_{e_j} * \sigma_{e_k}}, \tag{3}$$

where $\hat{e}_j$ and σ_{e_j} indicate the mean and standard deviation, respectively, of all points of the jth profile. Preprocessing was done to preserve only the stronger correlated gene interaction pairs. A gene along a weaker link (of correlation) is considered to be not interacting or regulating the other gene. Incorporation of additional knowledge in the form of biclustering helped focus our attention to a smaller subset of genes and/or time points, thereby reducing computational

[9] Reconstructing interactions in gene regulatory networks, using gene expression data.

burden while extracting the network structure. The relationship between the expression level variation over time of a transcription factor and that of its target was analyzed, in the framework of the evolutionary biclusters [64]. The relationship was represented in terms of rules, linking a transcription factor (TF) to the target gene that it regulates. Subsequently these rules were mapped to generate parts of the entire regulatory network.

5 Conclusions

In recent years, Natural Computing models have been successfully applied to several fields of Bioinformatics, including protein structure prediction, microarray data analysis, and biological networks modeling. Natural Computing models demonstrated to be a family of reliable data analysis methods able to transform available heterogeneous biological data into biological knowledge, especially when the task concerns the fusion of multisensor outputs, such as in the case of microarrays, or the fusion of multi-source biological information.

In the coming years the development and the application of new powerful Natural Computing data processing tools will become all the more crucial, given the fast growing volume of available biological data.

References

[1] F.S. Collins, V.A. McKusick, Implications of the Human Genome Project for Medical Science, The Journal of the American Medical Association, vol. 285, pp. 540–544, 2001.

[2] J.D. Watson, The human genome project: past, present, and future, Science (1990) vol. 248, pp. 44–49, 1990.

[3] M. Zvelebil, J. Baum, Understanding Bioinformatics, Garland Science, 2007.

[4] D.W. Mount, Bioinformatics: Sequence and Genome Analysis, Cold Spring Harbor Laboratory Press, 2004.

[5] P. Baldi, S. Brunak, Bioinformatics: The Machine Learning Approach, Second Edition, The MIT Press, 2001.

[6] S. Haykin, Neural Networks. A Comprehensive Foundation, Second Edition, Prentice Hall, Upper Saddle River, NJ, 1999.

[7] C.-T. Lin C.S.G. Lee, Neural Fuzzy Systems: A Neuro-Fuzzy Synergism to Intelligent Systems, Prentice Hall, Upper Saddle River, NJ, 1996.

[8] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2007.

[9] Z. Michalewicz, Genetic Algorithms + Data Structures = Evolution Programs, Springer, 1998.

[10] T. Baeck. Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithm, Oxford University Press, 1996.

[11] R.C. Eberhart, J. Kennedy, Y. Shi, Swarm Intelligence, Elsevier Science, 2001.

[12] Y. Ishida, Immunity-Based Systems, Springer, 2004.

[13] S. Kirkpatrick, C.D. Gelatt and M.P. Vecchi, Optimization by simulated annealing. Science, 220, 661–680 (1983).

[14] S. Mitra, S. Datta, T. Perkins, G. Michailidis, Introduction to Machine Learning and Bioinformatics, Chapman & Hall/CRC Press, 2008.

[15] H. Berman, et al., The protein data bank. Nucleic Acids Research, vol. 28, pp. 235–242, 2000.

[16] J. Berg, J. Tymoczko, L. Stryer, N. Clarke, Biochemistry (5-th ed.). San Francisco, CA: W.H. Freeman & Co., 2002

[17] J. Moult et al., Critical assessment of methods of protein structure prediction-Round VII, Proteins, vol. 69, pp. 3–9, 2007.

[18] Y. Zhang, Progress and challenges in protein structure prediction, Current Opinion in Structural Biology, vol. 18, pp. 342–348, 2008

[19] D. Baker, A. Sali, Protein structure prediction and structural genomics. Science, vol. 294, pp. 93–96, 2001.

[20] N. Qian, T.J. Sejnowski,Predicting the secondary structure of globular proteins using neural network models. Journal of Molecular Biology, vol. 202, pp. 865–884, 1988.

[21] H. Bohr, et al, Protein Secondary Structure and Homology by Neural Networks: The a-helices in Rhodopsin. FEBS Letters, vol. 241, pp. 223, 1988.

[22] L.H. Holley, M. Karplus, Protein secondary structure prediction with a neural network, PNAS, vol. 86, pp. 152–156, 1989.

[23] D.G. Kneller, F.E. Cohen, R. Langridge. Improvements in protein secondary structure prediction by an enhanced neural network. Journal of molecular biology, vol. 214(1), pp. 171–82, 1990.

[24] P. Stolorz, A. Lapedes, Y. Xia, Predicting protein secondary structure using neural net and statistical methods, Journal of Molecular Biology,vol. 225, pp. 363-377, 1992.

[25] F. Sasagawa, K .Tajima, Prediction of protein secondary structures by a neural network, Computer Applications in the Biosciences, vol. 9, pp.147-152, 1993.

[26] P. Fariselli, M. Compiani, R. Casadio, Predicting secondary structures of membrane proteins with neural networks, European biophysics journal vol. 22, pp. 41-51, 1993.

[27] B. Rost, C. Sander, Improved prediction of protein secondary structure by use of sequence profiles and neural networks, Proceedings of the National Academy of Sciences of the United States of America, vol. 90, pp- 7558–7562, 1993.

[28] S.K. Riis, A. Krogh, Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments, Journal of Computational Biology, vol. 3, pp. 163–183, 1996.

[29] D. Jones, Protein secondary structure prediction based on position- specific scoring matrices. Journal of Molecular Biology, vol. 292, pp. 195–202, 1999.

[30] K. Karplus, C. Barrett, R. Hughey, Hidden markov models for detecting remote protein homologies, Bioinformatics, vol. 14, pp. 846–856, 1998.

[31] P. Baldi, et al., Exploiting the past and the future in protein secondary structure predicton, Bioinformatics, vol. 15, pp. 937–946, 1999.

[32] G. Pollastri, A. McLysaght, Porter: A new, accurate server for protein secondary structure prediction, Bioinformatics. vol. 21, pp. 1719-1720, 2005.

[33] G. Pollastri, D. Przybylski, B. Rost, P. Baldi, Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins, vol. 47, pp. 228–235, 2002.

[34] A. Ceroni, P. Frasconi, G. Pollastri, Learning Protein Secondary Structure from Sequential and Relational Data. Neural Networks, vol. 18, pp.1029-39, 2005.

[35] J.T. Pedersen, J. Moult, Protein Folding Simulations with Genetic Algorithms and a Detailed Molecular Description, J. Mol. Biol. vol. 269, pp. 240–259, 1997.

[36] V. Cutello, G. Narzisi, G. Nicosia, A multi-objective evolutionary approach to the protein structure prediction problem, Journal of The Royal Society Interface, vol. 3, pp. 139-151, 2006.

[37] A. Blanco, D.A. Pelta, J.-L. Verdegay, Applying a Fuzzy Sets-based Heuristic to the Protein Structure Prediction Problem, International Journal of Intelligent Systems, vol. 17, pp. 629–643, 2002.

[38] D. Pelta, et al., A fuzzy sets based generalization of contact maps for the overlap of protein structures, Journal of Fuzzy Sets and Systems, vol. 152, pp. 103-123, 2005.

[39] J. Cheng, P. Baldi, A machine learning information retrieval approach to protein fold recognition, Bioinformatics, vol. 22, pp. 1456–1463, 2006.

[40] H. Bohr et al., A novel approach to prediction of the 3-dimensional structures of protein backbones by neural networks, FEBS Letters, vol. 261, pp. 43–46, 1990.

[41] M. Milik, A. Kolinski, J. Skolnick, Neural network system for the evaluation of side-chain packing in protein structures, Protein Engineering, vol. 8, pp. 225-236, 1995.

[42] K.T. Simons, C. Kooperberg, E. Huang, D. Baker, Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions, Journal of Molecular Biology, vol. 268, pp. 209-25, 1997.

[43] J. Meiler, D. Baker, Coupled prediction of protein secondary and tertiary structure, Proceedings of the National Academy of Sciences of the United States of America vol. 100, pp. 12105-10, 2003.

[44] K. Y. Yeung, D. R. Haynor, and W. L. Ruzzo, "Validating clustering for gene expression data," *Bioinformatics*, vol. 17, pp. 309–318, 2001.

[45] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum Press, 1981.

[46] D. Dembele and P. Kastner, "Fuzzy c-means method for clustering microarray data," *Bioinformatics*, vol. 19, pp. 973–980, 2003.

[47] A. P. Gasch and M. B. Eisen, "Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering," *Genome Biology*, vol. 3, pp. research0059.1–0059.22, 2002.

[48] G. A. Carpenter, S. Grossberg, and D. B. Rosen, "Fuzzy ART: fast stable learning and categorization of analog patterns by an adaptive resonance system," *Neural Networks*, vol. 4, pp. 759–771, 1991.

[49] S. Tomida, T. Hanai, H. Honda, and T. Kobayashi, "Analysis of expression profile using fuzzy adaptive resonance theory," *Bioinformatics*, vol. 18, pp. 1073–1083, 2002.

[50] S. Bleuler, A. Prelić, and E. Zitzler, "An EA framework for biclustering of gene expression data," in *Proceedings of Congress on Evolutionary Computation*, pp. 166–173, 2004.

[51] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 27, pp. 29–34, 1999.

[52] H. De Jong, "Modeling and simulation of genetic regulatory systems: A literature review," *Journal of Computational Biology*, vol. 9, pp. 67–103, 2002.

[53] P. D'haeseleer, S. Liang, and R. Somogyi, "Genetic network inference: From co-expression clustering to reverse engineering," *Bioinformatics*, vol. 16, pp. 707–726, 2000.

[54] D. Thieffry, A. M. Huerta, E. Pérez-Rueda, and J. Collado-Vides, "From specific gene regulation to genomic networks: A global analysis of transcriptional regulation in *escherichia coli*," *BioEssays*, vol. 20, pp. 433–440, 1998.

[55] J. Vohradsky, "Neural network model of gene expression," *FASEB Journal*, vol. 15, pp. 846–854, 2001.

[56] H. Ressom, D. Wang, and P. Natarajan, "Clustering gene expression data using adaptive double self-organizing map," *Physiol. Genomics*, vol. 14, pp. 35–46, 2003.

[57] P. J. Woolf and Y. Wang, "A fuzzy logic approach to analyzing gene expression data," *Physiol. Genomics*, vol. 3, pp. 9–15, 2000.

[58] H. Ressom, R. Reynolds, and R. S. Varghese, "Increasing the efficiency of fuzzy logic-based gene expression data analysis," *Physiol. Genomics*, vol. 13, pp. 107–117, 2003.

[59] B. A. Sokhansanj, J. P. Fitch, J. N. Quong, and A. A. Quong, "Linear fuzzy gene network models obtained from microarray data by exhaustive search," *BMC Bioinformatics*, vol. 5, p. 108, 2004.

[60] H. Iba and A. Mimura, "Inference of a gene regulatory network by means of interactive evolutionary computing," *Information Science*, vol. 145, pp. 225–236, 2002.

[61] D. C. Weaver, C. T. Workman, and G. D. Stormo, "Modelling regulatory networks with weight matrices," in *Proceedings of Pacific Symposium on Biocomputing*, pp. 112–123, 1999.

[62] E. Keedwell and A. Narayanan, "Discovering gene networks with a neural-genetic hybrid," *IEEE Transactions on Computational Biology and Bioinformatics*, vol. 2, pp. 231–242, 2005.

[63] S. Mitra, R. Das, H. Banka, and S. Mukhopadhyay, "Gene interaction - An evolutionary biclustering approach," *Information Fusion*, pp. –, 2008.

[64] S. Mitra and H. Banka, "Multi-objective evolutionary biclustering of gene expression data," *Pattern Recognition*, vol. 39, pp. 2464–2477, 2006.

[65] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. NJ: Prentice Hall, 1988.

[66] K. Torkkola, R. M. Gardner, T. Kaysser-Kranich, and C. Ma, "Self-organizing maps in mining gene expression data," *Information Sciences*, vol. 139, pp. 79–96, 2001.

[67] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Smitrovsky, E. S. Lander, and T. R. Golub, "Interpreting patterns of gene expression with self-organizing maps: Methods and applications to hematopoietic differentiation," *Proceedings of National Academy of Sciences USA*, vol. 96, pp. 2907–2912, 1999.

[68] P. Törönen, M. Kolehmainen, G. Wong, and E. Castrén, "Analysis of gene expression data using self-organizing maps," *FEBS Letters*, vol. 451, pp. 142–146, 1999.

[69] J. Herrero, A. Valencia, and J. Dopazo, "A hierarchical unsupervised growing neural network for clustering gene expression patterns," *Bioinformatics*, vol. 17, pp. 126–136, 2001.

[70] S. B. Cho and J. Ryu, "Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features," *Proceedings of the IEEE*, vol. 90, pp. 1744–1753, 2002.

[71] F. Chu, W. Xie, and L. Wang, "Gene selection and cancer classification using a fuzzy neural network," in *Proceedings of 2004 Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS 2004)*, vol. 2, pp. 555–559, 2004.

[72] M. E. Futschik, A. Reeve, and N. Kasabov, "Evolving connectionist systems for knowledge discovery from gene expression data of cancer tissue," *Artificial Intelligence in Medicine*, vol. 28, pp. 165–189, 2003.

[73] K. Deb, S. Agarwal, A. Pratap, and T. Meyarivan, "A fast and elitist multi-objective genetic algorithm : NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, pp. 182–197, 2002.