Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 5, May 2013, pg.202 – 207

RESEARCH ARTICLE

New Applications of Soft Computing, Artificial Intelligence, Fuzzy Logic & Genetic Algorithm in Bioinformatics

Dr. Tryambak A. Hiwarkar[1], R. Sridhar Iyer[2]
[1]Associate Prof, Computer Science and Engineering, MBITM, Dongargarh (C.G.), India
[2]Research Scholar, Computer Science, CMJ University, Shillong, India

Abstract— Soft computing is make several latent in bioinformatics, especially by generating low-cost, low precision (approximate), good solutions. Bioinformatics is an interdisciplinary research area that is the edge between the biological and computational sciences. Bioinformatics pact with algorithms, databases and information systems, web technologies, artificial intelligence and soft computing, information and computation theory, structural biology, software engineering, data mining, image processing, modeling and simulation, discrete mathematics, control and system theory, circuit theory, and statistics.
Bioinformatics is a promise and pioneering research field. Soft Computing is live a crucial role as it give techniques that are particularly well suited to obtain results in an efficient way and with a good level of quality. Soft Computing can also be useful to model the indistinctness and uncertainty that the Bioinformatics data and problems have. In this paper, we survey the role of different soft computing paradigms, like Fuzzy Sets (FSs), Artificial Neural Networks (ANNs), evolutionary computation, Rough Sets (RSs), and Support Vector Machines (SVMs), biologically inspired algorithm like ant colony system, swarm intelligence and others in bioinformatics systems and problems.

Key Terms: - Bioinformatics; Soft computing; artificial neural network; Fuzzy logic; Genetic algorithms

I. INTRODUCTION

Development in soft computing method reveal the high principles of technology, algorithms, and tools in bioinformatics for enthusiastic reason such as dependable and parallel genome sequencing, fast sequence comparison, search in databases, mechanical gene identification, efficient modeling and storage of mixed data, etc. The basic problems in bioinformatics like protein structure forecast, multiple arrangement, phylogenetic inference etc. are mostly NP-hard in nature. For all these problems, soft computing present on promising advance to attain efficient and dependable heuristic solution. On the other side the incessant development of high quality biotechnology, e.g. micro-array techniques and mass spectrometry, which provide complex patterns for the direct characterization of cell processes, offers further promising opportunities for advanced research in bioinformatics. So bioinformatics must cross the border towards a massive integration of the aspects and experience in the different core subjects like computer science and statistics etc. for an integrated understanding of relevant processes in systems biology.

1.1 Why Soft Computing Techniques In Bioinformatics:

Present are a number of reasons why soft computing advance are extensively used in practice, particularly in bioinformatics

1. Usually, a person being builds such an expert system by collect knowledge from specific experts. The experts can always explain what factors they use to charge a condition, however, it is often complicated for the experts to say what rules they use (for example, for disease analysis and control).

2. Systems often create results different from the preferred ones. This may be cause by unknown properties or functions of inputs throughout the design of systems. This situation always occurs in the biological world because of the difficulty and secrecy of life sciences. However, with its ability of dynamic development, soft computing can cope with this problem.
3. In molecular biology research, new data and concepts are generated every day, and those new data and concepts update or replace the old ones. Soft computing can be easily adapted to a changing environment. This benefits system designers, as they do not need to redesign systems whenever the environment changes.
4. Missing and noisy data is one characteristic of biological data. The conventional computer techniques fail to handle this. Soft computing based techniques are able to deal with missing and noisy data.
5. With advances in biotechnology, huge volumes of biological data are generated. In addition, it is possible that important hidden relationships and correlations exist in the data. Soft computing methods are designed to handle very large data sets, and can be used to extract such relationships.

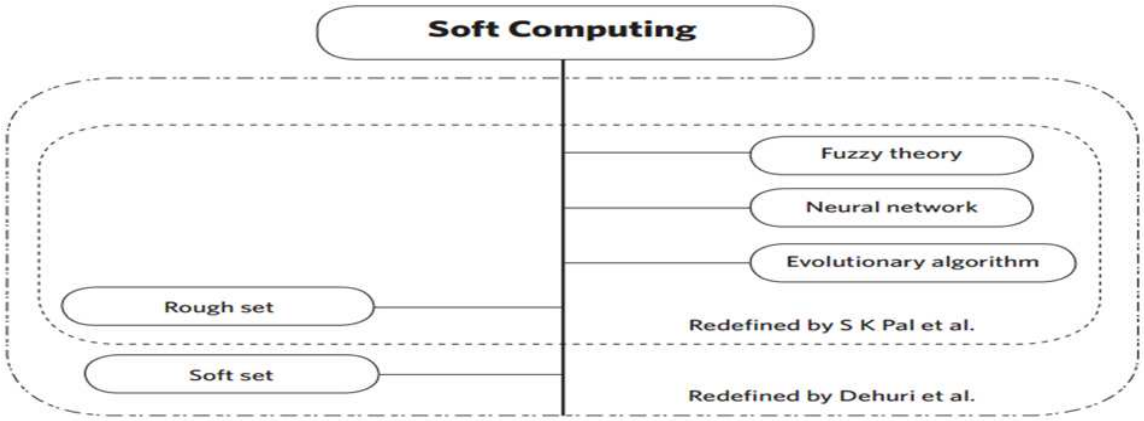

Fig. 1: Soft computing and its basic constituents

1.2 NOTION IN BIOINFORMATICS:

The cell is the basic unit of life (with the exception of viruses). Some organ-isms are single-cell and some are multi-cell. For example, a human being is calm of billions of cells. Although, all the cells are hereditarily identical, they can distinguish into different types of cells with different functionality to work together as a system. These complex systems not only need mechanisms that control the cells to work jointly, but also require device that regulate the processes within a cell. As a result, these intra-cellular and inter-cellular systems in turn define the organism. In fact, all this information is programmed in the genetic material and is passed down in generations. For many organisms, DNA is the genetic material, which contains all the essential information about this organism. RNA and protein are important intermediates that carry out the 'instructions' written in DNA. Be-low we will discuss DNA, RNA and protein in detail and explain the relationships between them.

The Deoxyribo Nucleic Acid is where the genetic information is stored. The DNA is a chain of small molecules called nucleotides. Each nucleotide is composed of three molecular components, a five-carbon sugar, a phosphate group and a nitrogenous base. In addition, there are four types of nitrogenous bases: adenine (A), thymine (T), guanine (G) and cytosine (C). Figure I.1 gives the molecular diagram of a DNA strand. The genetic information is encoded by alternating these four bases in the DNA strand. Moreover, this genetic information is usually represented as a DNA sequence, a string composed of characters from the alphabet{A, T, G, C}.A DNA strand is oriented, where one end is called the 5'-end (five prime end) and the other end is called the 3'-end (three prime end). Notice that some DNA strands are linear and some are circular. Regardless of the starting point of a DNA strand, the convention is to write the DNA sequence from the 5'-end to the 3'-end. That is, if we write TCGTA, the left end is the 5'-end and the right end is the 3'-end,i.e. 5'-TCGTA-3'. The length of a DNA strand can vary from a few thousands base pairs to a few millions base pairs in different organisms. DNA strands come in pairs and they are called complementary strands. Two complementary DNA strands complement of each other in the opposite orientation. Base and base Tare complements of each other, while base G and base C are complements. Two complementary DNA strands are bonded together in the shape of a double-helix by hydrogen bonds. Figure I.2 shows the double-helix complementary structure of DNAs. In fact, there are other higher level structures of the DNA strand, such as the super cooling structure. Except during some cellular events (such as transcription or duplication), DNA strands are coiled and packed up with a special type of protein called histone. This super-structure is visible under the electron micrograph and it is called the chromosome.

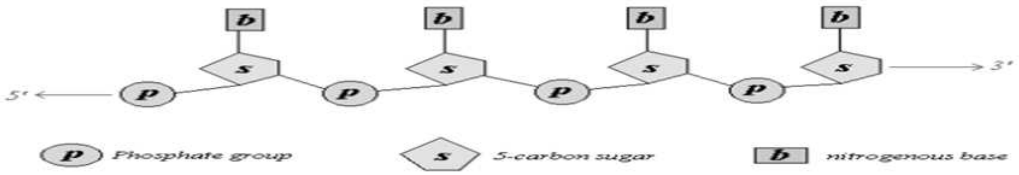

Figure I.1. The Molecular Structure of DNAs

Figure I.2. The Double Helix Structure of DNAs

1.3 UTILIZETION OF SOFT COMPUTING IN BIOINFORMATICS:

As soft computing are measured to handle vagueness, indecision and near optimality in large and complex search spaces use of soft computing gear for solving bioinformatics problems have been gained the attention of researchers., it involves genomic sequence, protein structure, gene expression microarray, and gene regulatory networks [17]. Most of the researches are woven around the tasks of pattern recognition and data mining like clustering, classification, feature selection, and rule generation, while classification pertains to supervised or unsupervised learning, clustering corresponds to unsupervised self -organization into homologous partitions. Feature selection techniques [24] aim at reducing the number of irrelevant and redundant variables in the dataset. Rule generation enables efficient representation of mined knowledge in human-understandable form. Many intangible parameters are mathematically modeled.

II. APPLICATION OF ARTIFICIAL NEURAL NETWORK IN BIOINFORMATICS

An Artificial Neural Network (ANN) [1] is an information processing model that is able to capture and represent complex input-output relationships. The motivation the development of the ANN technique came from a desire for an intelligent artificial system that could process information in the same way the human brain. Its novel structure is represented as multiple layers of simple processing elements, operating in parallel to solve specific problems. ANNs resemble human brain in two respects: learning process and storing experiential knowledge. An artificial neural network learns and classifies problem through repeated adjustments of the connecting weights between the elements.

An ANN learns from examples and generalizes the learning beyond the examples supplied. Artificial neural network [2,3] applications have recently received considerable attention. The methodology of modeling or estimation [4] is somewhat comparable to statistical modeling. Neural networks should not, however, be heralded as a substitute for statistical modeling but rather as a complementary effort (without the restrictive assumption of a particular statistical model) or an alternative approach to fitting non-linear data .Neural networks have been widely used in biology since the early 1990s. They can be used to:

(a) Prediction and the translation sites initiation in DNA sequences and proteins [2, 5].
(b) Explain the theory of artificial neural networks using applications in biology [5].
(c) Predict immunologically interesting peptides by combining an evolutionary algorithm [6].
(d) Study human TAP transporter [7].
(e) Carry out pattern classification and signal processing successfully in bioinformatics [8].
(f) Perform protein sequence classification [6,].
(g) Predict protein secondary structure prediction [7]

III. FUZZY LOGIC IN BIOINFORMATICS

Fuzzy logic [28,29] can be easily used to implement systems ranging from simple, small or even embedded up to large networked ones. Fuzzy logic is that it accepts the uncertainties that are inherited in the realistic inputs and it deals with these uncertainties in their affect is negligible and thus resulting in a precise outputs. Fuzzy Logic reduces the design steps and simplifies complexity that might arise since the first step is to understand and characterize the system behavior by using knowledge and experience. The concept of Fuzzy Logic (FL) was conceived by Lotfi Zadeh, FL provides a simple way to arrive at a definite conclusion based upon vague, ambiguous, imprecise, noisy, or missing input information. It mimics human control logic [9].

Fuzzy systems have been successfully applied to several areas in practice like for building knowledge-based systems, fuzzy logic-based and fuzzy rule-based models. They can control and analyze processes and diagnose and make decisions in biomedical sciences. There are many application areas in biomedical science and bioinformatics, where fuzzy logic techniques [10] can be applied successfully. Some of the important uses of fuzzy logic are listed below:

(a) Increasing flexibility of protein motifs [9].
(b) Studying differences between various poly nucleotides [11].
(c) Analyzing experimental expression data [6] using fuzzy adaptive resonance theory [9].
(d) Studying aligning sequences based on a fuzzy dynamic programming algorithm [7,4].
(e) Mathematical modeling of complex traits influenced by genes with fuzzy-valued in pedigreed populations.
(f) Finding cluster membership values to genes applying a fuzzy partitioning method using fuzzy C-Means and fuzzy c-hard mean algorithms [7].
(g) Generating DNA sequencing using genetic fuzzy and neuro-fuzzy systems by anticipating disturbances due to intangible parameters [9].
(h) Identifying the cluster genes from micro-array data [8].
(i) Predicting protein's sub-cellular locations fuzzy k- nearest neighbor's algorithm.
(j) Mapping specific sequence patterns to putative functional classes since evolutionary comparison leads to functional characterization of hypothetical proteins.
(k) Developing gene expression data.

IV. GENETIC ALGORITHMS TECHNIQUES IN BIOINFORMATICS

Genetic algorithms [12] (GA), are randomized search and optimization techniques guided by the principles of evolution and natural genetics [13]. The applications of GAs are for solving certain multi objective problems of bioinformatics, which yields optimization of computation requirements, and robust, fast and close approximate solutions. Moreover, the errors generated in experiments with bioinformatics data can be handled with the robust characteristics of GAs. To some extent, such errors may be regarded as contributing to genetic diversity, a desirable property. The problem of integrating GAs and bioinformatics constitutes a new research area. GAs [14] are executed iteratively on coded solutions (population) biological basic Operators: selection/reproduction, crossover, and mutation. They use objective function information and probabilistic transition rules for moving to the next iteration. Of all the evolutionarily inspired approaches, Gas seem particularly suited to implementation using DNA, protein and other bioinformatics tasks. This is because GAs is generally based on manipulating populations of bit-strings using both crossover and point-wise mutation.

The most suitable applications of GAs in bioinformatics are:

(a) Alignment and comparison of DNA, RNA, and protein sequences [4, 10].
(b) Gene mappings in chromosomes.
(c) RNA structure prediction
(d) Protein structure prediction and clustering [14].
(e) Molecular design and molecular docking [14].
(f) Gene finding and promoter identification from DNA sequences.
(g) Interpretation of gene expression and micro array data [15].
(h) Gene regulatory network identification [15].
(i) Construction of phylogenetic tree for studying evolutionary relationship [13].
(j) DNA structure prediction [13].

V. PHRASE OF SWARM INTELLIGENCE IN BIOINFORMATICS

Historically, the phrase Swarm Intelligence (SI) was coined by Beny & Wang in late 1980's [13] in the context of cellular robotics. SI systems are typically made up of a population of simple agents (an entity capable of performing/executing certain operations) interacting locally with one another and with their environment. Although there is normally no centralized control structure dictating how individual agents should behave, local interactions between such agents often lead to the emergence of global behavior [16]. Many biological creatures such as fish schools and bird flocks clearly display structural order, with the behavior of the organisms so integrated that even though they may change shape and direction, they appear to move as a single coherent entity. The main properties of the collective behavior can be given below: Individuals attempt to maintain a minimum distance between themselves and others at all times. This rule is given the highest priority and corresponds to a frequently observed behavior of animals in nature. If individuals are not performing, an avoidance maneuver they tend to be attracted towards other individuals (to avoid being isolated) and to align themselves with neighbors. The structure and function of fish schools identified four collective dynamical behaviors as [17]:

• Swarm: an aggregate with cohesion, but a low level of polarization (parallel alignment) among members.

• Torus: individuals perpetually rotate around an empty core (milling). The direction of rotation is random.

• Dynamic parallel group: the individuals are polarized and move as a coherent group, but individuals can move throughout the group and density and group form can fluctuate.

• Highly parallel group: much more static in terms of exchange of spatial positions within the group than the dynamic parallel group and the variation in density and form is minimal. A swarm can be viewed as a group of agents cooperating to achieve some purposeful behavior and achieve some goal. This collective intelligence seems to emerge from what are often large groups. According to Milonas, five basic principles define the swarm intelligence paradigm, phase transitions, and collective intelligence [1 ,18]:

(a) The proximity principle: the swarm should be able to carry out simple space and time computations.

(b) The quality principle: the swarm should be able to respond to quality factors in the environment.

(c) The principle of diverse response: the swarm should not commit its activities along excessively narrow channels.

(d) The principle of stability: the swarm should not change its mode of behavior every time the environment changes [49].

(e) The principle of adaptability: the swarm must be able to change behavior mote when it is worth the computational price.

VI. CONCLUSION

Through a volatile increase of the gloss genomic sequences in available form, bioinformatics has materialized as a challenging and attractive field of science. It presents the ideal agreement of statistics, biology and computational aptitude method for examine and dispensation biological in order in the form of gene, DNA, RNA and proteins. Soft computing algorithms on the other hand, have recently gained wide status among the researchers, for their amazing ability in finding near best answer to a number of NP hard, real world search problems.

Traditional deterministic search algorithms and the copied based optimization method are of no use for them as the search space may be enormously large and irregular at several points.

REFERENCES

[1] N. Qian and T.J. Sejnowski; "Predicting the secondary structure of globular proteins using neural network models", J. Mol. Biol., Vol. 202(4), pp. 865-884, 1988.

[2] V. Kecman; "Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models. Complex Adaptive Systems", Cambridge, MA: MIT Press, 2001 .

[3] S. K. Pal and S. Mitra; "Neuro-Fuzzy Pattern Recognition: Methods in Soft Computing", New York: Wiley, 1999.

[4] D. Wang and G. B. Huang; "Protein sequence classification using extreme learning machine",Proc. Int. Joint Conf. Neural Networks (IJCNN'05), Montreal, QC, Canada, pp. 1406-1411, August 2005

[5] Y. Huang and Y. Li; "Prediction of protein subcellular locations using fuzzy k-NN method", Bioinformatics, Vol. 20(1), pp. 21-28, 2004.

[6] Zou Xiu-fen, Pan Zi-shu, Kang Le-shan and Zhang Chu-yu; "Evolutionary computation techniques for Protein structure prediction: A Survey", Wuhan University Journal of Natural Sciences, Vol. 8(1B), 2003

[7] E.E. Snyder and G.D. Stormo; "Identification of protein coding regions in genomic DNA", J. Mol. Biol., Vol. 248, pp. 1 -18, 1995.

[8] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman; "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs", Nucleic. Acids Res., Vol. 25, pp. 3389-3402, 1997

[9] L. A. Zadeh; "Fuzzy logic, neural networks, and soft computing", Commun. ACM, Vol. 37, pp. 77-84, 1994.

[10] G. Pollastri, A.J. Martin, C. Mooney and A. Vullo; "Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information", BMC Bioinform., Vol. 8, pp. 201, 2007.

[11] Y. Huang and Y. Li; "Prediction of protein subcellular locations using fuzzy k-NN method", Bioinformatics, Vol. 20(1), pp. 21-28, 2004

[12] Sushmita Mitra and Yoichi Hayashi; "Bioinformatics with Soft Computing", IEEE transactions on systems, man, and cybernetics—part c: Applications and Reviews, Vol. 36(5), September 2006.

[13] D.E. Goldberg; "Genetic Algorithms in Search, Optimization and Machine Learning", Reading, MA: Addison- Wesley, 1989

[14] Zou Xiu-fen, Pan Zi-shu, Kang Le-shan and Zhang Chu-yu; "Evolutionary computation techniques for Protein structure prediction: A Survey", Wuhan University Journal of Natural Sciences, Vol. 8(1B), 2003

[15] G. Fogel and D. Corne (eds.); "Evolutionary Computation in Bioinformatics", San Francisco, CA: Morgan Kaufmann, 2002.

[16] Swagatam Das, Ajith Abraham, and Amit Konar; "Swarm Intelligence Algorithms in Bioinformatics", (SCI) 94, pp. 113-147, Springerlink., 2008.

[17] M. Carolina Hinestrosa, Kay Dickersin, Pamela Klein, Musa Mayer, Karin Noss, Dennis Slamon, George Sledge and Frances M. Visco; "Shaping the future of biomarker research in breast cancer to ensure clinical relevance", Nature Reviews Cancer, Vol. 7, pp. 309-315, 2007

[18] Rabindra Ku. Jena, Musbah M. Aqel, Pankaj Srivastava and Prabhat K. Mahanti; "Soft Computing Methodologies in Bioinformatics", European Journal of Scientific Research, Vol.26(2), pp. 189-203, 2009.

