Application of Data mining and Soft Computing in Bioinformatics

[1]P.K.Vaishali, [2]Dr.A.Vinayababu
[1]Department of Computer Science & Information Technology, Jyothishmathi Institute of Tech & Sciences, JNTU, Hyderabad, AP, INDIA.
[2]Professor of CSE, Director of Admissions JNTUH University, Hyderabad, AP, INDIA.

Abstract

The explosive growth of biological information generated by the scientific community all over the world has led to storage of voluminous data. This torrent of information has, in turn, led to an supreme requirement for computerized databases to store, categorize, and index the data and for specialized tools to view and analyze the data. Bioinformatics is the analysis of biological information using computers and statistical techniques; the science of developing and utilizing computer databases and algorithms to accelerate and enhance biological research. Computers are used to gather, store, analyze and integrate biological and genetic information which can then be applied to gene-based drug discovery and development. Data mining or knowledge discovery from data (KDD), is used to extract interesting, nontrivial, implicit, previously unknown and potentially useful information from data. Soft computing differs from conventional (hard) computing in that, unlike hard computing, it is tolerant of imprecision, uncertainty, partial truth, and approximation. The guiding principle of soft computing is to exploit the tolerance for imprecision, uncertainty, partial truth, and approximation to achieve tractability, robustness and low solution cost for the emerging field of conceptual intelligence. Soft computing includes the techniques such as fuzzy logic, neural networks, genetic algorithms,etc.The methodologies of soft computing are complementary rather than competitive and they can be viewed as a foundation component. This paper will focus on issues related to data mining and soft computing and relevance of these in bioinformatics. Further the paper focuses on some of its current applications.

Keywords: Datamining,SoftComputing,Bioinformatis , Fuzzy Logic, Neural Networks, Genetic Algorithm

I.INTRODUCTION

Bioinformatics, is a field committed to the interpretation and analysis of biological data using computational techniques, has evolved tremendously in recent years due to the explosive growth of biological information generated by the scientific community. Bioinformatics is the science of managing, mining, integrating, and interpreting information from biological data at the genomic, proteomic, phylogenetic, cellular, or whole organism levels. The need for bioinformatics tools and expertise has increased as genome sequencing projects have resulted in an exponential growth in complete and partial sequence databases. Data mining is the use of automated data analysis techniques to uncover previously undetected relationships among data items. Data mining often involves the analysis of data stored in a data warehouse. Data mining aids the scientists and the researches by providing sophisticated techniques to extract the useful information from the huge amount biological data at hand. Data mining refers to the extraction of useful information from a large set of data. It is a technique for the discovery of patterns hidden in large data sets, focusing on issues relating to their feasibility, usefulness, effectiveness, and scalability[2]. The term data mining refers to information elicitation. On the other hand, soft computing deals with information processing. If these two key properties can be combined in a constructive way, then this formation can effectively be used for knowledge discovery in large databases. Referring to this synergetic combination, the basic merits of data mining and soft computing paradigms are pointed out and novel data mining implementation coupled to a soft computing approach for knowledge discovery is presented. In this context, in the following two sections the properties of data mining and machine learning paradigms are pointed out. The present article provides an overview of the available literature on data mining, and its aspects. This is followed by Section -2 which discusses the state of art of soft computing and we discuss each of the soft computing methods in brief .the following by section-4 which brings out the relevance of different soft computing methods in data mining .finally we conclude with section -4, where the significane of soft computing in data mining is highlighted.[2][4].The purpose of this paper is to provide an overall understanding of Data mining and soft computing techniques and their application and usage in bioinformatics.

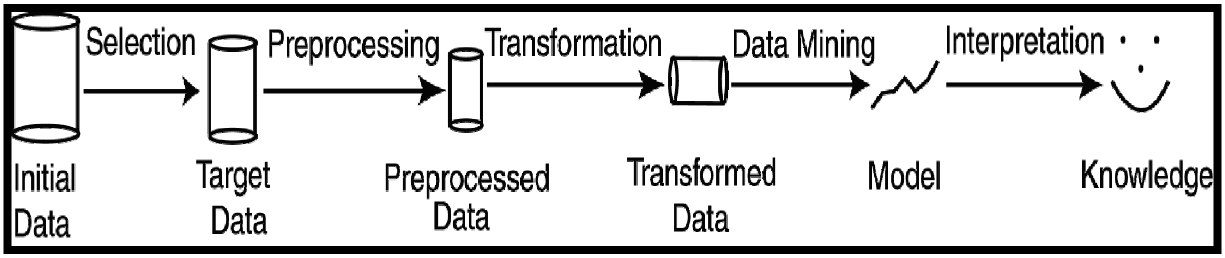
II. Bioinformatics:

Bioinfomatics is the field of science in which biology, computer science, and information technology merge to form a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned. At the beginning of the "genomic revolution", a bioinformatics concern was the creation and maintenance of a database to store biological information, such as nucleotide and amino acid sequences. Development of this type of database involved not only design issues but the development of complex interfaces whereby researchers could both access existing data as well as submit new or revised data. Ultimately, however, all of this information must be combined to form a comprehensive picture of normal cellular activities so that researchers may study how these activities are altered in different disease states. Therefore, the field of bioinformatics has evolved such that the most pressing task now involves the analysis and interpretation of various types of data, including nucleotide and amino acid sequences, protein domains, and protein structures. The actual process of analyzing and interpreting data is referred to as **computational biology**. Important sub-disciplines within bioinformatics and computational biology include the development and implementation of tools that enable efficient access to, and use and management of, various types of information the development of new algorithms (mathematical formulas) and statistics with which to assess relationships among members of large data sets, such as methods to locate a gene within a sequence, predict protein structure and/or function, and cluster protein sequences into families of related sequences.The National Center for Biotechnology Information (NCBI 2001) defines bioinformatics as:"Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline. There are three important sub-disciplines within bioinformatics: the development of new algorithms and statistics with which to assess relationships among members of large data sets; the analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures; and the development and implementation of tools that enable efficient access and management of different types of information."[ref-3]. The three terms bioinformatics, computational biology and bioinformation infrastructure are often times used interchangeably. These three may be defined as follows:

1. Bioinformatics refers to database-like activities, involving persistent sets of data that are maintained in a consistent state over essentially indefinite periods of time;
2. Computational biology encompasses the use of algorithmic tools to facilitate biological analysis.
3. Bioinformation infrastructure comprises the entire collective of information management systems, analysis tools and communication networks supporting biology. Thus, the latter may be viewed as a computational gibbet of the former two.

A. Biological Database

A biological database is a large, organized body of persistent data, usually associated with computerized software designed to update, query, and retrieve components of the data stored within the system. A simple database might be a single file containing many records, each of which includes the same set of information. For example, a record associated with a nucleotide sequence database typically contains information such as contact name, the input sequence with a description of the type of molecule, the scientific name of the source organism from which it was isolated, and often, literature citations associated with the sequence. For researchers to benefit from the data stored in a database, two additional requirements must be met: easy access to the information and a method for extracting only that information needed to answer a specific biological question.



Data mining task :Figure :1

B. Importance of Bioinformatics

The rationale for applying computational approaches to facilitate the understanding of various biological processes includes: a more global perspective in experimental design the ability to capitalize on the emerging technology of database-mining - the process by which testable hypotheses are generated regarding the function or structure of a gene or.protein of interest by identifying similar sequences in better characterized organisms

C. SCOPE AND USE OF BIOINFORMATICS:

Bioinformatics is used in analyzing genomes protein sequences, three-dimensional modelling of bio molecules and biologic systems, etc. Different biological problems considered within the scope of bioinformatics fall into main tasks which are given below

- Alignment and comparison of DNA, RNA, and protein sequences.
- Gene finding and promoter identification from DNA sequences.
- Interpretation of gene expression and micro-array data.
- Gene regulatory network identification.
- Construction of phylogenetic trees for studying evolutionary relationship.
- Protein structure prediction and classification.
- Molecular design and molecular docking.

Therefore the aims of bioinformatics are:
• To organize data in a way that allows researchers to create and access information
• To develop tools that facilitates the analysis and management of data.
• To use biological data to analyse and interpret the results in a biologically meaningful manner

III. DATA MINING

Data mining in general is a term which refers to extracting or “mining” knowledge from huge amounts of records. Data Mining (DM) is the discipline of finding novel remarkable patterns and liaison in vast quantity of data. Data mining is also sometimes called Knowledge Discovery in Databases (KDD)[4].Data mining is not specific to any industry,it has its presence felt inmost of the applications today. It requires sharp technologies and the compliance to discover the possibility of hidden knowledge that resides in the data.Data Mining approaches seem ideally suited for Bioinformatics, since it is data-rich, but lacks a comprehensive theory of life’s organization at the molecular level. The extensive databases of biological information crafts both challenges and opportunities for development of novel KDD methods. Mining biological data helps to extract useful knowledge from massive datasets gathered in biology, and in other related life sciences areas such as medicine and neuroscience ,etc. Knowledge discovery as a process is depicted in Figure 1 and consists of an iterative sequence of the following steps:

- Selection: Obtain data from various sources.
- Preprocessing: Cleanse data.
- Transformation: Convert to common format. Transform to new format.
- Data Mining: Obtain desired results.
- Interpretation/Evaluation: Present results to user in meaningful manner.

A. Data mining tasks:

The two "high-level" primary goals of data mining, in practice, are prediction and description. The main

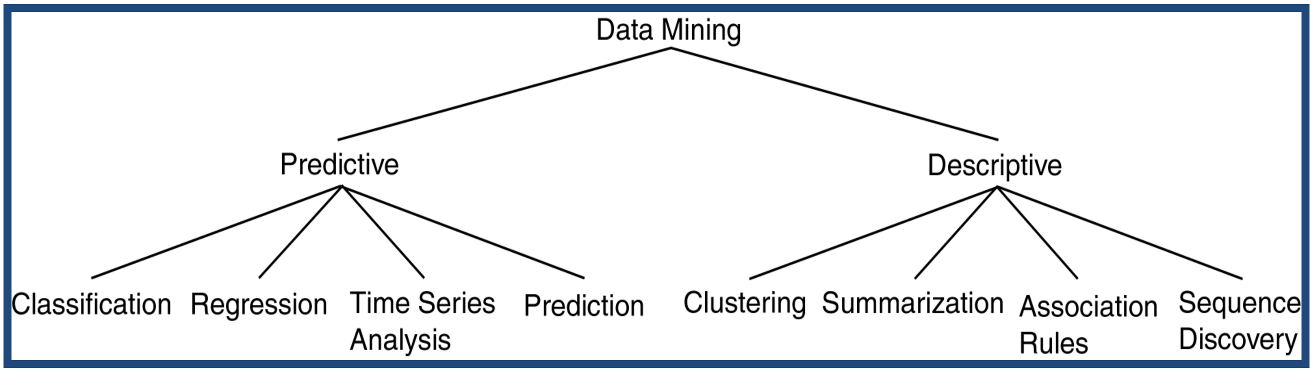


Figure 2:

tasks well apt for data mining, all of which involves mining meaningful new patterns from the data, are: Learning from data falls into two categories: directed (“supervised”) and undirected (“unsupervised”) learning given in figure2.The first three tasks – classification, estimation and prediction – are relationship among all the variables. Unsupervised learning attempts to find patterns without the use of a particular target field. The development of new data mining and knowledge discovery tools is a subject of active research. One motivation behind the development of these tools is their potential application in modern biology.

Classification

Classification is learning a function that maps (classifies) a data item into one of several predefined classes. 1. Classification [18], [19], [20], [21], [22]: classifies a data item into one of several predefined categorical classes. A classification problem is a supervised learning problem in which the output information is a discrete classification; i.e., given **an** object and its input attributes, the classification output is one of the possible mutually exclusive classes of the problem. The aim of the classification task is to discover some kind of relationship between the input attributes and the output class, so that the discovered knowledge can be used to predict the class of a new unknown object.

Regression

[8] [4] ,[10], [11], [12]: maps a data item to a real valued prediction variable. A regression problem is a supervised learning problem of building a more or

examples of supervised learning. The next three tasks – association rules, clustering and description & visualization – are examples of unsupervised learning. In unsupervised learning, no variable is singled out as the target; the goal is to establish some

less transparent model in which the output information is a continuous numerical value or a vector of such values rather than a discrete class. Then, given an object, it is possible to predict one of its attributes by means of the other attributes by using the built model. The prediction of numeric values may be done by classical or more advanced statistical methods and by symbolic methods often used in the classification task.

Estimation

Given some input data, coming up with a value for some unknown continuous variable.

Prediction

Same as classification & estimation except that the records are classified according to some future behavior or estimated future value).

Association rules

Determining which things go together, also called dependency modeling. association rules [13], [14], [15], [16]: describes association relationship among different attributes. Given a collection of items and a set of records, each of which contain some number of items from the given collection, an association function is an operation against this set of records which return affinities or patterns that exist among the collection of items. These patterns can be expressed by rules such as "72% of all the records that contain items A, B and C also contain items D and E." Association rule generators are a powerful data mining technique used to search through an entire data set for rules revealing the nature and frequency of relationships or association between data entities. The resulting associations can be used to filter the information for human analysis and possibly to define a prediction model based on observed behaviour.

Clustering

Segmenting a population into a number of subgroups or clusters. . Clustering [17], [18], [19], [20], [21], [22], [23], [24]:maps a data item into one of several clusters, where clusters are natural groupings of data items based on similarity metrics or probability density models

Rule generation

[25], [26], [27], [28], [29], [30], [31], [32]: extracts classification rules from the data. **A** clustering problem is an unsupervised learning problem that aims at finding clusters of similar objects sharing a number of interesting properties. It may be used in data mining to evaluate similarities among data, to build a set of representative prototypes, to analyze correlations between attributes, or to automatically represent a data set by a small number of regions, preserving the topological properties of the original input space

Summarization

[33], [34], [35], [36]: provides a compact description for a subset of data This task aims at producing compact and characteristic descriptions for a given set of data. It can take multiple forms: numerical (simple descriptive statistical measures like means, standard deviations), graphical forms (histograms, scatter plots), or the form of if-then rules. It may provide descriptions about objects in the whole database or in selected subsets of it. ***An*** example of summarization is “the minimum unitary price for all the transactions with energy is 70 price units”

Sequence analysis

[37], [38]: models sequential patterns,like time-series analysis. The goal is to model the states of the process generating the sequence or to extract andreport deviation and trends over time.

Dependency modelling

[37], [38]: describes significant dependencies among variables.

Description & visualization*:*

Representing the data using visualization techniques.

B.

C. Major Issues in Data Mining

T he following are the major issues related with the usage of the data mining functionalities:

1.Mining methodology and user interaction

Mining different kinds of knowledge in databases.
Interactive mining of knowledge at multiple
levels of abstraction.

2.Performance Scalability

Efficiency and scalability of data mining algorithms.
Parallel, distributed and incremental mining methods

3.Other Issues in Data Mining

Issues relating to the diversity of data types. Handling relational and complex types of data. Mining information from heterogeneous databases and global information systems (WWW). Issues related to applications and social impacts.. Application of discovered knowledge. Domain-specific data mining tools. Intelligent query answering. protection of data security, integrity, and privacy. Integration of the discovered knowledge with existing knowledge:
A knowledge fusion problem.

D. MAJOR CHALLENGES IN DATA MINING:

1.Mining methodology and user interaction :

Mining different kinds of knowledge in databases. Interactive mining ofknowledge at multiple levels of abstraction .Incorporation of background knowledge. Data mining query languages and ad-hoc data mining.Expression and visualization of data mining results Handling noise.

E. DATA MINING APPLICATION AREAS:

Data Mining techniques have been applied successfully in many areas from business to science to sports.

1. .Science applications

Data mining techniques have been used in astronomy, bioinformatics, drug discovery and many more.

2. Business applications

Many organizations now employ data mining as a secreat weapon to keep in pace or gain a competitive edge .Data mining has been used in advertising, CRM (Customer Relationship management), investments, manufacturing, sports/entertainment, telecom, e-Commerce, targeted marketing, health care, etc.

3. Other application

Data mininng has been successfully used for various other application such as Web: search engines,, Government ,law enforcement, profiling tax cheaters, anti-terror, credit approval, etc. Putting it together Data Mining is the step in the process of knowledge discovery in databases, that inputs predominantly cleaned, transformed data, searches the data using algorithms, and outputs patterns and relationships to the interpretation/evaluation step of the KDD process

IV Soft computing:

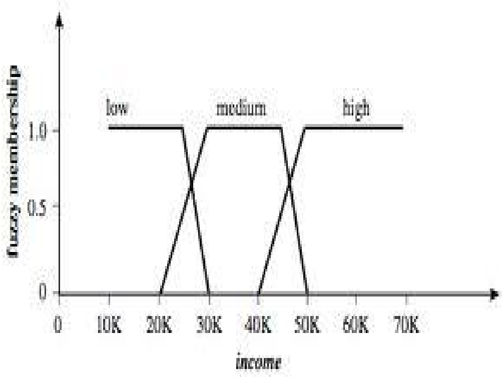
Soft computing is a consortium of methodologies that works synergistically and provides, in one form or another, flexible information processing capability [10][]. Its aim is to exploit the tolerance for imprecision, uncertainty, approximate reasoning, and partial truth in order to achieve tractability, robustness, and low-cost solutions[11][]. Methodologies like fuzzy sets, neural networks, genetic algorithms, and rough sets are most widely applied in the data. An excellent survey demonstrating the significance of soft computing tools in data mining problem is provided by Mitra et al. [40]. These methodologies of soft computing are complementary rather than competitive and they can be viewed as a foundation

A. Importance of Soft Computing

The complementarity of fuzzy logic, neural networks, genetic algorithms and probabilistic reasoning has an important consequence in many cases. A problem can be solved most effectively by using fuzzy logic, neural networks, genetic algorithms and probabilistic reasoning in combination rather than using them exclusively. A typical example for such combination is *neurofuzzy systems*.

Fuzzy Logic

Fuzzy logic is derived from fuzzy set theory dealing with reasoning that is approximate rather than precisely deduced from classical predicate logic. It can be thought of as the application side of fuzzy set. Fuzzy truth represents membership in vaguely defined sets, like likelihood of some event or condition and fuzzy sets are based on vague definitions of sets, not randomness [39]. Since, data mining is the process of extracting nontrivial relationships in the database, the association that are qualitative are very difficult to utilize effectively by applying conventional rule induction algorithms. Since fuzzy logic modeling is a probability based modeling, it has many advantages over the conventional rule induction algorithms. The advantage is that it allows processing of very large data sets which require efficient algorithms.[ref book] The concept of fuzzy logic was conceived by Lotfi Zadeh. Fuzzy logic is an organized method for dealing with imprecise data. This data is considered to be as fuzzy sets[2][]. Fig.2 shows how values for the continuous attribute income are mapped into the discrete categories flow, medium, high, as well as how the fuzzy membership or truth values are calculated. Fuzzy logic systems typically provide graphical tools to assist users in this step.



Fuzzy logic can be introduced into the system to allow fuzzy" thresholds or boundaries to be defined. Rather than having a precise cutoffs between categories or sets, fuzzy logic uses truth values between 0:0 and 1:0 to represent the degree of membership that a certain value has in a given category[2][]. In general, the use of fuzzy logic in rule-based systems involves the following:

1) Attribute values are converted to fuzzy values. Above figure shows how values for the continuous attribute income are mapped into the discrete categories flow, medium, high, as well as how the fuzzy membership or truth values are calculated. Fuzzy logic systems typically provide graphical tools to assist users in this step.

2) For a given new sample, more than one fuzzy rule may apply. Each applicable rule contributes a vote for membership in the categories. Typically, the truth values for each predicted category are summed.

3)The sums obtained above are combined into a value that is returned by the system. This process may be done by weighting each category by its truth sum and multiplying by the mean truth value of each category. The calculations involved may be more complex, depending on the complexity of the fuzzy membership graphs.

1. Role of Fuzzy Sets

Modeling of imprecise/qualitative knowledge. Transmission and handling uncertainties at various stages. Supporting, to an extent, human type reasoning in natural form.

2. Fuzzy Sets in Data Mining

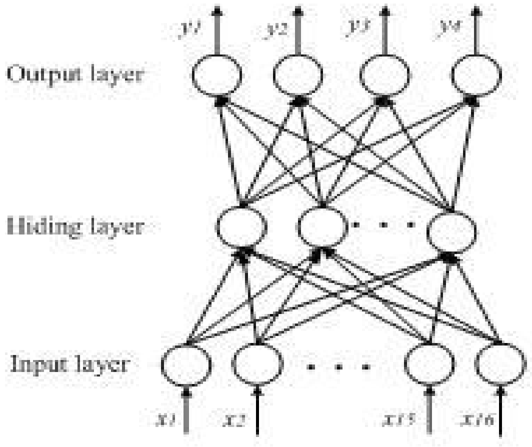
Classification/Regression/ Clustering. Discovering association rules (describing interesting association relationship among different attributes). Data summarization/condensation (abstracting the essence from a large amount of information).

Neural Networks

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain processes information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in union to solve specific problems. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons [41].

NEURAL NETWORKS

An artificial neural network (ANN), [2][], [4][] often just called a "neural network" (NN), is a mathematical model or computational model based on biological neural networks, in other words, is an emulation of biological neural system. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. The human brain is a highly complex structure viewed as a massive , highly interconnected network of simple processing elements called neurons. This behaviour of the neuron can be captured by a simple model which can be shown as:



In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the netwo Every component of the model bears a direct analogy to the actual constituents of a biological neuron and hence is termed as Ann. There are several classes of Neural Networks, classified according to their learning mechanisms.

a.Single layer feed forward network

b.Multilayer feed forward network:

c.Recurrent networks:

During the learning phase

Role of ANN

- Adaptivity, robustness, parallelism, optimality. Machinery for learning and curve fitting (Learns from examples).
- Initially, thought to be unsuitable for —black box nature —no information available in symbolic form (suitable for human interpretation).
- Recently, embedded knowledge is extracte in the form of symbolic rules making it suitable for rule generation.

ANNs provide Natural Classifiers/prediction based models having

- Resistance to Noise,
- Tolerance to Distorted Patterns /Images (Ability to Generalize).
- Superior Ability to Recognize Overlapping.
- Pattern Classes or Classes with Highly.
- Nonlinear Boundaries or Partially Occluded or Degraded Images.
- Potential for Parallel Processing.
- Non parametric

Genetic Algorithms

Genetic Algorithms have found a wide gamut of applications in data mining, where knowledge is mined from large databases. Genetic algorithms can be used to buildeffective classifier systems, mining association rules and other such datamining problems. Their robust search technique has given them a central place in the field of data mining and machine learning [42]. GA can be viewed as an evolutionary process where at each generation, from a set of feasible solutions, individuals or solutions are selected such that individuals with higher fitness have greater probability of getting chosen. At each generation, these chosen individuals undergo crossover and mutation to produce a population of the next generation. This concept of survival of the fittest proposed by Darwin is the main cause for the robust performance of GAs. Crossover helps in the exchange of discovered knowledge in the form of genes between individuals and mutation helps in restoring lost or unexplored regions in search space.[ref book]Genetic algorithm (GA)[1], belonging to a class of randomized heuristic and adaptive search techniques based on the principal of natural selection, is an attractive tool to find near optimal solutions for optimization problems. Genetic algorithms [2][]attempt to incorporate ideas of natural evolution. In general, genetic learning starts as follows. An initial population is created consisting of randomly generated rules. Each rule can be represented by a string of bits. As a simple example, suppose that samples in a given training set are described by two Boolean attributes, A1 and A2, and that there are two classes, C1 and C2. The rule \IF A1 and not A2 THEN C2" can be encoded as the bit string \100", where the two leftmost bits represent attributes A1 and A2, respectively, and the rightmost bit represents the class. Similarly, the rule \if not A1 and not A2 then C1" can be encoded as \001". If an attribute has k values where k > 2, then k bits may be used to encode the attribute's values. Classes can be encoded in a similar fashion. Based on the notion of survival of the fittest, a new population is formed to consist of the fittest rules in the current population, as well as offspring of these rules. Typically, the fitness of a rule is assessed by its classification accuracy on a set of training samples. Off spring are created by applying genetic operators such as crossover and mutation. In crossover, substrings from pairs of rules are swapped to form new pairs of rules. In mutation, randomly selected bits in a rule's string are inverted. The process of generating new populations based on prior populations of rules continues until a population P evolves where each rule in P satisfies a prespecified fitness threshold. Genetic algorithms are easily parallelizable and have been used for classification as well as other optimization problems. In data mining, they may be used to evaluate the fitness of other algorithms.

1. Role of Genetic Algorithms

- Robust, parallel, adaptive search methods —suitable when the search space is large.
- GAs : Efficient, Adaptive and robust Search Processes, Producing near optimal solutions and have a large amount of Implicit Parallelism.

- GAs are *Appropriate* and *Natural Choice* for problems which need – Optimizing Computation.
- Requirements, and Robust, Fast and Close Approximate Solutions

2. GAs in Data Mining

Many tasks involved in analyzing/identifying a pattern need Appropriate Parameter Selection and Efficient Search in complex spaces to obtain Optimal Solutions

B. Need of Soft Computing in Data Mining

Relevance of FL, ANN, GAs Individually to Data Mining problems is established. The hybrid methods provide a more power tool for data mining incorporating representation, learning and optimization features in the data mining model Soft Computing based techniques provide useful tools for datamining, making use of their main features[43][44]: Commonsense knowledge may sometimes be captured in an natural way using fuzzy rules. ANN: Machinery for learning and curve fitting (Learns from examples) GAs are Appropriate and Natural Choice for problems which need – Optimizing Computation Requirements, and Robust, Fast and Close Approximate Solutions.

V. DM challenges in bioinformatics[45]:

1.Non-traditional Feature Selection

- When the number of attributes >> number of samples?
- Highly imbalanced.

2.Explainable and Accurate Data Mining Methods

- NN, SVM→ Rules ?

3.Transfer Learning

- Can knowledge learned from one set of samples help data mining on another sample?

4.Exploiting the network structure

5.Individual i.i.d type of classification vs social networks?

VI.Application of Data Mining in Bioinformatics

Applications of data mining to bioinformatics include gene finding, protein function domain detection, function motif detection, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing, and protein sub-cellular location prediction. For example, microarray technologies are used to predict a patient's outcome. On the basis of patients' genotypic microarray data, their survival time and risk of tumour metastasis or recurrence can be estimated. Machine learning can be used for peptide identification through mass spectroscopy. Correlation among fragment ions in a tandem mass spectrum is crucial in reducing stochastic mismatches for peptide identification by database searching. An efficient scoring algorithm that considers the correlative information in a tuneable and comprehensive manner is highly desirable. We discuss some application here:

1. Genome analysis

It entails the prediction of genes in uncharacterized genomic sequences. The objective is to be able to take a newly sequenced uncharacterized genome and break it up into introns, exons, repetitive DNA sequences, etc. and other elements. The various components of Genome Analysis are:

- **Gene Evaluation:** Given a DNA sequence, what part of it codes for a protein and what part of it is junk DNA.
- **Genome Classification:** Classify the junk DNA as intron, untranslated region, transposons, dead genes, regulatory elements etc.
- **Gene Prediction:** Predict the coding regions in a newly sequenced genome into the genes (coding) and the non-coding regions

2.Protein Structure Prediction

Considered to be the holy grail of modern biology, a solution to the protein folding problem has enormous potentially beneficial impact on society. Prediction of three dimensional structures of drug targets, design of biocatalysts and nano biomachines are a few of the multitude of foreseeable applications. At IIT Delhi, we have been developing a computational protocol for modeling and predicting protein structures at the atomic level.

3.Drug Design

Design of a novel drug is one of the biggest challenges faced by the pharmaceutical industry. The use of computers accelerate the process of drug design which is a time intensive process, and also reduces the cost of whole process. Computational methods are used in various forms of drug discovery like QSAR, virtual screening and structure-based drug designing methods. Among these, structure based drug design is gaining importance due to rapid growth in structural data (available in RCSB & Nucleic acid Data Bank). This structural data can be used in molecular modeling to design lead molecules based on the structural features of the active site.

V.APPLICATION OF SOFT COMPUTING TO BIOINFOMATICS:

A. FUZZY LOGIC IN BIOINFOMATICS:

There are many application areas in biomedical science and bioinformatics, where fuzzy logic techniques can be applied successfully. Some of the important uses of fuzzy logic are listed below:

- To increase the flexibility of protein motifs (Anbarasu et al., 1998; Taria et al., 2008).
- To study differences between polynucleotides (Torres and Nieto, 2003).
- To analyze experimental expression data (Tomida et al., 2002) using fuzzy adaptive resonance theory.
- To align sequences based on a fuzzy recast of a dynamic programming algorithm (Schlosshauer and Ohlsson, 2002).
- DNA sequencing using genetic fuzzy systems (Cord'on et al., 2004).
- To cluster genes from micro-array data (Fickett, 1996; Belacel et al., 2004).
- To predict proteins sub-cellular locations from their dipeptide composition (Huang and Li, 2004) using fuzzy k- nearest neighbors algorithm.
- To simulate complex traits influenced by genes with fuzzy-valued e.ects in pedigreed populations (Carleos et al., 2003).
- To attribute cluster membership values to genes (Demb'el'e and Kastner, 2003) applying a fuzzy partitioning method, fuzzy C-means.
- To map specific sequence patterns to putative functional classes since evolutionary comparison leads to e.cient functional characterization of hypothetical proteins (Heger and Holm, 2003). The authors used a fuzzy alignment model.
- To analyze gene expression data (Woolf and Wang, 2000).
- To unravel functional and ancestral relationships between proteins via fuzzy alignment methods (Blankenbecler et al., 2003), or using a generalized radial basis function neural network architecture that generates fuzzy classification rules (Wang et al., 2003).

GA in Bioinfomatics

GA has been applied to

- Gene regulatory network identification (Chen et al., 1999; Ando and Iba, 2001; Behera and Nanjundiah, 1997; Ando and Iba, 2000; Leping et al., 2007; Tominaga et al., 1999; Lewis, 1998),
- Construction of phylogenetic tree for studying evolutionary relationship (Lemmon and Milinkovitch, 2002; Katoh et al., 2001; Matsuda, 1996; Skourikhine, 2000).
- DNA structure prediction (Baldi and Baisnee, 2000; Anselmi et al., 2000; Landau and Lifshitz, 1970; Becker and Buydens, 1997; Parbhane et al., 2000).
- RNA structure prediction (Adrahams and Breg, 1990; Waterman, 1988; Zuker and Stiegler, 1981; Batenburg et al., 1995; Gultyaev et al., 1995; Wiese and Glen, 2003; Shapiro and Navetta, 1994; Shapiro and Wu, 1996; Shapiro et al., 2001).
- Protein structure prediction and clustering (Ghou and Fasmann, 1978; Riis and Krogh,1996; Qian and Sejnowski, 1988; Salamov and Solovyev, 1995; Salzberg and Cost, 1992;Garnier et al., 1996; Unger and Moult, 1993; Schulze-Kremer, 2000; Morris et al., 1998;Chen et al., 1998).

- Molecular design and molecular docking (Rosin et al., 1997; Yang and Kao, 2000; Oshiro et al., 1995; Clark and Westhead, 1996; Venkatasubramanian et al., 1994; Deaven and Ho,1995; Jones et al., 1995; Jones et al., 1999; McGarrah and Judson, 1993; Hou et al., 1999;Hatzigeorgiou and Reckzo, 2004) etc.

ARTIFICIAL NEURAL NETWORKS:

Recently there has been increased interest in applications of artificial neural networks (ANNs) in biomedical researches (Hosseini et al., 2007, Amiri et al., 2009, Rafienia et al., 2010).

ANNs are used in pharmaceutical and pharmacokinetic areas to model complex relationships and to predict the nonlinear relationship between causal factors and response variables. The distinct features of the ANN make this approach very useful in situations where the functional dependence between the inputs and outputs is not clear. The basic concepts of the multiobjective simultaneous optimization technique of drug formulations, by utilizing ANN, were reviewed by Takayama and colleagues (2003).

The applicability of the ANN in modeling and predicting drug release profiles was investigated to evaluate an experimental study in transdermal iontophoresis (Lim et al. 2003). An ANN-based system was reported to predict peaks and troughs of gentamicin serum concentrations based on a set of empirical data, and the results were comparable with those using nonlinear mixed effect modeling (Brier et al. 1995).

Furthermore, some researchers focused on developing pharmacokinetic models to predict plasma drug concentration based on ANNs and calculate the estimated concentrations of heparin for patients undergoing hemodialysis treatment (Valafar & Valafar, 1999). The following article are examples of different research endeavors that utilize sub-symbolic soft computing techniques. [47]

Backpropagation neural nets and variants (classifiers)

- Ma, Q., Wang, J. T. L. and C. H. Wu. (2000). Application of neural networks to biological data mining: A case study in DNA sequence classification. Proceedings of SEKE-2000, 23-30.
- C. F. Allex, J. W. Shavlik, and F. R. Blattner. (1999). Neural network input representations that produce accurate consensus sequences from DNA fragment assemblies. Bioinformatics 15(9), 723--728.
- Jason T. L. Wang, Qicheng Ma, Dennis Shasta, and Cathy H. Wu, New Techniques for Extracting Features from Protein Sequences.
- Qicheng Ma and Jason T. L. Wang. (1999). Biological Data Mining Using Bayesian Neural Networks: A Case Study. International Journal on Artificial Intelligence Tools, Vol. 8, No. 4, 433-451.
- Lucas Tabesh, Application Of A Neural Network To The Prediction Of Transmembrane Regions Of Membrane Proteins.
- Lawrence B. Crowell, Fokker-Planck Theory and Knot topology as a Method for Computing the Folding of Proteins . A rather simplistic model by a physicist for how proteins fold with lots of formulas from statistical mechanics. The model is not validated with any data whatsoever. Outline of a proposed mechanism of protein folding. It uses Feynman rachets to model the mechanics of polypeptide formation, Fokker Planck equations to explain ATP hydrolysis and the Ising spin lattice model to analyse thermodynamic stability. The mathematical modelling uses the Yang Baxter relation in knot topology. A neural-network based algorithm is described to predict the 3D structure of a protein sequence in time O(n^2 + nlogn), but not implemented or tested.
- C.H. Wu, Artificial neural Networks for molecular sequence analysis. Compu. Chem. 21:237-256, 1997.
- Wu, C. H. and McLarty, J. M. (2000). Neural Networks and Genome Informatics. Methods in Computational Biology and Biochemistry, Volume 1, Series Editor A. K. Konopka, Elsevier Science.
- Wu, C. H. (1997). Artificial neural networks for molecular sequence analysis. *Computers & Chemistry*, 21(4), 237 - 256.
- Wu, C. H. , H. L. Chen and S. Chen. (1997). Counter-propagation neural networks for molecular sequence classification: Supervised LVQ and dynamic

node allocation. *Applied Intelligence*, 7 (1), 27-38.

- Wu, C. H., S. Zhao, H. L. Chen, C. J. Lo and J. McLarty. (1996). Motif identification neural design for rapid and sensitive protein family search. *CABIOS*, 12 (2), 109-118.
- Wu, C. H.. (1996).Gene Classification Artificial Neural System. *Methods In Enzymolog*, 266, 71-88.
- Wu, C. H., M. Berry, S. Shivakumar and J. McLarty. (1995). Neural networks for full-scale protein sequence classification: Sequence encoding with singular value decomposition. *Machine Learning*, 21, 177-193.
- Wu, C. H. and S. Shivakumar. (1994). Back-propagation and counter-propagation neural networks for phylogenetic classification of ribosomal RNA sequences. *Nucleic Acids Research*, 22(20), 4291-4299.
- Wu, C. H., G. Whitson, J. McLarty, A. Ermongkonchai and T. Chang. (1992). Protein classification artificial neural system. *Protein Science*, 1, 667-677.
- J. T. L. Wang, Q. Ma, D. Shasha, and C. H. Wu. Application of neural networks to biological data mining: A case study in protein sequence classification. In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2000, pp. 305—309.

Kohonen maps / self-organizing feature maps (categorizers).

- J. Herrero and A. Valencia and J. Dopazo. (2001). A hierarchical unsupervised growing neural network for clustering gene expression patterns, Bioinformatics, 17:126--136.
- Gabriela Guimarães and Wolfgang Urfer, Self-Organizing Maps and its Applications in Sleep Apnea Research and Molecular Genetics.
- W. Wriggers and R. A. Milligan and K. Schulten and J. A. McCammon. (1998) Self-Organizing Neural Networks Bridge the Biomolecular Resolution Gap, Journal of Molecular Biology, vol. 284, 1247--1254.

Associative memory NNs

- Silvio Bicciato, Giuseppe Didone, Mario Pandin, and Carlo Di Bello, Analysis Of An Associative Memory Neural Network For Pattern Identification In Gene Expression Data, BIOKDD01: Workshop on Data Mining in Bioinformatics

NN enhancement techniques

- Nitesh Chawla Thomas, Bagging-Like Effects for Decision Trees and Neural Nets in Protein Secondary Structure Prediction

. In N. e. a. Kasabov, editor, Proceedings of ICONIP '97. Springer, 1998.

VII.Conclusion and challenges

Bioinformatics and data mining are developing as interdisciplinary science. Data mining approaches seem preferably suited for bioinformatics, since bioinformatics is data-rich but lacks a widespread theory of life's organization at the molecular level. However, data mining in bioinformatics is hampered by many facets of biological databases, including their size,number, assortment and the lack of a standard ontology to assist the querying of them as well as the heterogeneous data of the quality and provenance information they contain. Another problem is the range of levels the domains of expertise present amongst potential users, so it can be difficult for the database curators to provide access mechanism appropriate to all. The integration of biological databases is also a problem. [46]Data mining and bioinformatics are fast growing research area today. It is important to examine what are the important research issues in bioinformatics and develop new data mining methods for scalable and effective analysis. Soft computing techniques are closely used in the field of bioinformatics to solve hard problems. Not only individual soft computing methodologies but their hybridization also have proved there value and strength and in the field of bioinformatics. Knowledge and ability to use neural networks method add definite advantage to bioinformaticians to solve many types of problems in the field of bioinformatics.

References:

[1] Dataminig Concepts and Techniques- Jiawei Han, Micheline Kamber

[2] Principles of Soft Computing-S.N.Sivanandam and S.N.Deepa

[3] bio refin II

[4]K.R.Venugopal, K.G. Srinivasa and L.M. Patnaik soft computing for data Mining Application

[5] T. M. Mitchell, "Machine learning and data mining," *Commun. ACM*,vol. 42, no. 11, 1999.

[6] U. Fayyad, G. P. Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data," *Commun. ACM*, vol. 39, pp. 27–34, 1996.

[7]neural networks,Fuzzy Logic and Genetic Algorithm Synthesis and Applications-S.Rajasekaran,G.A.Vijayalakshmi Pai.

[8]"Bioinfomatics World" magazine:www.bioinfomaticsworld.info

[21][9]

[10] V. Ciesielski and G. Palstra, "Using a hybrid neural/expert system for database mining in market survey data," in Proc. SecondInternational Conference on Knowledge Discovery and Data Mining(KDD-96), (Portland, OR), p. 38, AAAI Press, Aug. 2-4, 1996.

[11] K. Xu, Z. Wang, and K. S. Leung, "Using a new type of non linear integral for multi-regression: an application of evolutionary algorithms in data mining," in Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, (San Diego, CA), pp. 2326-2331, October 1998.

[12] E. Noda, A. A. Freitas, and H. S. Lopes, "Discovering interesting prediction rules with a genetic algorithm," in
Proceedings of IEEE Congress on Evolutionary Computation CEC 99, (Wash- ington DC), pp. 1322-1329, July 1999.

[13] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in Proceedings of 1993 ACM SIGMOD International Conference on Management of Data, (Washington D.C.), pp. 207-216, May 1993

[14] Q. Wei and G. Chen, "Mining generalized association rules with fuzzy taxonomic structures," in Proceedings of NAFIPS 99, (New York), pp. 477-481, June 1999.

[15]W. H. Au and K. C. C. Chan, "An effective algorithm for discovering fuzzy rules in relational databases," in Proceedings of IEEE International Conference on Fuzzy Systems FUZZ IEEE 98, (Alaska), pp. 1314-1319, May 1998.

[16]C. Lopes, M. Pacheco, M. Vellasco, and E. Passos, "Ruleevolver: An evolutionary approach for data mining," in Proceedings of RSFDGrC'99, (Yamaguchi, Japan), pp. 458-462, November

[17]I. B. Turksen, "Fuzzy data mining and expert system development," in Proceedings of IEEE International Conference on
Systems, Man, and Cybernetics, (San Diego, CA), pp. 2057-2061, October 1998.

[18] S. Russell and W. Lodwick, "Fuzzy clustering in data mining for telco database marketing campaigns," in Proceedings of NAFIPS 99, (New York), pp. 720-726, June 1999.

[19] W. Pedrycz, "Conditional fuzzy c-means," Pattern Recognition Letters, vol. 17, pp. 625-632, 1996.

[20]D. Shalvi and N. De Claris, "Unsupervised neural network approach to medical data mining techniques," in Proceedings of IEEE International Joint Conference on Neural Networks, (Alaska), pp. 171-176, May 1998.

[21]H. Kiem and D. Phuc, "Using rough genetic and Kohonen's neural network for conceptual cluster discovery in data mining," in Proceings of RSFDGrC'99, (Yamaguchi, Japan), pp. 448-452, November 1999.

[22]T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, and A. Saarela, "Self organization of a massive document collection," IEEE Transactions on Neural Networks, vol. 11, pp. 574-585, 2000.

[23]J. Vesanto and E. Alhoniemi, "Clustering of the selforganizing map," IEEE Transactions on Neural Networks, vol. 11,
pp. 586-600, 2000.

[24]D. Alahakoon, S. K. Halgamuge, and B. Srinivasan, "Dynamic self organizing maps with controlled growth for
knowledge discovery," IEEE Transactions on Neural Networks, vol. 11, pp. 601-614, 2000.

[25]A. B. Tickle, R. Andrews, M. Golea, and J. Diederich, "The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks," IEEE Transactions on Neural Networks, vol. 9, pp. 1057-1068, 1998.

[26]H. J. Lu, R. Setiono, and H. Liu, "Effective data mining using neural networks," IEEE Transactions on Knowledge and Data Engineering, vol. 8, pp. 957-961, 1996.

[27] S. Mitra and Y. Hayashi, "Neuro-fuzzy rule generation: Survey in soft computing framework,"

IEEE Transactions on Neural Networks, vol. 11, pp. 748-768, 2000.

[28]T. Mollestad and A. Skowron, “A rough set framework for data mining of propositional default rules,” Lecture Notes in Computer Science, vol. 1079, pp. 448-457, 1996.

[29]X. Hu and N. Cercone, “Mining knowledge rules from databases: A rough set approach,” in Proceedings of the 12th International Conference on Data Engineering, (Washington), pp. 96-105, IEEE Computer Society, Feb. 1996.

[30] A. Skowron, “Extracting laws from decision tables - a rough set approach,” Computational Intelligence, vol. 11, pp. 371-388, 1995.

[31] N. Shan and W. Ziarko, “Data-based acquisition and incremental modification of classification rules,” Computational Intelligence, vol. 11, pp. 357-370, 1995.

[32]Y. Q. Zhang, M. D. Fraser, R. A. Gagliano, and A. Kandel, “Granular neural networks for numerical-linguistic data fusion and knowledge discovery,” IEEE Transactions on Neural Networks, vol. 11, pp. 658-667, 2000

[33] D. H. Lee and M. H. Kim, “Database summarization using fuzzy ISA hierarchies,” IEEE Transactions on Systems Man and Cybernetics. Part B-Cybernetics, vol. 27, pp. 68-78, 1997.

[34]R. R. Yager, “On linguistic summaries of data,” in Knowledge Discovery in Databases (W. Frawley and G. Piatetsky-Shapiro, eds.), pp. 347-363, Menlo Park, CA: AAAI/MIT Press, 1991.

[35]J. Kacprzyk and S. Zadrozny, “Data mining via linguistic summaries of data: an interactive approach,” in Proceedings of IIZUKA 98, (Fukuoka, Japan), pp. 668-671, October 1998.

[36] R. George and R. Srikanth, “Data summarization using genetic algorithms and fuzzy logic,” in Genetic Algorithms and Soft Computing (F. Herrera and J. L. Verdegay, eds.), pp. 599-611, Heidelberg: Springer-Verlag, 1996.

[37] D. A. Chiang, L. R. Chow, and Y. F. Wang, “Mining time series data by a fuzzy linguistic summary system,” Fuzzy Sets and Systems, vol. 112, pp. 419-432, 2000.

[38]R. S. T. Lee and J. N. K. Liu, “Tropical cyclone identification and tracking system using integrated neural oscillatory leastic graph matching and hybrid RBF network track mining techniques,” IEEE Transactions on Neural Networks, vol. 11, pp. 680-689, 2000.

[39]. Jang, J.S.R., Sun, C.T., Mizutani, E.: Neuro-Fuzzy and Soft Computing. Pearson Education, London (2004)

[40]. Sushmita Mitra , Sankar K. Pal, Pabitra Mitra ,”Data Mining in Soft Computing FrameworkA Survey”, IEEE Transactions on Neural Networks, Vol. 13, No. 1, January 2002

[41].Mitchell, T.M.: Machine Learning. McGraw Hill International Editions, New York (1997)

[42]. Holland, J.H.: Adaptation in Natural and Artificial Systems. University of Michigan Press (2004)

[43] http://sci2s.ugr.es

[44] http://decsai.ugr.es/~herrera

[45] http://www.cse.ust.hk/~qyang/

[46]ISSN : 0976-5166 117 Khalid Raza / Indian Journal of Computer Science and Engineering Vol 1 No 2, 114-118

[47http://citeseer.nj.nec.com].