Pattern recognition techniques for the emerging fields in bioinformatics

K.M.K.P Kumarasinghe

Reg no: 2011cs176 Index no : 11001763
University of Colombo School of Computing
kkaveenp@gmail.com

SCS3017 : Literature Survey
Reference style: IEEE standard
Word count : 5569 words
Tools used: Word2013/Latex/Mendeley

Supervisor : Dr.D.A.S. Atukorale
Dec-12-2014

Abstract

The emerging field of bioinformatics has recently created much interest in the computer science and engineering communities. With the wealth of sequence data in many public online databases and the huge amount of data generated from the Human Genome Project, computer analysis has become indispensable. This calls for novel algorithms and opens up new areas of applications for many pattern recognition techniques. This paper describe current methods, tools, algorithms and show how pattern recognition techniques could be useful in these areas. This paper does not discuss much technical details about those techniques but overview of existing techniques and algorithms in emerging field of bioinformatics. It is my hope that this review article could demonstrate how the pattern recognition community could have an impact on the fascinating and challenging area of genomic research.

Contents

List of Figures

List of Abbreviations and Acronyms

GNB Gaussian Naive Bayes

RSA Relative solvent accessibility

PSSM Position scoring specific matrix

RNA Ribonucleic acid

mRNA messenger RNA

PSP Protein structure prediction

SVM Support vector machine

GSVM Granula support vector machine

k-NN k-nearest neighbour algorithm

NB Naive Bayesian classifier

DT Decision trees

PCA Principle component analysis

ACNN Associative clustering neural network

1 Introduction

1.1 Motivation

The significance of the bio-informatics is quite clear during the recent past, although it is a new research area. The treatments for the new diseases existed in the modern society; this field was used by the people as it is quite convenient. The main research area of this field is included the analysis of the Protein and DNA atoms, and the diagnosis of new diseases. The numerous biological data are diagnosed earlier and to do matching successfully the computational method was essential. For this, the pattern recognition technology and the techniques relevant to this were the most successful remedy. The Pattern Recognition is used worldwide as a successful and essential technology in the Bio-Informatics field.

1.2 Bioinformatics and Pattern recognition

Bioinformatics is the conceptualizing biology in terms of molecules and applying informatics techniques to understand and organize the information associated with the molecules, on a large scale[1]. It involves the development and advancement of algorithms using techniques including pattern recognition, machine learning, applied mathematics, statistics, informatics, and biology to solve biological problems usually on the molecular level. The researchers who are doing reaches in this field always facing major research problems including sequence alignment and analysis, gene finding, genome annotation, protein structure alignment and prediction, classification of protein clustering and dimensionality reduction of microarray expression data, protein-protein docking or interaction, modeling of evolution and so forth[2].

Bioinformatics also can be described as development and application of computational method to make biological discoveries[1]. The ultimate attempt of this filed is to develop new insights in to the science of life as well as creating a global perspective, from which the unifying principles of biology can be derived.

As classification, clustering, feature selection needed in this field pattern recognition techniques and tools widely using in this field. Pattern recognition technique and algorithms successfully applied in to emerging fields in bio informatics. Most of the pattern recognition techniques and tools are using in classification of protein, classification of DNA and RNA, novel diseases identification and bio image analyzing.

Pattern recognition methods are built on the assumption that some underlying characteristics of protein sequence or of a protein structure, can be used to identify similar traits in related proteins. In other words, if part of a sequence or structure is preserved or conserved this characteristic may be used to diagnose new family members. If such conserved traits are distilled from known protein families and stored in databases, then newly sequenced proteins may be rapidly analyze to determine whether they contain these previously recognized family characteristics. Searches of sequence pattern databases, and of fold template databases, are now routinely used to diagnose family relationships and hence to infer structure and functions of newly determined sequences[3]. This review article discus about the pattern recognition techniques and tools which are using in the emerging fields of bioinformatics. I‘ve recognized four main fields as the emerging fields, including DNA, PROTEIN, Novel disease identification and Gene expression. Under each topic this article discuss the novel techniques and tools which are existing now.

2 Overview of pattern recognition

Pattern recognition is an activity that human being normally excel in. The task of pattern recognition is encountered in a wide range of human activity. In a broader perspective, the term could cover any context in which some decision or forecast is made on the basis of currently available information. Mathematically, the problem of pattern recognition deals with the construction of a procedure to be applied to a set of inputs; the procedure assigns each new input to one of a set of classes on the basis of observed features. The construction of such a procedure on an input data set is defined as pattern recognition[4].

2.1 Pattern recognition methods

2.1.1 Statistical pattern recognition(SPR)

Statistical decision and estimation theories have been commonly used in pattern recognition for a long time. It is a classical method and it is based on feature vector distributing which getting from probability and statistical model. This model is defining by class conditional probability density functions. In statistical pattern recognition deals with features only without consider the relations between features[4].

2.1.2 Data clustering

Data clustering is an unsupervised method. In general, the method of data clustering can be partitioned two classes, one is hierarchical clustering, and the other is partition clustering.

2.1.3 fuzzy logic

The thinking process of human being is often fuzzy and uncertain, and the languages of human are often fuzzy also. And in reality, we can't always give complete answers or classification, so theory of fuzzy sets come into being. Fuzzy sets can describe the extension and intension of a concept effectively.

2.1.4 Neural Network

It is a data clustering method based on distance measurement; also this method is model-irrespective. The neural approach applies biological concept to machines to recognize pattern. The outcome of this effort is the invention of artificial neural networks which is set up by the elicitation of the physiology knowledge of human brain. Neural networks is composed of a series of different, associate unit. In addition, genetic algorithms applied in neural networks is a statistical optimized algorithms proposed by Holland (1975).

2.1.5 Structural pattern recognition.

Structural pattern recognition is not based on a firm theory which relies on segmentation and feature extraction. Structural pattern recognition emphases on the description of the structure, namely explain how some simple sub patterns compose one pattern. There are two main method in structural pattern recognition, structure matching and syntax analysis. The basic of structure matching is some special technique of mathematics based on sub-pattern.

2.1.6 Support vector machine

SVM is a relatively new thing with simple structure; it has been researched widely since it was discovered 1990's. SVM base on the statistical theory, and the method of SVM is an effective tools that can solve the problems of pattern recognition and function estimation, especially can solve classification and recognition such as face detection, verification and recognition, object detection and recognition, speech recognition etc.

2.1.7 A novel method and system of pattern recognition using data encoded as Fourier series and Fourier space

This novel method anticipate the signal processing of an ensemble of neurons as a unit and intends to simulate aspects of brain which bring capabilities like pattern recognition an reasoning that have not been produced with past approaches as neural networks[5].

2.2 Pattern recognition system

A pattern recognition system can be regarded as a process that allows it to cope with real and noisy data. Whether the decision made by the system right or not mainly depending on the decision make by the human expert.

A pattern recognition system based on any pattern recognition method mainly includes three mutual associate and differentiated process. One is data building the other two are pattern recognition analysis and pattern classification. Data building convert original information into vector which can be dealt with by computer. Pattern analysis's task is to process the data (vector) such as feature selection, feature extraction, data dimension compress and so on. The aim of pattern classification to discipline the computer in order to accomplish the classification.In its border senses pattern recognition is the heart of many scientific inquiries, including ourselves and the real world around us. By the rest of this paper will discuss the pattern recognition techniques and algorithms which are currently using in the emerging fields of bioinformatics.

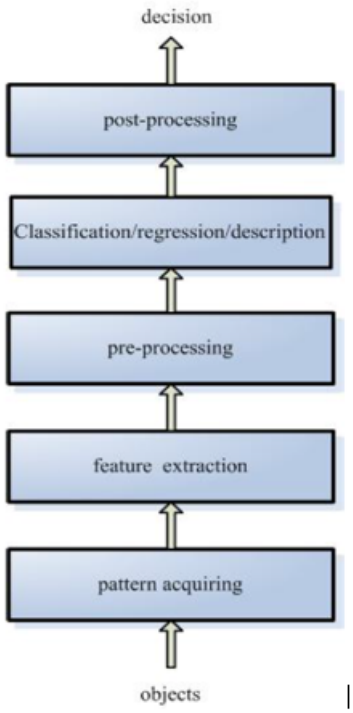

Figure 1: The composition of Pattern Recognition system

3 Pattern recognition techniques in DNA related researches

3.1 Introduction

Deoxyribonucleic acid (DNA) is a nucleic acid that contains the genetic instructions for the development and function of living things. DNA sequencing and research have progressed over the years, ultimately leading into the field of bioinformatics. The two major research areas related to DNA in bioinformatics are DNA sequence classification and sequence based prediction of DNA-binding proteins. In pattern recognition perspective sequence based prediction of DNA binding proteins has much more specialty than DNA sequence classification. And also newly introduced pattern recognition techniques are widely using under this topic.

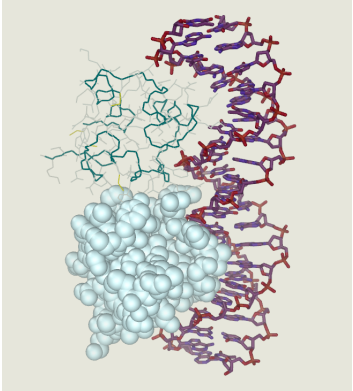

Figure 2: Co protein complex with DNA

3.2 Sequence based prediction of DNA-Binding Proteins

DNA-binding protein play key role in a variety of molecular functions, including recognizing specific nucleotide sequence. Maintenances of cellular DNA, transcriptional and transcriptional regulation, and DNA damage repair[6].Currently both computational and experimental techniques have been developed to identify the protein-DNA interactions. The experimental techniques such as filter binding assays, chip-chip, genetic analysis and X-ray crystallography can provide a detailed picture about the binding. However they are both time consuming and expensive .

Thus it is highly desired to develop automated computational methods for identifying the DNA-binding proteins from the extremely fast increased amount of newly discovered proteins[7].

So far number of predictors of DNA-binding proteins have been proposed. These methods can be divided into two categories structure based modeling and sequence based prediction. The pit fall of various structure based methods for predicting DNA-binding function is that they are all limited to a relatively small number of proteins for which high-resolution three-dimensional structures are available. In contrast, sequence based methods have the main advantage with no need for known structures and thus can be applied to large-scale datasets and genomic targets[8]

3.2.1 Techniques and tools

As mentioned above many Pattern recognition techniques are proposed to apply to sequence based prediction of DNA binding proteins. For instance, Szilgayi and Skolnick used logistic regression to predict the DNA-binding proteins from the amino acid composition[9]. Kumar et al. utilized support vector machine and coded the features from evolutionary profiles for the prediction of DNA-binding proteins. Another group, group Kumar et al. proposed DNA-Prot method for the classification of the DNA-binding proteins using random forest. The latest work by Zou et al. provide a comprehensive feature analysis using support vector machine for the prediction of DNA-binding proteins[8].

As a summery, sequence based prediction methods for DNA-binding proteins have been investigated with several pattern recognition methods such as logistic regression[8], random forest and support vector machine and Gaussian Naïve Bayes. When considering these techniques Random forest and Gaussian Naïve Bayes plays major role in sequence based prediction.

Random forest has been widely used for pattern recognition in bioinformatics. It can provide not only the high prediction performance but also information on variable importance for classification task. The algorithm of random forest is based on the ensemble of a large number of decision trees. Where each trees gives a classification and the forest choose the final classification having the most votes (over all the trees in the forest). In the most commonly used type of random forests, split selection is performed based on the so-called decrease of Gini impurity. In this study, the random forest is used to rank the features using Gini importance [8]that is implemented with the machine learning platform.

Naïve Bayes is a set of supervised learning algorithms that applies Bayes' theorem with the "naive" assumption of independence between every pair of features[8]. A Naïve Bayes classifier calculates the probability that a given instance belongs to a certain class. Given an instance X, described by its feature vector and a class target y, Bayes theorem allows us to express the conditional probability as a product of simpler probabilities using the naïve independence assumption:

$$P(y|X) = \frac{P(y)P(X|y)}{P(X)}$$
$$= \frac{P(y)\,\Pi_{i=1}^{n} P(x_i|y)}{P(X)|}$$

Figure 3: Simpler probabilities using the Naïve independence assumption

Since P(X) is a constant for a given instance, the following rule is used to classify the DNA sample

$$\hat{y} = \arg\max_{y} P(y) \prod_{i=1}^{n} P(x_i|y)$$

Figure 4: DNA classification rule

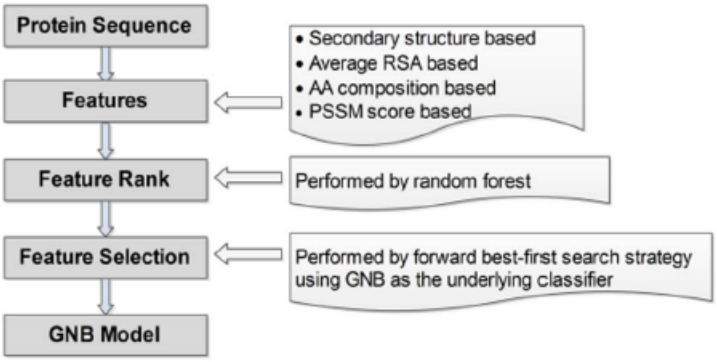

Figure 5: Flow chart of using Random Forest and Gaussian Naïve Bayes algorithm

4 Pattern recognition techniques in Protein related researches

4.1 Introduction

Proteins are the molecular devices, in the nanometer scale, where biological function is exerted[10]. They are the building blocks of all cell in our bodies and in all living creatures of all kingdoms. Although the information necessary for life to go on is encoded by the DNA molecule, the dynamic process of life maintenance, replication, defense and reproduction are carried out by proteins. As the protein plays major role in human being it has been invoked big research areas in bioinformatics. Most of the researches and scientists are interested in protein classification, protein structure prediction and sequence comparison. Among these fields pattern recognition techniques are success fully used in protein classification and protein structure prediction (PSP). Rest of this chapter will discuss about the pattern recognition techniques are using in protein structure prediction.

4.2 Protein Structure Prediction

Protein structure prediction (PSP) is one of the most important goals pursued in bioinformatics and theoretical chemistry. Its aim is prediction of the three dimensional structure of proteins from their amino acid sequence, sometimes including additional relevant information such as the structure prediction of related proteins. In other words, it deals with

the prediction of a protein's tertiary structure from is primary structure. Protein structure prediction is of high importance in medicine and bio technology.

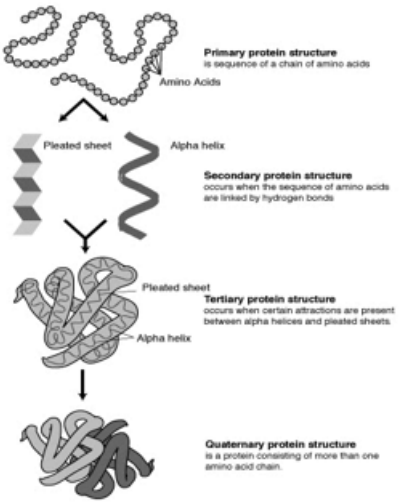

Figure 6: Protein structure prediction. Methodology

4.2.1 Techniques and tools

There have been many successful research projects focusing on this problem[11]. As an example Tang et al. address a problem if predicting protein homology between given two proteins. They propose a leaning method that combines the idea of association rule with their previous method called Granular Support Vector Machine (GSVM)[11], which systematically combines a SVM with granular computing. The method, called GSVM-AR, uses association rules with high enough confidence and significant support to find suitable granules to build a GSVM with good performance. The authors compared their method with SVM by KDDCUP04[12] protein homology prediction data. From the experimental results, GSVM-AR showed significant improvement compared to a single SVM.

The interface between combinatorial optimization and fuzzy sets-based methodologies is the subject of a very active and increasing research. In this context, Balanco et al[13] describe a fuzzy adaptive neighborhood search (FANS) optimization heuristic that uses a fuzzy valuation to qualify solutions and adapts its behavior as a function of the search state. FANS may also be regarded as a local search framework. The authors show an application

of this fuzzy sets based heuristic to the protein structure prediction problem in two aspects: 1) To analyze how the codification of the application of the solutions affects the result and 2) To confirm that FAN is able to obtain as good result as a genetic algorithm

Both result shed some light on the application of heuristics to the protein structure prediction problem and show the benefits and power of combination basic fuzzy sets ideas with heuristic techniques.

Ron Roger [14] reviewed a general frame work of genetic algorithms can be used for structure prediction problem. Using this frame work, significant studies that were published in recent years are discussed and compared. Application of genetic algorithms to the related question of protein alignments are also mentioned. The rationale of why genetic algorithms are suitable for protein structure prediction is presented, and future improvements that are still needed are discussed.

5 Pattern recognition in novel diseases identification

5.1 Introduction

Prediction of novel diseases is an important issue in biomedical research. At early days, annotation based methods were proposed for this problem. In next stage, with high through put technologies, data of interaction between genes/protein has grown quickly and covered almost genome and proteome, and therefore network-based methods for the issue is becoming prominent. Beside those two methods, the prediction problem can be also approached using pattern recognition because it can be formulated as a classification task. To date, a number of pattern recognition techniques and various machine learning methods has been success fully using to solve this issue [15].

5.2 Disease gene prediction

Disease gene prediction, the task of identifying the most plausible candidate disease genes, is an important issue in biomedical research, and a variety of approaches have been proposed.

5.2.1 Tools and techniques

Pattern recognition techniques has been successfully applied to solve various impotent biomedical problems they have been applied to identifying disease associated genes[13][14]. First the problem is formulated as a classification problem, where the task is to learn classifier from training data. Then the learned classifier is used to predict whether or not a test/candidate gene is a disease gene. Figure 5 shows a common schema of classification based approaches for disease gene prediction. Training data are usually known disease gene/proteins, however, some studies have also used unknown genes in the training task. These gene and unknown genes are annotated by-omits data.

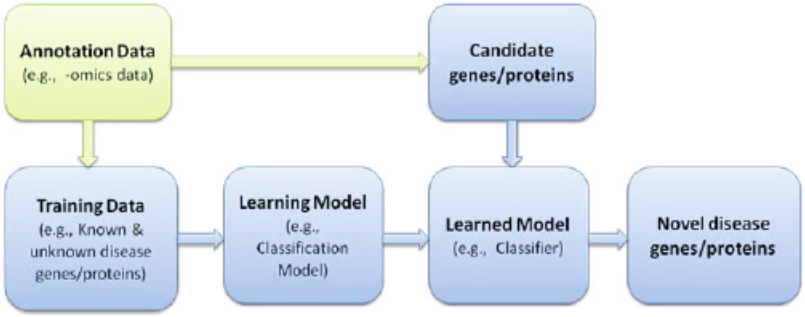

Figure 7: Common schema of classification- based approaches

To date, the binary-class classification techniques have largely been applied to the problems. The early applications of such techniques were of decision trees (DT)[15][16] using distinctive sequence features of known disease protein compared to all human proteins. With growth of interaction data between proteins k-nearest neighbor algorithm (k-NN), an instance based classifier was introduced and it was based on topological properties of protein on a human protein interaction network. Naïve Bayesian classifier (NB) was also used to identify human disease genes by integrating multiple types of genomic, phenotypic and interatomic data. In particular, a NB classifier was built based on eight different genomic dataset to identify human mitochondrial diseases. Based on both interaction and sequence data of protein, support vector machine was also used for the problem[17] and showed that SVM's performance was better than the k-NN. SVMs were subsequently used in a number of studies for disease gene prediction. Moreover unlike the methods mentioned above where the classifier was trained on all disease ontology term[18].Likewise the SVM classifier was used to identify genes associated to a specific disease. Furthermore an artificial neural network (ANN) was proposed to identify novel disease genes for four complex diseases.

5.3 Pattern recognition system for the diagnosis of Gonorrhea Disease

Sexually transmitted diseases (STDs) share common symptoms and can be classified as confusable disease, as such become difficult for physicians to correctly diagnose them. This PR system has provided an efficient answer for this problem.

5.3.1 Tools and techniques

The architecture of the pattern recognition system (PRS) for diagnosis of gonorrhea disease is shown in figure 6. This system consist of three major components; Knowledge base, Inference engine and Pattern classifier[19]

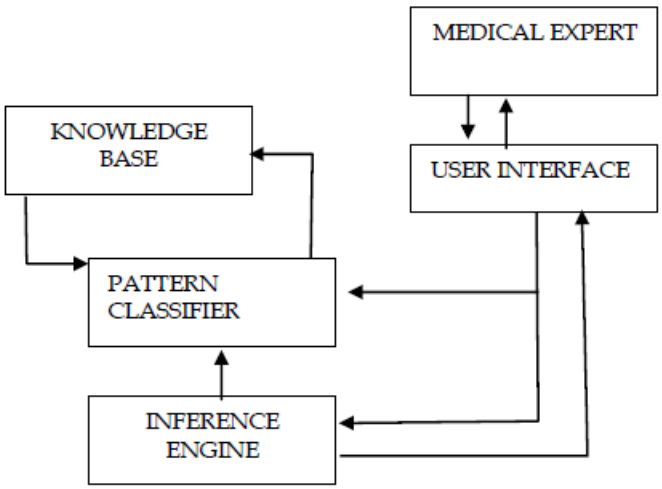

Figure 8: Architecture of PRS for diagnosis of gonorrhea

5.3.2 Knowledge base

Keeps track of relevant knowledge required for the diagnosis of gonorrhea. The user through the user interface supplies fact and information to the expert system or receives expert advice from the system. The knowledge base contains knowledge about the problem domain and database as its component.

5.3.3 Inference engine

The process of drawing conclusion from existing details called inference. The pattern recognition system inference uses the knowledge in the knowledge base to draw conclusions and decide whether the patient is infected with gonococcus bacterium or not. The system applies a probabilities output such that the probabilistic pattern recognition algorithm is effectively incorporated into a larger machine learning tasks, in a way that partially or

completely eliminate the problems or error propagation in the diagnostic process of sexual transmitted disease. The pattern recognition system algorithm classifies symptoms into: Number of classes (c=2), Features vector dimension (d=7), Classification coefficient (m=2). The PRS algorithm for the diagnosis of gonorrhea diseases is designed using the formula: Randomly initialized matrix[19] u = uij Where I = 1, 2..., N and j = 1, 2...M

$$Cj = \frac{\sum_{i=1}^{N} U^2 ijxi}{\sum_{i-1}^{N} U^2 ij}$$

$$uij = \frac{1}{\sum_{i-1}^{N} ((//xi - ci)^2/(//xi - cx//))}$$

Figure 9: Randomly initialized matrix

5.3.4 Pattern Classifier

Pattern recognition has to do with the assignment of some sort of output value (or label) to the series of input value (for instance), according to some specific algorithm. In here classification algorithm use to provide some reasonable answer for all possible inputs and to do "fuzzy" matching of inputs. The pattern matching algorithms look for exact matches in the input with preexisting patterns. In here features are categorize consisting of one of a set of unordered items such as a gender of "Male" or "Female", or a blood type "A","B","AB","O", ordinal (consisting of set of ordered items, e.g.: small, large, medium or small), integer –valued (e.g. the number of occurrences of a particular word) or real valued (e.g. a measurement of blood pressure). The recognizer algorithm is probabilistic nature, it produces a probabilistic output of the instance as described by the labels in (4) and (5). Figure 7 shows the endpoint of each membership function

$$P(\text{label} / x, \theta) = f(x, j\ \theta) \qquad (4)$$

Such that

$$P(\text{label}/x) = \int p(\text{label}/x, \theta)\ p(\theta/D) f\theta \qquad (5)$$

Figure 10: The recognizer algorithm

With the rate of sexual transmitted diseases in our generation today, medical doctors find it difficult to handle the diagnosis of a given class of STD and as such patients are being

	A	B	C
Class 1 (no testis pain)		36.45	36.62
Class 2 (slight testis pain)	27.21	38.20	30.12
Class 3 (high testis pain)	38.22	39.62	31.01
Class 4 (very high testis pain	40.26	42.22	

Figure 11: End point of each membership function

forwarded to the laboratory. Sometimes, the patients may be in a window state as such can prove such diagnosis wrong. This pattern recognition system for the diagnosis of gonorrhea with better performance, reliability and increase efficiency and availability.

6 Pattern recognition in Gene expression

6.1 Introduction

Gene expression refers to through which the coded information of a gene is converted into structures operating in the cell. It provides the physical evidence that a gene has been turned on or activated. Expressed genes include those that are transcribed into mRNA and then translated into protein and those that are transcribed into RNA but not translated into protein (e.g. transfer and ribosomal RNAs)[23] [24].The expression levels of thousands of genes can be measured at the same time using the modern microarray technology[20].

By this chapter this article will provide a substantial review of the state of the art research, which focuses on the application of computational intelligence to different bioinformatics related Gene expression problems.

6.1.1 Tools and techniques

In the field of pattern recognition clustering refers to the process of partitioning a dataset in to a finite number of groups according to some similarity measure. Currently it has become a widely used process in microarray engineering for understanding the functional relationship between group's genes. Clustering was used for example to understand the

functional differences n cultured primary hepatocytes relative to the intact liver. In another study, clustering techniques were used on gene expression data for tumor and normal colon tissue probed by oligonucleotide arrays[21].

A number of clustering algorithms including hierarchical clustering. Principle component analysis (PCA), genetic algorithms and artificial neural networks have been used to cluster gene expression data. However in 2002 Yuhui et al. proposed a new approach to analysis of gene expression data using Associative Clustering Neural Network (ACNN). ACNN dynamically evaluates similarity between any two gene samples through the interaction of a group of gene samples. It exhibits more robust performance than the methods with similarities evaluated by direct distances which has been tested on the leukemia data set. The experimental result demonstrate that ACNN superior in dealing with high dimensional data (7129 genes).The performance can be further enhanced when some useful feature selection methodologies are incorporated. The study has shown ACNN can achieve 98.61(percent) accuracy on clustering the Leukemias data set with correlation analysis[21].

7 Conclusion

Pattern recognition has increasingly gained attention in bioinformatics research and computational biology. With availability of different types of PR algorithms, it has become common for researchers to apply the off-shelf systems to classify and mine their databases. At present with various PR methods available in the literature, scientists are facing difficulties in choosing the best method that could be applied to specific data set. Researches need tools which present the data in a comprehensible fashion, annotated with context, estimates of accuracy and explanation. The terms bioinformatics and computational biology mean about the same. Recently however the USA national institute of health (NIH) came up with slightly different definitions.

Bioinformatics: Research, development or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data including those to acquire, store, organize, archive, analyze or visualize such data.

Computational Biology: The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral and social system.

7.1 Challenge

The problem of cancer classification is a major challenge face by scientist when designing PR algorithms. It can divide into two related but separate challenges i) Class prediction ii) Class discovery

Class prediction refers the assignment of sample to one of several previously define classes. Class discovery refers to defining a previously unrecognized tumor subtypes in expression data. Both of these tasks are challenging and require computational assistance. Class prediction via cluster analysis is typically used to infer the function of novel genes by grouping them with genes of well-known functionality in gene expression profiling. Genes that show similar activity patterns are often related functionally and are controlled by the same mech-

anisms or regulation. A major obstacle to the eventual utility of microarrays is the lack of efficient methods for cataloging the data into expressed groups. A new way of processing numeric data with large number of attributes versus low number of objects turns out to be well-suited to the gene expression data. Furthermore tumors are not identical even when they occur in the same organ and patients may need different treatment depending on their particular subtype of cancer. Identification of tumor subgroups is therefore important for diagnosis and design of medical treatment.

Most medical classification systems for tumors are currently based on clinical observations and the microscopically appearance of the tumors. These observations are not informative with regard to the molecular characteristic of the cancer[21]. The genes, whose expression levels are associated with the tumor subtypes, are largely unknown. A better understanding of the cancer could be achieved if these genes were identified. Furthermore, the disease may manifest itself earlier on the molecular level than on a clinical level. Hence, gene expression data from microarrays may enable prediction of tumor subtype and outcome at an earlier stage than clinical examination. Thus microarray analysis may allow earlier detection and treatment of the disease, which again may increase the survival rate[21].

Another challenge is to combine gene expression research with noninvasive imaging techniques. To address the challenges of relating gene expression to imaging, the researches followed a three step methodology and created an association map between imaging features on three-phase contrast enhanced CT scans and gene expression patterns of 28 human hepatocellular carcinomas (HCC). First, the researchers defined and quantified 138 units of distinctiveness named traits present in one or more HCCs. Second, the module networks algorithm was implemented. The algorithm systematically search for associations between expression levels of 6,732 well-measured genes determined by microarray analysis and combinations of imaging traits. Third, the statistical significance of the association map was validated by comparison with permuted data sets, and by testing the prediction of the association map in an independent set of tumors[21].

7.2 Future Direction

Most of the universities, companies and organizations have identified these current issues and already spent resources for the improvement of this research field. As an example Wyeth Company (a global leader in pharmaceuticals, consumer health care products, and animal health care products) is investing almost $86 million for developing new reliable PR related techniques and algorithms for above calcification problems[21]. And researchers are also need to pay attention to investigate reliable techniques for identifying disease blood cells which are caused to leukemia.

Cancer is not the only one disease which needs big attention. But when considering existing pattern recognition algorithms they can be easily apply to identify cancer cells at the primary age of the cancer. But the biological researchers cannot use pattern recognition techniques as it is in their day to day life. So there should be simple tools which are using this great technology. And it should easy to use for the biological researches.

The main purpose of this paper was to present the existing pattern recognition techniques and algorithms which are using in emerging fields in bioinformatics. And to inspire further research and development on new applications and new concept in new trend-setting directions in pattern recognition.

References

[1] D. Luscombe, "What is bioinformatics?. an introduction and overview," *Department of Molecular Biographyics and iochemistry Yale University, New Haven, USA.*

[2] S. P. P. Maji, "Scalable pattern recognition algorithms," *Applications in Computational Biology and Bioinformatics.*

[3] G. T. K. Attwood, "Introduction to bioinformatics," 2014.

[4] J. Lui, "Pattern recogniton: An overveiw," *IJCSNS International Journal of compute-science and network security*, vol. 6, pp. 5–100, 2006.

[5] R. Mills, "Novel methods and system for pattern recognition and processing using data encoeded fourier seris and fourier space," 2001.

[6] L. L. RE, "Boosting the prediction and understanding of dna binding domain from sequence," 2010.

[7] F. L. WZ, "Crystal structure of the hypertermophic arheal dna-binding protein," 2001.

[8] W. L. W, "Sequence based prediction of dna-binding proteins based on hybrid feature selection using," 2014.

[9] S. S. A, "Efficient prediction of nucleic acid inding function from low resolution protein structure," 2006.

[10] A.Lesk, "Introduction to potein architecture," 2001.

[11] J. Tang, "Granular support vector machine with association rules mining for protein homology prediction," *Artifical intelligence in medicine*, 2005.

[12] vailable At Available at `http://www.sigkdd.org/kdd-cup-2004-particle-physics-plus-protein-homology-prediction"`, YEAR = 2014,.

[13] D.R.D.Rider, "Pattern recognition in bioinformatics," *Briefing bioinformatics*, 2013.

[14] L.-B. N, "Genome wide identification of gene likly to be involved in human genatic disease," 2011.

[15] O. L.-B. N, "Genome wide identification of genelikely to be involved in human genatic diseases," *Nucleic acid research*, 2004.

[16] A. e al, "Speeding dieses gene discovery by sequence base candidate prioritzation," *BMC bioinformatics*, 2005.

[17] "Human disease-gene classification with integrative sequence-based," *IEEE International Conference on Bioinformatics and Biomedicine*, p. 216, 2007.

[18] R. P, "An integrated approach to inferring gene-disease association," 2008.

[19] J. G. G. O. U. U. U. U. A, "Design of pattern recognition sytem for the diagnosis of gonorrhea disease," *INTERNATIONAL JOURNAL OF SCIENTIFIC,TECHNOLOGY RESEARCH*, vol. 1, p. 6, 2012.

[20] J. Quackenbush, "Computational analysis of microarray data. national review of genatics," vol. 2, p. 418, 2001.

[21] M. G. M. T. G. S. A.-E. Hassanien, "Computational intelligence in solving bioinformatics problems reviews perspectives and challenges," 2008.