Machine Learning in Bioinformatics: A Novel Approach for DNA Sequencing

Pooja Dixit
Department of Information Technology
Shri S'ad Vidya Mandal Institute of Technology
Bharuch, India
dixit.pooja44@gmail.com

Ghanshyam I. Prajapati
Department of Computer Science and Information Technology
Shri S'ad Vidya Mandal Institute of Technology
Bharuch, India
Giprajapati612@gmail.com

Abstract— **Machine learning is the adaptive process that makes computers improve from experience, by example, and by analogy. So It is a discipline of methodologies that provides, in one form or another, intelligent information processing capabilities for handling real life. Bioinformatics is one of the application of Machine Learning. Bioinformatics is the interdisciplinary science of interpreting biological data using information technology and computer science. Machine learning (ML) focuses on automatic learning from data set. Machine learning includes the learning speed, the guarantee of convergence, and how the data can be learned incrementally. We usually refer to methods like Artificial Neural Networks (ANNs), Genetic algorithms (GAs), and Fuzzy systems along with hybrid methods including a combination of some of these methods. One of the major problems is to classify the normal genes and the invalid genes which are infected by some kind of diseases. In genomic research, classifying DNA sequences into existing categories is used to learn the functions of a new protein. So, it is important to identify those genes and classify them. In order to identify the infected genes and the normal genes with the use of classification methods here we use the machine learning techniques. This paper gives a review on the mechanisms of gene sequence classification using Machine Learning techniques, which includes a brief detail on bioinformatics, literature survey and key issues in DNA Sequencing using Machine Learning.**

Keywords— Bioinformatics; DNA Sequencing; Classification; Machine Learning Techniques

I. INTRODUCTION

The origin of Bioinformatics can be from the Mendel's discovery of genetic inheritance in 1865.Since 1953 big revolution achievements took place by James Watson and Francis Crick as they determined the structure of DNA [1]. Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data.

As an interdisciplinary field of science, Bioinformatics combines computer science, statistics, mathematics and engineering to study and process biological data. Uses of bioinformatics include the identification of candidate gens and nucleotides. Such identification is made up with the aim of better understanding the genetic basis of disease. Gene sequence classification plays an important role it also tries to understand the principles within nucleic acid and protein sequences.

In the nucleus of each cell, the DNA molecule is packaged into thread-like structures called chromosomes. Each chromosome is made up of DNA tightly coiled many times around proteins called histones that support its structure. Deoxyribo Nucleic Acids (DNA) which carry the genetic information. It is nucleic acid alongside proteins and carbohydrates. It has three molecules components: 5 carbon sugar, phosphate and nitrogenous base. It contains nucleotides Each nucleotide is composed of a nitrogen containing nucleobase- Adenine(A), Thymine(T), Cytosine(C) or Guanine(G).the sequence of these four nucleobase encodes the biological information. Adenine & Guanine called purin base (double ringed) And Thymine & Cytosine called pyrimedine (single ringed) [4]. Bioinformatics is the application of computer science in the field of biology to get information from the biological data. In the recent years, rapid development in genomics and proteomics has generated large amount of data [14].

Drawing conclusion from these data it requires sophiscated computational analysis that is done by bioinformatics. The major area of research including into bioinformatics are protein sequence analysis, analysis of protein expression, classification of gene sequences and prediction etc. a particular active area of research in bioinformatics is the application and development of machine learning techniques to solve the biological problems. Over the past few years rapid development in genomics and proteomics research technologies and developments in information technologies have combined to get information about genes and its sequences in molecular biology. The main purpose of this is to increase the understanding of biological processes. Bioinformatics plays an important role in gathering, identifying, analyzing, storing and classify the genetic data. There are millions of genes data. Nowdays it is essential to understand the functions of genome sequences because there is huge volume of gene data are available for identify and analyze that genes or it needs to monitor the patterns of thousands of genes.for this purpose need to apply some computational methods.Many researchers have been studying many problems of gene clasiification and

2327-0659/15 $31.00 © 2015 IEEE
DOI 10.1109/ACCT.2015.73

attempting to propose the optimal classification technique to work out these problems.some produce better results than others, but that have been still no comprehensive work to compare the possible classifiers.we need through effort to give the evaluation of possible methods to solve the problems of genes data. In this paper we attempt to explore many classifier which is more efficient for classification.

This paper contains the survey on different approaches to gene sequence classification using machine learning techniques. Section II contains machine learning in Bioinformatics. Section III contains the literature review on some latest approaches using machine learning techniques for gene sequence classification. After that in section IV & V contains the challenges & issues in Bioinformatics. Section VI contains DNA Sequencing. Section VII summarizes the comparative study of different bioinformatics problems on the bases of different techniques and database employs with their pros and cons.

II. Machine Learning In Bioinformatics

Computational intelligence techniques have many characteristics such as adaption and fault tolerance that made them attractive for research on bioinformatics. A machine learning approach is introduced for classifying network. The objective of machine learning is to discover and learn and then adapt to the circumstances that might change over time and therefore improving the performance of the machine. In the field of bioinformatics, the reference input is used for the algorithms of machine learning so that they "learn". The ability of soft computing techniques to deal with uncertain and partially true data makes them attractive to be applied in bioinformatics

- Machine learning techniques can be used here to train the network for better performance and enhancing the accuracy of the system.
- Moreover, Machine learning tools are used to decrease false positive rates.

Machine learning can be defined as the ability of the computing machine to increase its performance based on previous results. The figure represents the various machine learning tools that plays an important role in the field of Bioinformatics. Nowadays classification problem become a dangerous task to handle the biological data and it is not possible by using the traditional methods. Artificial Neural Network is the widely used machine learning tool in the bioinformatics. Neural network provides learning capability and it is one of the important components of soft computing. A neural network will consist of one input layer, one or more number of hidden layers and an output layer. Neural networks are used in bioinformatics for the purpose of property prediction and classification of genes in different classes.

There are supervised learning and unsupervised learning methods. In supervised learning the network learn by examples while in unsupervised it learns by itself. There are unknown class labels. When the class labels are known then it produce a model that can classify the new examples that have not seen by learner. And the evaluation of this learning method is done by N-Cross Validation. In this the 90% of examples are considered as learning algorithm while remaining is used to estimate the future accuracy of the learned model. There are many applications of biology where the neural network is applied [5].

- The coding region recognization & gene identification problem
- Identification and analysis of the signals binding sites or regulatory sites
- Sequence classification & feature detection

III. Literature Study

Genomic sequence, protein structure, gene expression microarrays, and gene regulatory networks are some of the application areas described. Since the work entails processing huge amounts of incomplete or ambiguous biological data, we can utilize the learning ability of the Machine learning techniques to solve this kind of problems. The machine learning techniques will train the system to classify the genes data [16].

The machine learning techniques enhances the problem which occurs in the Biological areas, there is a need for modern techniques which handles the genes data. There are many machine learning methods are used for identification, selection, prediction, recognization and also in classification of the DNA Sequences. A Neural Network based multi classifier are used for the identification of the gene in the DNA Sequences. Genes in DNA is preceeded by promoter sequences. A promoter which frequently appears before its associated gene in DNA sequence governs the expression of genes which identitifying genes that are based on their DNA sequences. Promoter which are used to identify the genes. For predicting the location of the promoter neural network is used because it classifies the promoter either presence of promoter or the absence of the promoter in the sequences. Data used from E.Coli promoters which contains the 324 known promoters and 429 unknown promoters that is non-promoters.

Use of artificial neural network classifier to predict the promoter of DNA sequences and evaluate their performances. Similarly in another paper they proposed a new hybrid learning system which is used to recognize the promoters in the DNA Sequences which involve in the bindings of RNA to initiate the process of transcription. A promoter which frequently appears before its associated gene in DNA sequence governs the expression of genes which identitifying genes that are based on their DNA sequences. In this paper promoter production occurs in two processes: transcription and translation. Transcription and Translation plays an important role to predict the position of promoter in DNA sequencing. Promoters are DNA sequence

that defines the transcription of genes begins the initiate of transcription process. Transcription is the process of making RNA copy in gene sequence which is known as messenger RNA (mRNA) after that translation process occurs which translates the sequence of mRNA that is known as cytoplasm. Use of artificial neural network classifier to predict the promoter of DNA sequences and evaluate their performances. They used SVM and ANN method. In SVM method which classify the unlabeled input samples by using non-linear kernel functions [15].

It shows that this classifier is better than other existing techniques for indentifying promoter regions. Use of ANN they classify into two classes promoters and non-promoters. They classify by using four neural networks and obtained results shows that the classification accuracy was high. It performs better than the single NN. It is more efficient and effective technique for classify the promoters [3] [6].

In this paper they proposed a methodology for selecting the genes that have a role in mediating some diseases and certain cancer also. Here gene selection refers to the task of selecting some important genes. It is not possible to focus on a small number of genes that have different pattern expression in diseased samples. So, use of effective and efficient gene selection method for selecting important genes from the whole genome which plays an important direct/indirect role in causing the diseases. hybrid model is used. It means the concepts of fuzzy sets and ANN is used as a proposed model which is known as a NeuroFuzzy Model. In disease mediating genes, those genes which cause disease. So, they change their behaviour from the normal condition when the symptoms of the disease are not present [7].

In biological processes, the Lipid Binding Proteins (LBPs) have an important role. The major problem in LBPs is to sequence, structure and function which results in low accuracy. There is need for developing prediction methods to identify the LBPs from the Non-LBPs. In this paper there is comparative study of the performance of Support Vector machine and Artificial Neural Network is done. It means the performance of SVM and ANN are compared to classify the LBPs from the Non-LBPs. It has been found that SVM was more successful between the LBPs and the Non- LBPs than the ANN. Lipids are like some energy homeostasis, cell signalling, and formation of membranes. So some diseases are related to this disorder of the lipids. It is one type of functional proteins such as cell growth, regulation of gene expression, lipid transport, lipid metabolism which response to the bacterial function.

The machine learning methods are used for this purpose in two steps: first is extracting the dimensional features vectors with class label and second step is use of any machine learning methods as classifier for prediction the class labels. So, the ANN and SVM is used to predict the LBPs classes which is divided into nine classes which described the characteristics of each protein the prediction accuracy of the Non-LBPs was higher than LBPs classes for this situation SVM is fit for this [8].

Support Vector Machines is one of the non parametric controlled classifier it is a two class classification method because basically it is divided into two groups linear support vector machines and Non-Linear support vector machines. One of the main tasks in DNA sequencing is to classify the quality of DNA sequencing data .it needs to be classified as high or low quality. If the quality of data is known to be low data might be reproduced by repeating necessary reactions. It means we can use SVM for automatic control the quality of DNA chromatograms. So here developed new quality evaluation technique where the quality of DNA chromatograms is classified as low or high. It is two class classification problem that's why we choose Support Vector Machine . SVM is trained on a training set to learn the hyperplane that is why it is called SVM learning .then run on testing set to create a confusion matrix. it is only used as automatic screening of DNA data. it means where large number of DNA sequencing data are available we can use SVM as a powerful classification technique [9].

Hierarchical classification is a problem with many areas. So there is need for a algorithm. This paper represents a algorithm for hierarchical classification called Multi label Hierarchical Classification using a Competitive Neural Network (MHC-CNN). Generally classification done in two ways: conventional classification and hierarchical classification. In conventional the set of classes, where each class is independent of the other classes and in hierarchical classification classes are in hierarchy structure like tree structure and they are dependent on each other. Also they are two types of prediction of classes: mandatory prediction (possible in both conventional and hierarchical classification) and non-mandatory prediction (possible only in hierarchical classification). So, the possible solution has been given in this paper to explore the hierarchical classification: Flat hierarchical classification (to predict the leaf node of the hierarchy structure), local hierarchical classification (for per node or parent node) and the global or big bang classifier [10].

Classification in bioinformatics is an important area of research and the most important task is the performance of classification. In this paper a new hybrid model was proposed that combines artificial intelligence into fuzzy, the unique advantage is both fuzzy and the classification power of artificial neural network, to construct an efficient and accurate classifier. It means using of fuzzy parameters without using of crisp parameter. The amount of data for network training depends upon the input data but complexity and the time consuming in the training the network was poor. So we have to apply some fuzzy rules. Proposed model gives an accurate result than the traditional neural networks where inadequate data are available. ANN is universal classifier that can be applied to the classification problems with the high degree of accuracy. This model faces certain problems like we have to train the network

over a short span of time. So in this paper a new ANN is proposed using the advantage of fuzzy logic for classifying the genes data. Five well known statistical and intelligent classification models: Linear Discriminant Analysis, Quadratic Discriminant Analysis, K-Nearest Neighbour, Support Vector machine and Artificial Neural Network are used in this paper which is compared with the new hybrid proposed model. It shows the effectiveness and appropriateness of the proposed model for gene expression classification. It can be applied for solving problems in scant data, specifically when higher classification accuracy is needed [11].

We know that the DNA Sequence is one of the main issues in bioinformatics. So there is need for a proper classification of the sequences. Because DNA Sequence are used to determine the function of DNA. In this paper they show the difference between the traditional method which uses Artificial Neural Network and the new classification method. There are some problems when we use the traditional method. The main existing problems are all the frequency of the nucleotides bases in the DNA Sequence are calculated twice. So the result of the classification is not good and when use the ANN method, it is waste of time and cannot suitable. So there is greater difference on the precise classification between the difference data. To use DNA sequence classification analysis, it needs to be extracted. Then the next step is redundancy removal which is used to reduce the complexity. Then we used k-means algorithm in dataset for clustering. Those objects which are in same cluster are of higher similarity and different are smaller. A mean value, centre of gravity is generated. Input the parameter and the DNA Sequence, and then the data is clustered into small clusters. It is used to cluster the large dataset into smaller ones. After the clustering process we used a classification method that is SVM, which is one of the machines learning method by training. When SVM used the data need to be normalization within the given range [12].

DNA sequence is Splice site junction (introns-exon or exon-intron) sites. For the successful gene prediction it is important that splice junction sites are detected in the DNA sequences. So, in this paper they used SVM for classification of DNA and Splice site junction. The information like motifs, clusters, promoters, genes and protein signatures are considered as a splice-junction site recognization. Splice is said as modification of genetic information after the transfer of information from DNA to RNA, in which introns are removed and the exons are added. So, for indentifying that which sequences has splice sites or not. The splice site in the upstream part of an intron is known as Donor and in the downstream part known as Acceptor. SVM classifier is best for identification of the splice junction sites in bioinformatics. In this paper the main aim is to classify the DNA sequences into two classes as splice/negative and EI/IE. So SVM is labelled as positive or negative [13].

IV. CHALLENGES IN BIOINFORMATICS

The first challenges facing the bioinformatics today is the intelligent and efficient storage of this massive data. it is essential to provide easy and reliable access to this data. The data itself is meaningless if there is no relevant data are available then it is need to analyze and interpret properly.

1) Precise, predictive model of transcription initiation and termination: ability to predict where and when transcription will occur in a genome
2) Precise, predictive model of RNA splicing/alternative splicing: ability to predict the splicing pattern of any primary transcript
3) Precise, quantitative models of signal transduction pathways: ability to predict cellular response to external stimuli
4) Determining effective protein-DNA, protein-RNA and protein-protein recognition codes
5) Rational design of small molecule inhibitors of proteins
6) Mechanistic understanding of protein evolution: understanding exactly how new protein functions evolve
7) Mechanistic understanding of speciation: molecular details of how speciation occurs
8) Continued development of effective gene ontologies systematic ways to describe the functions of any gene or protein (Infrastructure and education challenge)

So, the intelligent computer tools must be developed to allow the extraction of meaningful biological information. There are three biological processes around which bioinformatics tools must be developed.

- DNA Sequence which determines protein sequence.
- Protein sequence which determines protein structure.
- Protein structure which determines protein function.

V. ISSUES IN BIOINFORMATICS

Biological sequence has become one of the main issues related to the sequencing problem such as DNA, RNA and protein sequence. There are various issues of bioinformatics which leads to major problems [2]. All are explained below:

A. *Sequence Analysis:*

1) *Genome Sequencing:* Advances in sequencing technologies provide opportunities in bioinformatics for management, processing and analyzing the sequences. Each of these sequencing has significant analytical challenges for bioinformatics in terms of experimental design, data interpretation and analyze of data.

2) *Gene Finding and Genome Annotation*: Gene finding refers to prediction of introns and extrons in a segment of DNA sequence. Many of the computer programs are used for identifying protein-coding genes are available. Important aspect of genome annotation is the analysis of repetitive DNA. Which are copied of identical or nearly identical sequences present in the genome.
3) *Sequence comparison:* comparing the sequence provides a foundation for many bioinformatics tools and may allow inference of the function, structure and evolution of genes and genomes.

B. *Transcriptome Analysis:*

The main goal is to learn about how changes in transcript abundant control growth and development of an organism. DNA microarrays proved a best technology for observing transcriptome.

1) *Microarray Analysis:* Microarray analysis allows the simultaneously measurement of transcript abundance for thousands of genes. This issues are in microarray analysis are in processing and normalization data. Some requires multiple biological replicates and statically valid results before publishing the microarray results.
2) *Tiling Arrays:* Microarray sample known and predicted genes. It cover the genome at regular intervals to measure the transcription without bias toward known or predict gene structures, discovery of polymorphisms, analysis of alternative splicing and identification of transcription factor binding sites.
3) *Regulatory Sequence Analysis:* Interpreting the results of microarray experiments involves discovery why genes with similar expression profile behave in a coordinated fashion. It is for extracting motifs that are shared between the upstream sequences of these genes.

C. *Computational Proteomics:*

It is used for the qualitative and quantitative characterization of proteins and their interaction on a genome scale. It includes identification and quantification of all protein types in a cell or tissues and association with other proteins.

1) *Electrophoresis Analysis:* It can qualitatively and quantitatively investigate expression proteins under different conditions. Several bioinformatics tools have been developed for 2 D electrophoresis analysis. It includes limited ability to identify proteins and low accuracy in detecting protein abundance.
2) *Protein Identification through Mass Spectrometry:* After protein separation using 2D electrophoresis, proteins are identified by mass spectrometry. The limitation in mass spectrometry protein identification is lack of open source software. Most of results are unreliable.

VI. DNA SEQUENCING IN BIOINFORMATICS

DNA sequencing may be used to determine the sequence of individual genes, larger genetic regions (i.e. clusters of genes or operons), full chromosomes or entire genomes. Sequencing provides the order of individual nucleotides in DNA or RNA (commonly represented as A, C, G, T, and U) isolated from cells of animals, plants, bacteria, archaea, or virtually any other source of genetic information. This is useful for:

- Molecular biology – studying the genome itself, how proteins are made, what proteins are made, identifying new genes and associations with diseases and phenotypes, and identifying potential drug targets
- Evolutionary biology – studying how different organisms are related and how they evolved
- Metagenomics – Identifying species present in a body of water, sewage, dirt, and debris filtered from the air.

Less-precise information is produced by non-sequencing techniques like DNA fingerprinting. This information may be easier to obtain and is useful for:

- Detect the presence of known genes for medical purposes (see genetic testing)
- Forensic identification
- Parental testing

One of the most common and robust techniques performed in molecular biological laboratories. Unfortunately, it does not always work and when it doesn't it can be very difficult to work out what went wrong. Fortunately, most failed (or sub-optimal) DNA sequencing results have only a fairly limited number of causes. To help in the troubleshooting of sequencing problems we have created a series of guides for identifying the most common causes of various sequencing problems.

Identifying the cause of a poor DNA sequencing result can often be very difficult as a particular sequencing problem may have many different causes, or be the result of multiple interacting factors. Often the only way to work out the real cause of a particular problem is to perform a process of elimination. This process can be greatly simplified by visually examining both the raw and processed data chromatograms of the problematic sequencing traces.

So, sequence is like ATCG. Now, if person is affected with any disease, actually its protein or genes get altered or affected. This causes alteration in gene sequence. This

altered gene is termed as abnormal gene which shows abnormal behaviour. So, it's necessary to detect these genes as soon as possible to cure person from dangerous disease. But, the problem here is how to detect this abnormal gene and which technique is most prominent in doing it effectively. Traditionally detection of abnormal genes is detected using extensive analysis and using medical tools to perform various tests. The techniques used were completely based on trial and error. Clearly, this is not a Good practice to do it. Sooner many data mining algorithms and machine learning algorithms were used to simplify the task with great accuracy and better performance. Many machine learning algorithms like ANN and SVM are already used in solving the issue of classifying normal genes from abnormal genes.

VII. COMPARATIVE STUDY

The comparative analysis of classification of gene data's using machine learning techniques is given in the table. The table shows the technique they use in each phase, bioinformatics database used, key points of the system and advantage and disadvantages of that system.

TABLE I
COMPARATIVE SURVEY ON GENE SEQUENCE CLASSIFICATION USING MACHINE LEARNING TECHNIQUES

Sr. no.	Context in Use/Dataset	Techniques	Key Points +Pros and – Cons
1	For splice site recognization in DNA Sequences [2] **Dataset:** UCI Repository of machine learning databases from GenBank	Support Vector Machine (SVM)	+It performs better result for identifying the Splice sites. -It needs appropriate kernel functions for training the data otherwise leads to poor classification
2	For classification in DNA Sequences **Dataset:** USA National Center for Biotechnology (NCBI) from GenBank	Support Vector Machine (SVM) & Artificial Neural Network (ANN)	+This algorithm having high precise than the traditional ANN method. -Only for scant data.
3	For Multi Label Hierarchical classification for protein function prediction **Dataset:** Cell cycle, Church, Derisi, Einsen and Spo Datasets	Artificial Neural Network (ANN)	+It is for Multi Label classification. -It is only suitable for hierarchy structure not for flat.
4	For classification in gene expression data analysis **Dataset:** microarray datasets	Support Vector machine & Artificial Neural Network	+It is more efficient than the traditional methods with highest classification accuracy. -For large datasets not for small dataset.
5	For Quality control in DNA Sequences **Dataset:** DNA chromatograms from InSNP databases	Support Vector Machine (SVM)	+It is used to classify quality of DNA chromatograms either high/low. -It is only for automatic screening of DNA chromatograms.
6	For prediction promoter location in DNA Sequences **Dataset:** E-Cole Promoter Genes Dataset	Artificial Neural Network (ANN)	+If no. of neurons increased then the classification accuracy also increases. -It is suitable for the identification of promoters
7	For selection of genes that have direct/indirect role in Cancer **Dataset:** Human lung, colon, breast cell,soft tissue sarcoma, lymphocytes and plasma cell expression data	Artificial NeuroFuzzy inference System (ANFIS)	+It performs better for identifying important genes in mediating cancers. -Complexity issues in making the important groups
8	For prediction of Liquid Binding Proteins in Sequence Homology **Dataset:** Lipid Binding Proteins	Artificial Neural Network (ANN) & Support Vector Machine (SVM)	+It is for classify the LBPs from the Non-LBPs -ANN has poor performance than the SVM. It has lowest accuracy than the SVM

VIII. ISSUES

There are a lot of methods and algorithms that are already available for DNA Sequencing problems, but they have their own benefits and limitations as given in the previous section. From these studies some key issues that affect is highlighted. These issues have an important impact on design of Gene Sequence Classification.

The machine learning methods must be chosen very carefully. Machine learning techniques are more effective and appropriate for the classification of genes so, there is need for a mechanism which improves the problems which are faces by traditional methods. Classifier's efficiency

indirectly depends on what type of input is given. Most of the existing system takes a long **training times** to train the system or delay to give a desired output, which is unacceptable in real world problems. Many existing systems have used same dataset for train a classifier and test classifier which is not an effective way to build and test a classifier. When gene data's taken for testing is different than the data used for training gives a more effective classifier and it would be a more challenging task. It is also known as **database independency**.

IX. CONCLUSION

From extensive survey on machine learning and problem areas of gene sequencing, we conclude that advances in machine learning technique can be one of the prominent solution in solving classification/regression issues in identifying abnormal genes distorting gene sequences.

The paper shows a survey of recent trends to automatic classification of gene data's using machine learning. It is the most attractive field nowadays and proves effective techniques to the problem of classification, prediction, optimization, pattern recognition, image processing, etc. From this observation we can conclude that this techniques are fit for the Bioinformatics datasets. Many machine learning algorithms like ANN and SVM are already used in solving the issue of classifying normal genes from abnormal genes. But, problem with Ann is that it is unable to perform better for non- linear data and SVM is very complicated and costly to implement in this scenario. So, on the basis of my extensive literature survey I have decided to use some technique which can deal with non-linear data and fuzzy behavior of data. ANFIS is such algorithm. It is used in various fields like signal processing, medical diagnosis etc. From the survey it conclude that the ANN & SVM are mostly used in bioinformatics areas and they perform well also gives better result but there are certain problems which are faces by this technique. So, Artificial NeuroFuzzy Inference System is good as compare to these methods and gives effective results.

REFERENCES

[1] "A Survey of Existing Literatures on Bioinformatics Research", International Journal of advanced studies in Computer Science and Engineering IJASCSE, Volume 3, Issue 5, 2014

[2] "Bioinformatics and Its Applications in Plant Biology", Annual Review of Plant Biology Volume 57, 2006

[3] "A neural network based multi-classifier system for gene identification in DNA sequences",Romesh Ranawana E Vasile Palade, Neural Comput & Applic (2005) 14: 122–131

[4] "Towards cognitive Analysis of DNA", Witold Kinsner, Proc. 9th IEEE Int. Conf. on Cognitive Informatics

[5] "Using Machine Learning To Design And Interpret Gene-Expression Microarrays", Michael Molla, Michael Waddell, David Page, And Jude Shavlik

[6] "Predicting Functional Regions in Genomic DNA Sequences Using Artificial Neural Network", Gunay Karli, Adem Karadag, International Journal of Engineering Inventions Volume 3, Issue 6 (January 2014)

[7] "Selection of genes mediating certain cancers, using a Neuro fuzzy approach", Anupam ghosh, Bibhas Chandra dhara and Rajat k.ve, Elesevier, Science Direct, Neurocomputing 133 (2014)

[8] "Neural network and SVM Classifiers accurately predict lipid binding proteins, irrespective of sequence homology", Mehdi Khashei, Ali Zeinal Hamadani, Mehdi Bijari, Elsevier, Science Direct, Journal of Theoretical Biology (2014)

[9] "Support vector machines for quality control of DNA sequencing", Ersoy Oz and Huseyin Kaya, Journal of Inequalities and Application 2013 (a Springer Open journal)

[10] "Multi-Label Hierarchical Classification using a Competitive Neural Network for Protein Function Prediction", Helyane Bronoski Borges and Julio Cesar Nievola , WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15, 2012

[11] "A fuzzy intelligent approach to the classification problem in gene expression Data analysis", Mehdi Khashei, Ali Zeinal Hamadani, Mehdi Bijari, Elsevier, Knowledge-Based Systems 27 (2012)

[12] "A New Method for Classification in DNA Sequence", Qingda Zhou, Qingshan Jiang and DanWei , The 6th International Conference on Computer Science & Education (ICCSE 2011) August 3-5, 2011 IEEE

[13] Splice Site Recognition in DNA Sequences Using K-mer Frequency Based Mapping for Support Vector Machine with Power Series Kernel", Robertas Damasevicius, International Conference on Complex, Intelligent and Software Intensive Systems, 2008 IEEE

[14] "Application of Data mining in Bioinformatics", Indian journal of computer of science and engineering Volume 1 No.2 114-11

[15] "Integration of Knowledge-Discovery and Artificial-Intelligence Approaches For Promoter Recognition in Dna Sequences", Yin-Fu Huang Chia-Ming Wang, International Conference On Information Technology And Applications, 2005 IEEE

[16] "Bioinformatics With Soft Computing", Sushmita Mitra And Yoichi Hayashi, IEEE Transactions On Systems, Man, And Cybernetics 2006