International Journal of Computational Intelligence and Informatics, Vol. 3: No. 4, January - March 2014

Naïve Bayes Classification for Predicting Diseases in Haemoglobin Protein Sequences

S Vijayarani
Department of Computer Science
Bharahtiar University, Coimbatore, India
vijimohan_2000@yahoo.com

S Deepa
Department of Computer Science
Bharahtiar University, Coimbatore, India
deepasamiappan@gmail.com

***Abstract*- The development of sequencing techniques led to an exponential growth of protein sequences in the public databases. The sequential information has been successfully applied to unveil the structures, functions, evolutionary relationships, etc. Lot of computational methods have been developed to classify the protein sequences and to predict the diseases based on their sequence information. The classification of biological sequences is one of the significant challenges in bioinformatics as well in genomics and proteomics. The existence of these sequence data in huge masses and their indistinctness and especially the high costs for lab experiments make use of data mining in disease prediction methods which are applied instead of laboratory experiments. Since a wide number of diseases are based on proteins and their sequences, the protein sequence analysis has been of great attention recently. The use of data mining techniques in protein sequence analysis provides an efficient way for examining the proteins to identify their characteristics and it also provides a way for better drug designing. In this research work, the hemoglobin protein based diseases are predicted by applying Naïve Bayes classifier. The performance of this classifier is analyzed by the factors classification accuracy and execution time.**

Keywords- Sequences, Classification, Protein data bank, Naïve bayes approach

I. INTRODUCTION

In recent trends, technologies such as computers, satellites and many others have led to an exponential growth of collected data in various areas. It is clear that traditional data analysis techniques do not have sufficient power to process large amounts of data efficiently. In this case data mining technology is only way that can extract knowledge from large amount of data. In recent times the collection of biological data like protein sequences and the DNA sequences is increasing at explosive rate due to improvements of existing methods and technology. So data mining techniques are used to extract the meaningful information from the huge amount of biological data sequences such as the DNA or protein sequences etc. One of the important problems in this area of research is protein sequence classification in which protein sequences are classified into different classes or sub classes or families.

Classification is the most important technique to identify a particular character or a collection of them. A number of classification methods or algorithms have been proposed by different researchers to classify the protein sequences. Some of the well known protein sequence classification techniques are stepwise regression-based feature selection, ICA-based feature transformation which involves extraction of specific features from the sequences. These features are based on the structural and functional properties of amino acids. Other methods and techniques used for this classification includes Fuzzy ARTMAP, Genetic algorithm, Rough Set based Classifier, neural networks, etc.

Through the protein data bank, a protein sequence is easily obtained and analysed for various studies and applications. The conventional methods such as neural network and genetic algorithm have been applied for the protein classification. But these methods require a number of empirical parameters and cannot have reasonable predictions for various cases. The development of methods to assess the impact of amino acid mutations in proteins on human health has become an important goal in biomedical research. The computational methods constitute a valuable tool because they can easily process large amounts of mutations and give useful information on their pathological character [7].

The rest of the paper is organized as follows: Section 2 describes the related work. The problem objective is given in Section 3. Section 4 discusses the details about the dataset, and Naïve bayes classification approach for disease prediction. The performance evaluation is given in Section 5 and the conclusion is given in Section 6.

ISSN: 2349 - 6363

II. LITERATURE REVIEW

Algorithms that have been used for protein sequence classification can be classified roughly into several types, depending on whether they are based on the K-Nearest Neighbor (K-NN) approach, the Hidden Markov Model (HMM) approach, or the Support Vector Machine (SVM) approach or any other classification algorithms.

In the context of protein sequence classification, Markov models (MMs) are used to capture dependencies between the neighboring sequence elements. MMs are among the most widely used generative models of sequence data [4]. In a kth order MM, the sequence elements satisfy the Markov property: each element is independent of the rest given the K preceding elements. Begleiter et al.[2] have examined methods for prediction using variable order MMs, including probabilistic suffix trees, which can be viewed as variants of abstraction wherein the abstractions are constrained to share suffixes.

Due to its simplicity, the K-NN approach to classification (Fix et al., 1949) [8] has been popular in the biological domain (Deshpande et al., 2002; Lu et al., 2003) [4, 11]. Given a database of pre-classified sequences, a new sequence can be classified by finding k sequences in the database. It is then assigned to the class that the majority of these k sequences belong to. The key step in building a K-NN classifier is to determine how similar two sequences are, and the measure of similarity is usually determined by computing global or local alignment scores. For this purpose, the most popular algorithm used is the Smith-Waterman dynamic programming algorithm (Smith et al., 1981) [15]. This algorithm is relatively accurate, but it is not very computationally efficient. Heuristic algorithms such as BLAST (Altschul et al., 1990) [1] and FASTA (Pearson, 1990) [13] have therefore been developed to trade reduced accuracy for improved efficiency.

The HMM-based approaches (Rabiner, 1989) [14] to protein sequence classification have been shown to be effective in detecting for conserved residue patterns in a set of protein sequences (Eddy et al., 1995; Hughey et al., 1996 ;) [6], [9]. A typical HMM consists of a chain of match, insert, and delete states in a Markov chain, with all transitions between states and all residue costs in the insert and match states trained to specific probabilities. When the HMM is trained on a set of sequences that are members of a given protein family, the model parameters are learned via an expectation-maximization approach and a form of dynamic programming is used to detect for similarity. The resulting HMM can identify the positions of residues that can describe conserved primary structures of a family and it can then be used to discriminate between family and non-family members.

The SVM-based approaches (Cristianini et al., 2002; Vapnik, 1998) [3] to classification use both positive and negative examples when training a classifier. They perform protein sequence classification by mapping the input training sequences into a high dimensional feature space and try to locate in the feature space a plane that maintains a maximum margin from any point in the training set. Then the class label of the unclassified sequence is predicted by mapping it into the feature space and deciding on which side of the separating plane, the given sequence lies. The SVM-pairwise method (Schölkopf et al., 2004) [16] also requires that a given set of protein sequences be converted into fixed-length vectors first. SVM is then trained from the vectorized protein sequences. A list of pairwise sequence similarity scores are computed by using the dynamic programming algorithm.

The K-NN-, HMM-, and SVM-based algorithms are most commonly used for protein sequence classification. In K-NN-based approach, the number of k needs to be determined in advance. Also, pairwise alignment is computationally inefficient and the reliability of similarity detection falls rapidly whenever the pairwise sequence identity drops below 30%. In HMM based algorithms, many parameters are needed to be estimated accurately and this requires a large amount of training data which may not always be readily available. For SVM based approaches, all input sequences need to be aligned beforehand in order to transform them into fixed-length vectors in the feature space and the alignment process can be difficult and time-consuming.

III. PROBLEM OBJECTIVE AND METHODOLOGY

Since the proteins are made up of amino acids sequence, a single change in the sequence of amino acid may cause certain diseases. The diseases considered for this study is based on the sequences of hemoglobin protein in the human. The change in the sequences of the hemoglobin protein may cause a variety of diseases namely sickle cell anaemia, Haemophilia, Thrombophilia, Retinal Regeneration, Thalassemia, Thrombosis, Cystic Fibrosis, Mild hemolytic symptom, etc. The sequential information taken for this research work mainly involves five types of diseases namely Sickle cell anaemia, Mild hemolytic symptom, Haemophilia, Cystic Fibrosis and Thalassemia.

The main objective of this research work is to classify the sequences as diseased and normal, based on the amino acid composition of the sequences. In this research work, when the dataset (Patient information such as patient id, Patient name, Patient sequence, etc) is given as input, they are classified into diseased and normal sequences by using Naïve bayes approach.

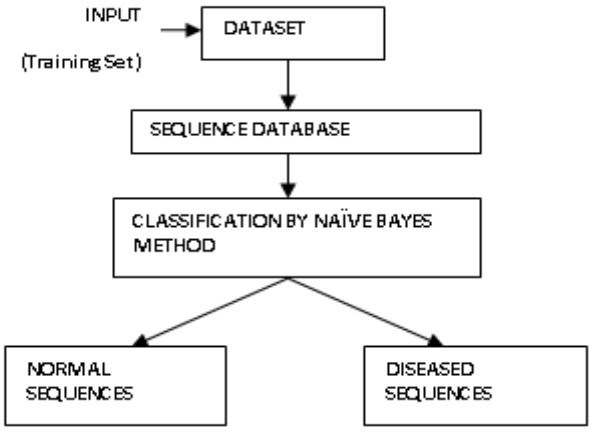

Figure 1. System architecture for the classification of sequences

IV. PROTEIN SEQUENCE CLASSIFICATION

Sequence classification methods require the knowledge of at least part of the amino acid or nucleotide sequence for a protein. Various algorithms and methods are then used to compare and classify the sequences. The resulting families contain the proteins that are related by the sequences. An obvious shortcoming of sequence-based classifications is that they can only be applied to proteins for which the sequence information is available. The sequence based classification techniques allow the classification of proteins for which no biochemical evidence has been obtained such as the thousands of uncharacterized sequences of carbohydrate-active enzymes that originate from genome sequencing efforts worldwide.

A. Dataset

The dataset taken for this research work is a synthetic dataset. It is created with three attributes namely Patient Id, Patient Name, age, gender, blood group and Protein sequence information of the patient. The length of the protein sequences may be 150, 300 and even up to 2000. The protein sequences consists of the combinations of different amino acids. In this study, the sequences of hemoglobin proteins are only considered. The total number of patient information is taken initially as 20 and then increased to 50 to analyze the performance of the naïve Bayes approach for classification. A sample of a hemoglobin protein which consists of 147 amino acids is shown below:

Hemoglobin protein sequence: VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDL
STPDAVMGPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELVDPENF
RLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH

B. Sequence Database

The dataset used for evaluating the performance of Naïve Bayes method for disease prediction is taken from PDB (Protein Data Bank), which is a universal repository that stores the sequences of all identified proteins. The sequence of a particular protein can be taken by giving the protein name in to the PDB and the sequences are retrieved easily. In this research work, the sequences for hemoglobin proteins are taken from the patients and stored in a sequence database. A sample of the sequence database is given below:

TABLE I. SAMPLE OF SEQUENCE DATABASE

S.No	Patient Id	Patient Name	Age	Gender	Blood Group	Protein sequence
1	P01	Alwin	43	Male	B -ve	VHLTPEEKSAVTALWGKVNVDEVGGE ALGRLLVVYPWTQRFFESFGDLSTPDA VMGNPKVKAHGKKVLGAFSDGLAHL DNLKGTFATLSELVDPENFRLLGNVLV CVLAHHFGKEFTPPVQAAYQKVVAGV ANALAHKYH
2	P02	Farry	32	Male	A1 -ve	VHLTPEEKSAVTALWGKVNVDEVGGE ALGRLLVVYPWTQRFFESFGDLSTPDA VMGNPKVKAHGKKVLGAFSDGLAHL DNLKGTFATLSELHCDKLHVDPENFRL LGNVLVCVLAHHFGKEFTPPVQAAYQ KVVAGVANALAHKYH

C. Naive Bayes Classification [12]

Naive Bayes Classifiers are one of the simplest and one of the most effective methods for classification. This method is based on the idea of Bayesian Networks which is a probabilistic graphical model representing a set of random variables and their conditional independencies. In the Bayesian Networks, there are several efficient algorithms that perform inference and learning. The only requirement for it to be applicable is the features of the dataset should be independent. The features in the dataset are dependent to each other because of the evolution of species but the dependence does not seem to be a strong one. So by considering that the features of the used dataset are independent, the classification is done based on the Naïve bayes method. Initially this approach creates the probability value for each of the sequences in the sequence database, Sndb. When the input dataset is given, the probability value for the sequence of each patient is calculated. This value is compared with the probability value in the Sndb database. Thus the sequences are classified based on the value of the probabilities of the sequences. Naïve Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong independence assumption.

The probability of data record X having the class label Cj is:

$$P(Cj|X) = \frac{P(X|Cj)*P(Cj)}{P(X)} \quad (1)$$

The class label Cj with largest conditional probability value determines the category of the sequence in the given input dataset.

V. EXPERIMENTAL RESULTS

In this research work, Naïve Bayes approach is used for disease prediction based on the protein sequence information which is extracted from the patients. The algorithm has been used with the aim of getting a better accuracy to classify the normal and diseased proteins using the Naïve Bayes approach. With the aim of getting better accuracy, input and training has been given to the Naïve Bayes method with 20 and 50 instances and the accuracy value is calculated. The input dataset consists of the patient details such as patient id, patient name, protein sequence information. The input protein sequences will be compared with the trained data and they are classified into normal and diseased sequences. The accuracy value of the naïve bayes classifier is calculated by identifying the number of patient information that has been correctly classified. It can be calculated as

Accuracy of naïve bayes classifier = (Number of records correctly classified / Total number of records) *100

TABLE II. ACCURACY VALUE FOR VARIOUS SIZED DATASETS

Classification by Naïve Bayes Approach	
Dataset size	Accuracy Value (in %)
20	85%
50	84%

The above table shows the accuracy value for the classification of diseased and normal sequences. The following graph shows the value of accuracy for different sized datasets.

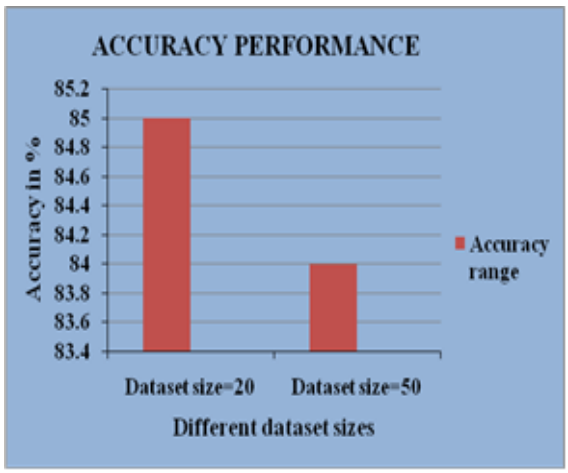

Figure 2. Classification accuracy based on naïve bayes method for datasets of different sizes

The following table shows the execution time taken to classify the diseases using Naïve Bayes Classification method.

TABLE III. EXECUTION TIME FOR CLASSIFICATION

Dataet Size	Execution Time(in MilliSeconds)				
	Sickle Cell Anemia	Thales-semia	Mild Hemolytic Symptom	Haemo-philia	Cystic Fibrosis
20	15.4	10.3	15.2	20	25.3
50	61.56	56.65	33.6	70	50.4

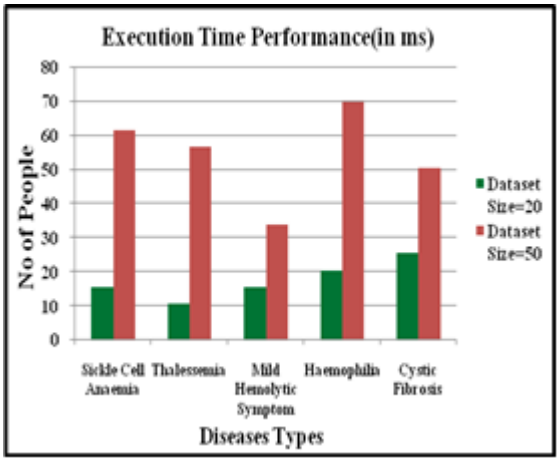

Figure 3. Execution time for classification

The naïve bayes classifier analyzes each protein sequence information and identifies whether the particular sequence is diseased or normal. Initially, the classifier calculates the probability value of the normal sequence and the diseased sequences from the training data. When new data is given for classification, the probability value of the new sequence is calculated and it is compared with the probability values that has been already calculated. When the probability of the new data matches with the probability value of the already calculated sequences, then the new data can be classified to its corresponding group.

The following table shows the ratio of people affected by each of the diseases in a certain period of time for various dataset sizes.

TABLE IV. RATIO OF DISEASED PEOPLE FOR VARIOUS DATASET SIZES

Dataet Size	TYPES OF DISEASES				
	Sickle Cell Anemia	Thales-ssemia	Mild Hemolytic Symptom	Haemo-philia	Cystic Fibrosis
20	3	2	3	4	5
50	12	11	6	4	9

The following graph shows the ratio of people affected by each of the diseases in a certain period of time.

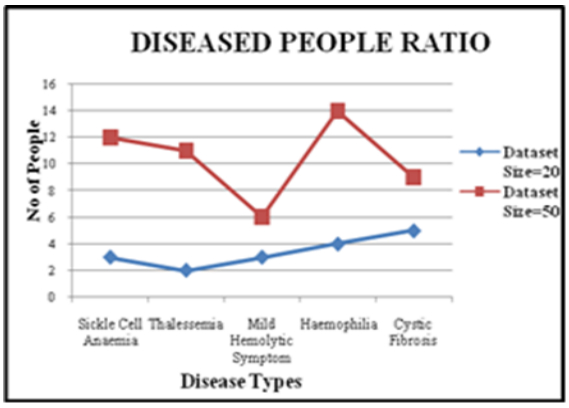

Figure 4. Number of people affected by different diseases for various dataset sizes

VI. CONCLUSION

Data mining approaches ideally suited for bioinformatics since bioinformatics is data-rich but lacks a comprehensive 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, diversity and the lack of a standard ontology to aid the querying of them as well as the heterogeneous data of the quality and provenance information they contain. This work provided the technique for the classification of diseased protein sequences using naïve bayes method. Since a large number of diseases are related with the sequential information, the use of naïve bayes method proves to be an effective method for the diseased sequence classification with better accuracy.

REFERENCES

[1] S.F. Altschul, , W. Gish, and W. Miller, "A basic local alignment search tool", J. Mol. Biol. 215, 403–410, 1990.

[2] R. Begleiter, R. El-Yaniv, and G. Yona, "On prediction using variable order markov models", Journal of Artificial Intelligence Res., vol. 22, pp. 385–421, 2004.

[3] N. Cristianini, and J. Shawe-Tahlor, "An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods", Cambridge University Press, New York, 2002..

[4] M. Deshpande, and G. Karypis, "Evaluation of techniques for classifying biological sequences", PAKDD 2002, 417–431, 2002.

[5] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, "Biological sequence analysis: Probabilistic Models of Proteins and Nucleic Acids", Cambridge University Press., 2004.

[6] S.R Eddy, "Multiple alignment using hidden Markov models", ISMB 114–120, 1995.

[7] C. Ferrer-Costa, M. Orozco, X. de la Cruz, "Sequence-based prediction of pathological mutations".

[8] E. Fix, and J.L Hodges, "Discriminatory Analysis, Non-Parametric Discrimination: Consistency Properties", Technical Report 21-49-004, USAF School of Aviation Medicine, 1949.

[9] R. Hughey, and A. Krogh, "Hidden Markov models for sequence analysis: extension and analysis of the basic method", Comput. Appl. Biosci. 12, 95–107, 1996.

[10] Y. Li and H-M. Lu, "A Computational Efficient Algorithm for Protein Sequence Classification".

[11] Y. Lu, and J. Han, "Cancer classification using gene expression data" Inform. Syst. 28, 243–268, 2003.

[12] Paul Helman, Robert Veroff, R. Susan Atlas and Cheryl Willman "A Bayesian Network Classification Methodology for Gene Expression Data", Journal of Computational Biology 11(4): 581-615. doi:10.1089/cmb.2004.11.581, 2004.

[13] W. R. Pearson, "Rapid and sensitive sequence comparisons with FASTP and FASTA. Methods Enzymol", 183,63–98, 1990.

[14] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", Proc. IEEE 77, 257–286, 1989.

[15] T.F. Smith, and M.S. Waterman, "Identification of common molecular subsequences", J.Mol. Biol. 147, 195–197, 1981.

[16] B. Schölkopf, K. Tsuda, and J. P, Vert, "Kernel Methods in Computational Biology", MIT Press, Cambridge, MA, 2004.

[17] Yang Yang, Bao-Liang Lu, Wen-Yun Yang, "Classification Of Protein Sequences Based On Word Segmentation Methods", Proceedings, October 3, 2007.