PROTEIN SEQUENCE CLASSIFICATION IN DATA MINING– A STUDY

Dr.S.Vijayarani[1] and Ms. S.Deepa[2]

[1]Assistant Professor, Department of Computer Science, School of Computer Science & Engineering, Bharathiar University, Coimbatore, Tamil Nadu
[2]M.Phil Research Scholar, Department of Computer Science, School of Computer Science & Engineering, Bharathiar University, Coimbatore, Tamil Nadu

ABSTRACT

Since the computerized applications are used all around the world, there occurs the collection of a vast amount of data. The important information hidden in vast data is attracting the researchers of multiple disciplines to make study in developing effective approaches to derive the hidden knowledge within them. Data mining may be considered to be the process of extracting or mining the useful and valuable knowledge from large amounts of data. There are various different domains in data mining such as text mining, image mining, sequential pattern mining, web mining and etc. Among these, sequence mining is one of the most important research area which helps to finding the sequential relationships found in the data. Sequence mining is applied in wide range of application areas such as the analysis of customer purchase patterns, web access patterns, weather observations, protein sequencing, DNA sequencing, etc. In protein and DNA analysis, sequence mining techniques are used for sequence alignment, sequence searching and sequence classification. In the area of protein sequence analysis, the researchers are showing their interest in the field of protein sequence classification. It has the ability to discover the recurring structures that exist in the protein sequences. This paper explains various techniques used by different researchers in classifying the proteins and also provides an overview of different protein sequence classification methods.

KEYWORDS

Sequence, Feature hashing, Neural Network Model, Rough set theory, Fuzzy modeling, Classification

1. INTRODUCTION

In recent years, developments in the field of genomics and proteomics is proceeding at a rapid rate and it have generated a large amount of biological data. Obtaining conclusions from these data requires the application of sophisticated computational analyses methods. Bioinformatics is the interdisciplinary science that interprets the biological data using information technology and computer science. The significance of this new field of research will grow as large quantities of genomic, proteomic and other data continue to be generated and integrated. Bioinformatics and computational biology are fields that require skills from a variety of fields to enable the gathering and storing, handling and analyzing, interpreting and spreading of biological information. It requires the usage of high performance computers and innovative software tools to manage and

DOI : 10.5121/ijitmc.2014.2201

organize enormous quantities of genomic and proteomic data. It also comprises the development and application of innovative algorithmic techniques necessary for the analysis and interpretation of sequential data. It is also used for the classification and prediction of sequential data which provides insight into the design and validation of experiments for the life sciences. This is required now with the rapid increase of the amounts of data generated on a daily basis.

Bioinformatics is a field that is committed to the interpretation and analysis of the biological data using computational techniques. Since there had been a fiery growth of biological information which has been generated by the scientific community, the bioinformatics field has developed to a greater extent. [14]. A number of data mining related tasks such as manipulation, searching and data mining of DNA sequence data are performed efficiently. The advancement in techniques for storing and searching DNA sequences have led to the widely applied advances in computer science. Several methods such as string searching algorithms, machine learning and database theory have been introduced and used effectivey. Solving the biological problems is the particular active area of research in bioinformatics where the data mining techniques are applied effectively.

Both the data mining and bioinformatics fields are fast expanding research frontiers. It is important to examine the important research issues in bioinformatics and develop new data mining methods for scalable and effective bio-data analysis [7]. Data mining can assist the researcher in finding out the new knowledge from plenty of biological data at the molecular level [5]. Sequential pattern mining[1] is an important field of data mining, in which a very small alphabet (i.e., 4 for DNA sequences and 20 for protein sequences) and long patterns with a typical length of few hundreds or even thousands frequently appear.

A number of different algorithms have been introduced to carry out the pattern mining task well. The main task involves the finding out of all the sequential patterns with higher or equal support to a predefined minimum support threshold in a sequence database [1]. Sequential data occurs frequently in many scientific, medical, security, business and other applications. Sequence data mining provides the necessary tools and approaches for unlocking useful knowledge hidden in the mountains of sequence data. A sequence is normally an ordered list of objects.

The term sequence in proteins denotes the arrangement of sequence of amino acids that constitutes protein. The sequences of proteins that are identified are being stored in various databases. Each database stores a specific classification of protein sequences. Different classification methods or algorithms have been projected by different scientists to classify the protein sequences. Researchers apply some well-known classification techniques like Naive bayes classification, neural networks, Genetic algorithm, Decision tree Classifier etc for accurate protein classifications.

The rest of the paper is organized as follows: Section 2 describes the basics of protein sequences and the need for classification of protein sequences into families. Section 3 describes the various research challenges related to classification of protein sequences. Section 4 deals with various methods developed by researchers for protein sequence classification and conclusion is given in section 5.

2. PROTEIN SEQUENCES:

Proteins are large molecules that are composed of one or more chains of amino acids in a specific order. The normal size of a protein molecule may be hundred amino acids, while the large proteins can have a thousand amino acids. 20 amino acids, i.e., {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}, make up the diverse array of proteins found in living things. Protein Sequence is considered as a very important part of biological sequence data where the analysis and study have become an important research direction and content in bioinformatics domain. This technique involves the determination of the amino acid sequence that makes up the protein. Some research can be performed on a protein sequence or a protein family sequence making the protein sequence pattern mining a much important task in this field. One of the basic tools for the analysis of proteins is similarity searching [11]. The process can be implemented in the amino acid sequence or in the spatial structure of the protein.

Searching for similarity among the proteins may have different applications. Based upon the area , the proteins can be analyzed at the level of amino acid sequence or with respect to various features of their structures. Comparative analysis of the protein sequences may be helpful for the identification of proteins, identification of their functions and determination of their fundamental physical-chemical properties. The comparative analysis of the protein sequences brings much more information and is extremely important in the processes such as the prediction of the properties of the newly discovered proteins that are difficult to be identified on the basis of amino acid sequence.

2.1 PROTEIN SEQUENCE DATABASES:

Researchers have found out that the sequences of various proteins and their details are stored in databases. Some well-known databases include Protein Data Bank (PDB), UniProt, Swiss-Prot, SCOP, etc. The sequence of the proteins can be extracted from these databases.

An example of protein sequence of Haemoglobin beta chain is shown below:

```
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQ
RFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLD
NLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
KEFTPPVQAAYQKVVAGVANALAHKYH
```

2.2 CLASSIFICATION IN PROTEIN SEQUENCES:

There is a need for the advancements in the development of an intelligent system in order to classify the incoming protein into a particular super family. Proteins are composed of a sequence of 20 different types of amino acids. The gene sequence is defined by the sequence of aminoacids which is encoded in the genetic code. During the synthesis of proteins, the residues in the proteins are often chemically altered. This is due to the posttranslational modification which may change the physical and chemical properties of the proteins and also alters their folding activity. Finally it changes the function of the proteins resulting in the disease conditions.

Two proteins can be classified into a same super family if they acquire similar sequence of amino acids, which may therefore be functionally and structurally related. If two protein sequences have high similarity i.e. have most of the features in common, then they can be classified in to the same class [11]. This plays important part in classifying new proteins based upon amino acid sequences, identifying diseased protein sequences, aligning multiple sequences, etc. The aim of protein super family classification is to predict the super family to which input protein belongs to. Protein classification aims on predicting the function or the structure of new proteins.

2.3 NEED FOR CLASSIFICATION TECHNIQUES IN PROTEIN SEQUENCES:

Different classification techniques have been used to classify the protein sequence into its particular class or sub class or family. Some of the features of the proteins can be extracted, then the value of these features can be matched and finally the protein sequence can be classified. In order to classify a data using classification or clustering techniques, a set of data is analyzed and a set of grouping rules are generated. For instance, one may classify diseases and provide the symptoms which describe each class or subclass.

In the analysis of biological data, an important problem is to classify the bio sequences or structures based on their critical features and functions. For instance, the protein and gene sequences that are isolated from the diseased and healthy tissues can be compared to identify the critical differences between the two classes of genes. Such features can be used for the classification of bio-data and prediction of the behaviors. For bio-data classification, several methods have been developed. For example, one can first retrieve the protein sequences from the two tissue classes and then find and compare the frequently occurring patterns of each class.

Analysis and interpretation of biological sequence data is a fundamental task in bioinformatics. The techniques such as classification and prediction are one way to deal with such task. In fact, the researchers are often involved in identifying the family to which a newly sequenced protein belongs. This helps in knowing the evolution of this protein and to discover its biological functions. In addition, the study of proteins and the prediction of proteins are very useful in biology and medicine for many reasons. Biologists also try to spot out the active sites in proteins and enzymes to classify parts of DNA or protein sequences into coding or non-coding zones or to determine the function of the nucleic sequences such as the identification of the promoter sites and the junction sites and the identification of position of mutation which results in diseased conditions [3].

3. RELATED WORK:

A wide range of classification techniques have been developed to classify the protein sequence into its particular class, sub class or family. All these methods have a tendency to extract some of the features of the proteins and then these feature values are compared and finally the protein sequence are classified.

Jason T. L. Wang, Qic Heng Ma, Dennis Shasha, Cathy H Wu In [6] had used a neural network model in their work for classifying the existing protein sequences. To implement this, some characteristics were extracted from the protein sequences to be used as an input. The high level features are extracted using some encoding techniques from the sequences considering both global and local similarity of the sequences. The methods that were used to find the global similarity are 2-gram encoding method and 6-letter exchange group. Certain user defined variables were used to achieve local similarity. Minimum description length (MDL) principle was also applied to calculate the significance of motif. Some features that were already defined were used as intermediate layers or hidden layers of the neural network. This model produces 90% to 92% accuracy [6].

Ramadevi Yellasiri, C.R.Rao [12] had proposed a new classification model called Rough Set Classifier for classifying the voluminous protein data based on structural and functional properties of protein. This model is fast and accurate and it can be used as an efficient classification tool than the others. This Classifier provides 97.7% accuracy. It is a hybridized tool comprising Sequence Arithmetic, Concept Lattice and Rough Set Theory. It can reduce the domain search space to 9% without losing the potentiality for the classification of proteins. The information about the family is identified using special arithmetic and utilize it for reducing the domain search space is proposed. The rules that are generated are stored in Sequence Arithmetic database.

Suprativ Saha and Rituparna Chaki [13] had proposed a three phase model for classifying the unknown proteins into known families. In first phase, the input dataset is reduced in order to remove the noisy sequences. It in turn will increase the accuracy and minimizes the computational time. In second phase, the necessary features such as molecular weight and isometric point are obtained and then feature ranking algorithm is applied to rank the features. This ranking is used to classify the sequences. In the final phase, the input sequence is classified in to the particular class or family using Neighbourhood Analysis. This rule has a power to extract the particular association relationships between the protein sequence and classes, sub classes and families. This type of classification produces the knowledge based information besides the data analysis technique.

Cornelia Caragea, Adrian Silvescu, Prasenjit Mitra [4] had used the feature hashing technique in their research work for the classification of protein sequences. Three issues are handled in this method. They are i) the influence of hash size on the performance of protein sequence classifiers that use hash features and the hash size at which the performance starts degrading due to hash collisions (ii) the effectiveness of feature hashing on prohibitively high dimensional representations (iii) the performance of feature hashing compared to that of feature selection by average mutual information. The results of the experiments on the protein data sets showed that feature hashing is an effective approach for reducing the dimensionality on protein sequence classification tasks.

Muhammad Mahbubur Rahman, Arif Ul Alam, Abdullah-Al-Mamun, Tamnun E Mursalin [9], had considered certain properties of protein for its classification namely structure comparison, Cluster Index, Connectivity, Interactivity, Taxonomic, etc. The probability of each protein that newly enters in the system is found against the already existing proteins in the system and they are added together. The aggregated probabilistic value is identified and it is compared with the probabilistic value of the protein that resides at the root. Based on guided search algorithm, the node which has the highest probability of level 1 is chosen. Then the probabilistic values of level 2 of selected node from level 1 are compared. Then the node having highest probability is chosen and continued until getting the exact position of newly entered protein reveals some of the functions based on the properties of proteins. These things are yet to be derived in different bioinformatics research lab such as cluster index, connectivity and interactivity. Then this can be efficiently used for protein classifications.

Xing-Ming Zhao1, De-Shuang Huang, Yiu-ming Cheung, Hong-qiang Wang and Xin Huang [15] had proposed a hybrid GA/SVM algorithm for protein sequence classification. The most important steps in the classification of protein sequences are the identification of the most informative features and reducing the dimensionality of the feature vectors. A novel hybrid GA/SVM system has been proposed that selects the features of proteins containing accurate and enough discriminative information for classification and train the SVM classifier simultaneously. Experimental results for protein sequences classification of six protein super families obtained from the Protein Information Resource (PIR) database showed that the hybrid GA/SVM system outperforms BLAST and HMMer techniques.

The above literature study clearly shows that till now, there is no technique that has been developed to classify the proteins based on the amino acid sequence of the proteins. Hence a new approach can be implemented for classifying the proteins based on sequences rather than considering the structural and functional characteristics.

4. RESEARCH CHALLENGES:

Recently, the assortment of biological data like protein sequences, DNA sequences etc. is growing at explosive rate due to the introduction of new ones. The results of recent research prove that it is very difficult to classify large amount of biological data like protein sequences using traditional database system. Data mining technique are suitable to handle the large data sources. A number of different techniques are already identified for classifying protein sequences. One important area of research is to classify protein sequences into different families, classes or sub classes. Some important research challenges include the following:

- The most important challenge in the area of protein sequence classification involves the organization of a huge volume of data such as the ones generated by the human genome project and to improve database design. For database access and manipulation and device data-entry procedures, several softwares need to be developed which in turn compensates for the varied computer procedures and systems used in different laboratories.
- In pattern matching, the entire protein sequence patterns cannot be matched accurately and the similarity among sequences are very difficult to identify. Hence computational work with data mining tasks must be developed which in turn minimizes the work of sequence comparison.
- Various classification methods are yet to be developed which provides easier way to classify the protein sequences based upon the user’s needs.
- Fuzzy modeling [13] helps in the sequence data analysis although storage and time requirement are high. But in the construction of fuzzy sets, the computational complexity is added in each iteration as well. Hence the fuzzy modeling can be enhanced to perform better.
- The protein sequence pattern mining process can be extended further which in turn determines the biological significance of the patterns. This will allow discriminating the weight of each pattern in the overall classification results [10].
- The researchers are preparing for a larger study to examine the notion that more extensive data mining studies of the protein database will yield both statistically significant and verifiable improvements within environments as well as meaningful comparisons across them.

- One of the most important technique for drug design and the design of novel enzymes is protein structure prediction. A general clarification to such predictions remains an open problem for the researchers [8].
- Another important research area in protein sequence classification is the usage of feature hashing technique to other types of biological sequence data, e.g., DNA data, and other tasks [4].

5. CONCLUSION:

In recent trends, analysis of large amount of biological data like protein sequences is very difficult using traditional database system. Data mining techniques can be used to classify the unknown protein sequence. But the different models, which are used to classify the protein sequence is not perfect regarding the both accuracy level and computational time. Data mining approaches prove to be efficiently well suited for the bioinformatics applications since the bioinformatics is data-rich but there is no comprehensive theory of life's organization at the molecular level. This paper provided an overview of the ongoing research works involving different techniques to classify the protein sequences. It has been observed that knowledge based and analysis of data form an integral part of protein sequence classification. The classification of protein sequence produces knowledge based information besides data analysis techniques. In future, different analysis has to be done for better performance of protein sequence classifications.

REFERENCES:

[1] Anita Zala, Mehul Barot," Mining Sequential Pattern with Time-Constraint", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 7, January 2013, ISSN: 2277-3754

[2] Arabinda Panda, "Bio-Data Mining:Concepts & Applications", International Journal of Computer Science and Management Research Conference Issue DPDM 2012, ISSN 2278-733x

[3] Christian Schaefer, Yana Bromberg, Dominik Achten, Burkhard Rost, "Disease-Related Mutations predicted to Impact Protein Function", From SNP-SIG 2011:Identification and Annotation of SNPs in the Context of Structure, Function and Disease Vienna, Austria, 15 July 2011.

[4] Cornelia Caragea, Adrian Silvescu, Prasenjit Mitra, "Protein Sequence Classification using Feature Hashing".

[5] Hui-Huang Hsu, "Introduction to Data Mining in Bioinformatics", Idea Group Publishing, ITB12936.

[6] Jason T. L. Wang, Qic Heng Ma, Dennis Shasha, Cathy H Wu, "Application of Neural Networks to Biological Data Mining: A case study in Protein Sequence Classification", pp: 305-309, KDD, Boston, MA, USA (2000).

[7] Jiawei Han, "How can Data Mining Help Bio-Data Analysis?", BIOKDD02: Workshop on Data Mining in Bioinformatics (with SIGKDD02 Conference)

[8] KJiawei Hanhalid Raza, "Application of data mining in Bioinformatics", Indian Journal of Computer Science and Engineering, Vol 1 No 2, 114-118.

[9] Muhammad Mahbubur Rahman, Arif Ul Alam, Abdullah-Al-Mamun, Tamnun E Mursalin, "A more appropriate protein classification using Data mining", Journal of Theoretical and Applied Information Technology.

[10] Padro Gabriel Ferreira, Paulo J.Azevedo, "Protein sequence classification through Relevant sequence Mining and Bayes Classifiers"

[11] Rabie Said, Mondher Maddouri , Engelbert Mephu Nguifo, "Protein sequences classification by means of feature extraction with substitution matrices"

[12] Ramadevi Yellasiri, C.R.Rao, "Rough Set Protein Classifier", Journal of Theoretical and Applied Information Technology, (2009).

[13] Suprativ Saha and Rituparna Chaki, “A Brief Review of Data Mining Application Involving Protein Sequence Classification”

[14] P.K.Vaishali, Dr.A.Vinayababu,” Application of Data mining and Soft Computing in Bioinformatics”, International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622

[15] Xing-Ming Zhao1, De-Shuang Huang, Yiu-ming Cheung, Hong-qiang Wang and Xin Huang, “A Novel Hybrid GA/SVM System for Protein Sequences Classification”

Authors

Dr. S.Vijayarani has completed MCA, M.Phil and Ph.D in Computer Science. She is working as Assistant Professor in School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of research interest are data mining, privacy, security, bioinformatics and data streams. She has published papers in the international journals and presented research papers in international and national conferences.

Ms. S.Deepa has completed M.Sc in Computer Science. She is currently pursuing her M.Phil in Computer Science in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of interest are Sequence pattern mining and privacy preserving data mining.