MANTECH Publications

Journal of Data Mining and Knowledge Engineering
Volume 6, Issue 2, July-December 2021

Interaction of Protein Sequence Classification with Data Mining

Aarti Shukla[1], Deepak Kumar[2]

Student[1], Professor[2]

Department of CSE

BBS College of Engineering and Technology

Corresponding Author's E-mail: *aarti.shukla004@gmail.com[1]*

Abstract

The employment of digital applications around the world has resulted in the collecting of massive amounts of information. The valuable information buried inside huge amounts of data is drawing scholars from a wide range of fields to conduct research into creating efficient techniques to extracting the hidden knowledge contained within them. When it comes to data mining, the process of extracting or mining meaningful and valuable knowledge from enormous volumes of data may be defined as follows: When it comes to data mining, there are many different domains to choose from, including text mining, picture mining, sequential pattern mining, web mining, and others. Sequence mining is one of the most significant study areas in this field, since it aids in the discovery of sequential relationships that have been discovered in data. It is used in a broad variety of applications, including the study of consumer purchase trends, online access patterns, weather observations, protein sequencing, DNA sequencing, and other types of data mining. For sequence alignment, sequence searching, and sequence classification in protein and DNA studies, sequence mining techniques are employed in conjunction with other approaches. Researchers in the field of protein sequence analysis are demonstrating an increased interest in the field of protein sequence categorization, according to their findings. It has the capability of identifying recurrent structures in protein sequences that are not previously known. This study describes the many methodologies employed by different researchers in the classification of proteins, as well as providing an overview of the numerous protein sequence classification methods available today.

Page 82-92 © MANTECH PUBLICATIONS 2021. All Rights Reserved

Keywords: *Data mining, Information, DNA, Protein sequence, Sequence mining*

INTRODUCTION

During the last several years, advances in the fields of genomics and proteomics have progressed at a rapid pace, resulting in the generation of a significant amount of biological information. It is necessary to use advanced computer analysis tools to draw inferences from this data in order to reach these findings. Informatics and computer science combine to generate the multidisciplinary field of bioinformatics, which is concerned with the interpretation of biological data. As huge volumes of genomic, proteomic, and other data continue to be collected and integrated, the relevance of this emerging field of study will expand. To enable the collection and storage of biological information, as well as the management and analysis, interpretation, and dissemination of that information, bioinformatics and computational biology are professions that need the acquisition of skills from a wide range of disciplines.

When dealing with massive amounts of genomic and proteomic data, it is necessary to make use of high-performance processors as well as cutting-edge software tools. The invention and deployment of unique algorithmic approaches, which are required for the analysis and interpretation of sequential data, are also included in this category. It is also utilised in the biological sciences for the categorization and prediction of sequential data, which gives insight into the design and validation of studies in the field of biomedicine. The significant growth in the volume of data collected on a daily basis necessitates the implementation of this measure right now.

It is the discipline of bioinformatics that is dedicated to the interpretation and analysis of biological data via the application of computer tools. Since there has been a rapid increase in the amount of biological information supplied by the scientific community, the area of bioinformatics has advanced to a larger level. A variety of data mining-related operations, such as the modification, searching, and data mining of DNA sequence data, are completed quickly and efficiently.

The development of tools for storing and searching DNA sequences has resulted in significant improvements in computer science that have been extensively

Page 82-92 © MANTECH PUBLICATIONS 2021. All Rights Reserved

implemented. There have been a number of approaches invented and successfully employed over time, such as string searching algorithms, machine learning, and database theory. The use of data mining techniques to biological issues is a particularly busy field of research in bioinformatics, and it is one in which the approaches are being used efficiently.

Both the data mining and bioinformatics areas are seeing rapid expansion of their respective study horizons. A critical component of bioinformatics research is the examination of major research questions as well as the development of novel data mining algorithms for scalable and effective bio-data processing. Data mining can aid researchers in discovering new information from a large amount of biological data collected at the molecular level [5, for example].

Serial pattern mining is an important field of data mining in which a very small alphabet (for DNA sequences, 4 letters, and 20 letters for protein sequences) and long patterns with a typical length of a few hundreds or even thousands of characters frequently appear. Sequential pattern mining is a subfield of data mining in which a very small alphabet (for DNA sequences, 4 letters) and long patterns with a typical length of a few hundreds or even thousands of characters frequently appear. A variety of different algorithms have been developed in order to effectively carry out the pattern mining process. It is the primary duty to identify all sequential patterns in a sequence database that have more or equal support to a predetermined minimum support threshold in order to do the search. Sequential data is regularly encountered in a wide range of scientific, medical, security, business, and other applications. When it comes to uncovering meaningful knowledge hidden in mountains of sequence data, sequence data mining provides the tools and methodologies that are required to do so. A sequence is often a list of things that is presented in an organised fashion.

Sequence in proteins refers to the organisation of amino acids in a certain order that results in the formation of the protein. The sequences of proteins that have been discovered are being maintained in a number of different databases. Each database has a distinct categorization of protein sequences that is unique to that database. There have been a variety of classification approaches or algorithms proposed by different academics in order to categorise distinct protein sequences. For reliable protein classifications,

Page 82-92 © MANTECH PUBLICATIONS 2021. All Rights Reserved

researchers use well-known classification approaches such as Naive Bayes classification, neural networks, genetic algorithms, decision tree classifiers, and other similar techniques, among others.

In the next sections, you will find the rest of the paper: Section 2 discusses the fundamentals of protein sequences, as well as the necessity of categorising protein sequences into groups called families. Section 3 discusses the different research obstacles that have been identified in the field of protein sequence categorization. Section 4 discusses the numerous approaches for protein sequence categorization that have been established by researchers, and Section 5 concludes the discussion.

Protein Sequences:

Proteins are huge molecules that are made up of one or more chains of amino acids that are arranged in a precise order on the molecule. The usual size of a protein molecule may be a few hundred amino acids, but big proteins may have several thousand amino acids or more. The amino acids A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y combine to form the vast variety of proteins found in living things. A total of 20 amino acids are present in all living things.

When it comes to biological sequence data, protein sequences are regarded to be a highly essential component, and their analysis and study have become a significant research area and topic in the bioinformatics sector. This approach entails determining the amino acid sequence that is used to construct the protein. It is possible to do some study on a protein sequence or on sequences belonging to a protein family, which makes protein sequence pattern mining a critically significant work in this discipline.

When it comes to protein analysis, similarity searching is one of the most important techniques. Depending on the amino acid sequence used, or the spatial structure of a protein, this method can be accomplished.

Similarity searches among proteins may have a variety of uses depending on their structure. Because of this, proteins may be studied at the amino acid sequence level or with respect to numerous structural aspects of their structures, depending on the region. The comparative study of protein sequences may be useful in the identification of proteins, the identification of their activities, and the discovery of their fundamental physical-chemical characteristics, among other applications.

Page 82-92 © MANTECH PUBLICATIONS 2021. All Rights Reserved

It is particularly significant in procedures such as the prediction of the attributes of newly found proteins that are difficult to identify just on the basis of amino acid sequence to conduct a comparative study of protein sequences since it provides considerably more information.

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:

VHLTPEEKSAVTALWGKVNVDEVGG
EALGRLLVVYPWTQ
RFFESFGDLSTPDAVMGNPKVKAHG
KKVLGAFSDGLAHLD
NLKGTFATLSELHCDKLHVDPENFRL
LGNVLVCVLAHHFG
KEFTPPVQAAYQKVVAGVANALAHK
YH

Classification in Protein Sequences:

It is necessary to make improvements in the construction of an intelligent system in order to categorise incoming proteins into a certain super family, and this can only be accomplished through collaboration. Proteins are made up of a sequence of amino acids that contain 20 distinct types of amino acids. As described by the genetic code, the amino acid sequence defines the sequence of amino acids present in a given gene. During the production of proteins, the residues in the proteins are frequently subjected to chemical modification. This is due to the posttranslational alteration of proteins, which may modify their physical and chemical characteristics as well as their folding activity, among other things. Finally, it alters the function of proteins, which results in the development of illness states.

Two proteins can be categorised into the same super family if they acquire a sequence of amino acids that is comparable to one another. This indicates that they are functionally and structurally connected to one another. When two protein sequences show a high degree of similarity, that is, when they have the majority of the same characteristics, they can be categorised into the same class. A significant role is played in the classification of novel proteins based on amino acid sequences, the identification of pathogenic protein sequences, the

Page 82-92 © MANTECH PUBLICATIONS 2021. All Rights Reserved

alignment of numerous sequences, and other applications. The purpose of protein super family classification is to determine which super family an input protein belongs to based on its sequence. Protein categorization seeks to anticipate the function or the structure of novel proteins by analysing their sequence and structure.

Need For Classification Techniques In Protein Sequences:

Different classification strategies have been used to the protein sequence in order to categorise it into a certain class, subclass, or family of proteins. Several proteins' characteristics may be retrieved, and the values of these characteristics can be matched, leading to the classification of a protein's sequence in the final step. Classification or clustering techniques are used to categorise data. A set of data is studied and a set of grouping rules is developed in order to classify data. For example, one may categorise diseases and then list the symptoms that correspond to each class or subclass of disorders.

One of the most challenging problems in the study of biological data is the classification of bio sequences or structures based on their key characteristics and functions. The protein and gene sequences that have been recovered from sick and healthy tissues, for example, can be compared in order to find the significant changes between the two classes of genes. Such characteristics can be employed for the categorization of bio-data as well as the prediction of behavioural patterns. A number of classification approaches have been developed for bio-data categorization. Example: To locate and compare commonly recurring patterns in each class, one may first get the protein sequences from both tissue classes and then find and compare the protein sequences from both tissue classes.

In bioinformatics, the analysis and interpretation of biological sequence data is a vital activity that must be completed. It is possible to deal with this job using techniques such as classification and prediction. In fact, when a fresh sequenced protein is identified, the researchers are frequently engaged in determining which family it belongs to. This information is useful in understanding the evolution of this protein as well as discovering its biological roles. There are a plethora of reasons why the study of proteins and the prediction of proteins are extremely valuable in biology and medicine. Additionally, biologists attempt to identify active sites in proteins and enzymes in

Page 82-92 © MANTECH PUBLICATIONS 2021. All Rights Reserved

order to classify parts of DNA or protein sequences into coding or non-coding zones, or to determine the function of nucleic sequences, such as the identification of promoter sites and junction sites, as well as the identification of the location of mutations that result in diseased conditions.

RELATED WORK:

A diverse variety of categorization approaches have been developed to group protein sequences into certain classes, subclasses, and families based on their amino acid sequences. All of these approaches have a propensity to extract some of the features of the proteins, and then the values of these features are compared, and lastly the protein sequences are classed according to their classification.

Ramadevi Yellasiri and C.R.Rao established a novel classification model called Rough Set Classifier for categorising massive protein data based on structural and functional features of proteins. This model is faster and more accurate than the others, and it may be used as a more efficient categorization tool. This Classifier has an accuracy of 97.7 percent. It is a hybrid tool that combines Sequence Arithmetic, Concept Lattice, and Rough Set Theory. It can minimise the domain search space by 9 percent while retaining the ability to classify proteins. The information about the family is recognised using special arithmetic and is offered for use in lowering the domain search space. The created rules are stored in the Sequence Arithmetic database.

Suprativ Saha and Rituparna Chaki presented a three-phase methodology for categorising novel proteins into recognised families. The input dataset is reduced in the first step to eliminate the noisy sequences. As a result, the accuracy improves while the computing time is reduced. In the second step, the relevant characteristics such as molecular weight and isometric point are acquired, and then a feature ranking algorithm is used to rank the features. This rating is used to categorise the sequences. In the last phase, the input sequence is categorised into a certain class or family using Neighbourhood Analysis. This rule may extract the specific connection associations between the protein sequence and classes, subclasses, and families. In addition to the data analysis approach, this sort of categorization generates knowledge-based information.

Cornelia Caragea, Adrian Silvescu, and Prasenjit Mitra employed the feature hashing approach to classify protein

Page 82-92 © MANTECH PUBLICATIONS 2021. All Rights Reserved

sequences in their research. This solution addresses three difficulties. They are as follows: I the effect of hash size on the performance of protein sequence classifiers that use hash features, as well as the hash size at which performance begins to degrade due to hash collisions; ii) the effectiveness of feature hashing on prohibitively high dimensional representations; and iii) the performance of feature hashing versus that of feature selection by average mutual information. Experiments on protein data sets revealed that feature hashing is an efficient method for decreasing dimensionality in protein sequence classification tasks.

Jason T. L. Wang, Qic Heng Ma, Dennis Shasha, and Cathy H Wu are among those who have contributed to this work. In their study, employed a neural network model to categorise known protein sequences. To do so, various properties were retrieved from protein sequences and utilised as input. The high level characteristics are retrieved from the sequences utilising several encoding approaches that take into account both global and local similarity. The 2-gram encoding technique and the 6-letter exchange group were used to determine global similarity. To achieve local similarity, some user-defined variables were employed. The minimum description length (MDL) method was also used to determine motif importance. Some previously determined characteristics were employed as intermediate or hidden layers of the neural network. This model achieves 90 to 92 percent accuracy.

Structure comparison, Cluster Index, Connectivity, Interactivity, Taxonomic, and other features of proteins were evaluated by Muhammad Mahbubur Rahman, Arif Ul Alam, Abdullah-Al-Mamun, and Tamnun E Mursalin for protein categorization. The likelihood of each new protein entering the system is compared to the probability of the previously existing proteins in the system, and they are put together. The aggregated probabilistic value is identified and compared to the probabilistic value of the root-residing protein. The node with the highest likelihood of being at level 1 is picked using the guided search method.

The probability values of level 2 of the selected node from level 1 are then compared. Then the node with the highest probability is picked, and the process is repeated until the precise position of the newly inputted protein discloses some of the functions based on protein attributes. These things, such as cluster index, connection, and interaction, have yet to be

Page 82-92 © MANTECH PUBLICATIONS 2021. All Rights Reserved

discovered in various bioinformatics research labs. This may then be used to efficiently classify proteins. The preceding literature review clearly demonstrates that no approach has been devised to categorise proteins based on their amino acid sequence. As a result, rather than evaluating structural and functional properties, a novel method to protein classification based on sequences may be adopted.

RESEARCH CHALLENGES:

Since the advent of new biological data, such as protein sequences, DNA sequences, and so on, the amount of biological data has grown exponentially. Recent studies have shown that standard database systems fail miserably when it comes to classifying enormous amounts of biological data, such as protein sequences. Large data sets are well-suited for data mining techniques. Numerous methods for categorising protein sequences have previously been discovered. The classification of protein sequences into families, classes, or subclasses is an important research subject. The following are some of the most pressing research issues:

- The most significant problem in protein sequence categorization is the management of a large volume of data, such as that created by the human genome project, and the improvement of database architecture. Various laboratories utilise a variety of computers, therefore different software must be built to accommodate the different computer techniques and systems in use.

Protein sequence patterns cannot be correctly matched and the similarity between sequences is extremely difficult to discern in pattern matching. Data mining jobs must be designed in order to reduce the amount of time spent on sequence comparison.

In the near future, new approaches for classifying protein sequences will be developed that are more user-friendly.

Even though it requires a lot of storage space and processing time, fuzzy modelling can be used to analyse sequence data. Fuzzy sets, on the other hand, increase in computational complexity with each iteration. Consequently, the fuzzy modelling may be improved to work more efficiently.

- The process of protein sequence pattern mining may be extended

Page 82-92 © MANTECH PUBLICATIONS 2021. All Rights Reserved

further, which in turn determines the biological importance of the patterns.

- The researchers are preparing for a bigger investigation to explore the concept that more thorough data mining studies of the protein database will generate both statistically significant and verifiable improvements within settings as well as relevant comparisons across them.

- Protein structure prediction is a crucial tool for drug development and the development of new enzymes. The researchers are still working on a broad explanation for such predictions.

- The application of feature hashing techniques to other forms of biological sequence data, such as DNA data, and other jobs, is another major study topic in protein sequence classification.

CONCLUSION

In recent years, conventional database systems have found it challenging to analyse enormous amounts of biological data such as protein sequences. The unknown protein sequence can be classified using data mining techniques. However, the many models used to categorise protein sequences are not perfect in terms of both accuracy and computing time. Because bioinformatics is data-rich, yet there is no complete theory of life's structure at the molecular level, data mining methodologies appear to be well suited for bioinformatics applications.

This publication gave an overview of current research projects that use various strategies to categorise protein sequences. Protein sequence categorization has been reported to include knowledge-based and data-analysis components. In addition to data analysis tools, protein sequence classification generates knowledge-based information. Different analyses will be required in the future to improve the effectiveness of protein sequence classifications.

REFERENCES

1. K Jiawei Hanhalid Raza, ”Application of data mining in Bioinformatics”, Indian Journal of Computer
2. Science and Engineering, Vol 1 No 2, 114-118.
3. 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

Page 82-92 © MANTECH PUBLICATIONS 2021. All Rights Reserved

Classification", pp: 305-309, KDD, Boston, MA, USA (2000).

4. Arabinda Panda, "Bio-Data Mining:Concepts & Applications", International Journal of Computer
5. Science and Management Research Conference Issue DPDM 2012, ISSN 2278-733x
6. Cornelia Caragea, Adrian Silvescu, Prasenjit Mitra, "Protein Sequence Classification using Feature Hashing".
7. 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.
8. Anita Zala, Mehul Barot," Mining Sequential Pattern with Time-Constraint", International Journal of
9. Engineering and Innovative Technology (IJEIT) Volume 2, Issue 7, January 2013, ISSN: 2277-3754
10. Hui-Huang Hsu, "Introduction to Data Mining in Bioinformatics", Idea Group Publishing, ITB12936.
11. 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
12. Jiawei Han, "How can Data Mining Help Bio-Data Analysis?", BIOKDD02: Workshop on Data
13. Mining in Bioinformatics (with SIGKDD02 Conference)
14. Suprativ Saha and Rituparna Chaki, "A Brief Review of Data Mining Application Involving Protein Sequence Classification"

Page 82-92 © MANTECH PUBLICATIONS 2021. All Rights Reserved