ISSN 2278-3091
Volume 8, No.5, September - October 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse67852019.pdf
https://doi.org/10.30534/ijatcse/2019/67852019

Food Waste Protein Sequence Analysis using Clustering and Classification Techniques

U. Vignesh[1], G. Sivanageswara Rao[2], B. Manjula Josephine[3], Puvvada Nagesh[4]
[1,2,3,4] KL University, India, vigneshbun@gmail.com[1], sivanags@kluniversity.in, manjulajosephine@gmail.com, pnagesh@kluniversity.in

ABSTRACT

The rapid development of data mining techniques takes place in 1980. A large number of newly invented technologies came into the computer field such as satellites, high storage of data medium, etc. This improvement paved the way for a destination in one place for the huge collected data. There were many methods followed earlier for the analysis of data but they failed to prove efficiency in higher gathered data and also in noisy data compared to data mining. This paper clearly describes about the various biological data mining techniques and their inter relations for prediction with the basic concepts of clustering, classification and alignment techniques. This provides the characteristics of protein sequence analysis in motif finding.

Key words: motif, prediction, protein sequence, data mining

1. INTRODUCTION

Data mining is defined as useful information from the database also said to be KDD (Knowledge Discovery in Databases). It is the process of extracting meaningful information from a large database. One can get useful patterns to form the database using the techniques of data mining which are efficient and accurate. This data mining has a very large scope to grow and which in hand can be used for society. The use of the database is required for any sector whether it is a government or non-government organization. Considering the example of DNA bank sectors, the data of each and every individual should be maintained properly. Efficient data storage is done through a database where a KDD process can be applied for the retrieval of data from it. The KDD process deals with the different stages of data access as shown in Figure 1.

1.1 Biological Data Mining and Motif

Data mining is an interdisciplinary area which includes many domain areas interconnected into it such as prediction, information retrieval, database systems, statistical calculations, and machine learning concepts. The applications of data mining are 'n' in numbers which are uncountable and unpredictable, some of them are medicine, biology, pharmacology, etc. In biology, the microarray concepts have to lead to an increase in new databases and different strategies involved in management systems. Thus, the gen bank such as NIH (National Institute of Health), NCBI (National Centre of Biotechnology Information) started in 2004, etc. have become thousands in numbers and increased substantially compared with the past twenty years. These prospects make the bioinformatics as an important field integrated with computer science as shown in Figure 2.

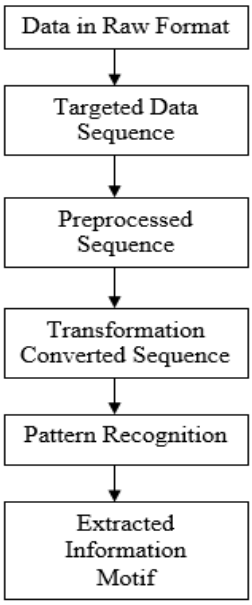

Figure 1: Knowledge Discovery Database Model

Bioinformatics allows researchers and inventors to refer the existing data and submit new data into the database entries. It provides a way to develop biological software analysis tools, which in turn results in a new and varied biological insights for reference. There are many applications directly connected with the bioinformatics in real time such as agriculture, the farming and organized study of biomaterials, etc.

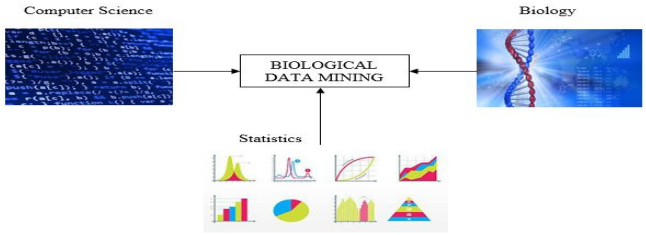

Figure 2: Components of Biological Data Mining

Data mining deals with the action of data fusion in the majority, enabling different types of data to be organized in a consistent manner and efficiently with accurate information obtained from the single source or multiple data source. Sampling also has its part in the data mining domain. In the process of sample analysis in the digital form if whether it is not in the digital format then the conversion process takes place. The analysis can also be provided using multi-resolution processes for finding the occurrence of redundancy in data from the database. In the pre-processing aspect, the concept of de-noising which removes the noise present in the data and access efficiently. Since data mining is the extraction process, it starts with the extraction of feature information.

The research on biological data mining provides a deep understanding of how the data mining techniques can be applied on biological data, which is enormous in amount made into digital format since the 1970s to maintain database efficiently and accurately. Retrieval of the data from public databases such as NCBI, PDB (Protein Data Bank), etc. is done by these data mining algorithms providing a major contribution since the retrieved data information should be meaningful information said to be a biological data mining. Data mining concepts provide very useful data without any redundancy containing meaningful information. It can also be able to represent all collected data stored in a group for easier reference. In the pre-processing stage of data mining, dimensionality reduction has its part to avoid or reduce the number of the available random attributes by selecting the feature provided by a set of principal design decisions. The gene analyses application of data mining is a very much challengeable one since the human body is made of more than 100, 000 genes. Genes are the acting units that have their place in the chromosome where the mutations are permitted. Genes contain the data in the form of nucleotides and nucleus. Nucleotides containing the phosphate group builds the group in the structured format known as DNA (Deoxyribose Nucleic Acid).

The nucleus is the genetic material that occurs in the eukaryotic part represented as center part in the symbol of the sphere which looks like a double membrane. Data mining proves its major contribution to the banking sector continuously for many years in their financial related data collection maintenance and retrieval.

The banking sector also widely uses classification techniques and clustering techniques of data mining for the management of its customer related activity which includes marketing for their individual services, etc. Data mining also helps banks to identify fraudulent processes seen in performing the currency oriented crimes. Retail state sector also has its application in data mining through which the customer can buy anything he needs. The customer provides his individual accessing data relating to the consistent buying items so that in future the industry recommends some offers to ensure customer maintenance. Telecommunication industry is a vast ocean where the data are stored, retrieved and perform numerous activities of a broad spectrum helping for the government organizations in resource allocation to various sectors for various perceptions and also for the private organization to provide the communication services in the various phenomenon of mobile phones, broadband, etc.

Knowledge discovery process deals with precise data and uncertainty data. The uncertainty data includes the various data extraction methods as shown in **Figure 3.** Data integration concept handled in the data mining during the pre-processing stage searches for the various data from different sources and gets them integrated into a single form for delivering the information in a useful manner. Data cleaning is the process of detecting and correcting (or removing) records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the coarse data. Data selection is defined as the process of determining the appropriate data type and source, as well as suitable instruments to collect data. During the extraction process, pattern evaluation is done based on performance measures such as score, time, etc.

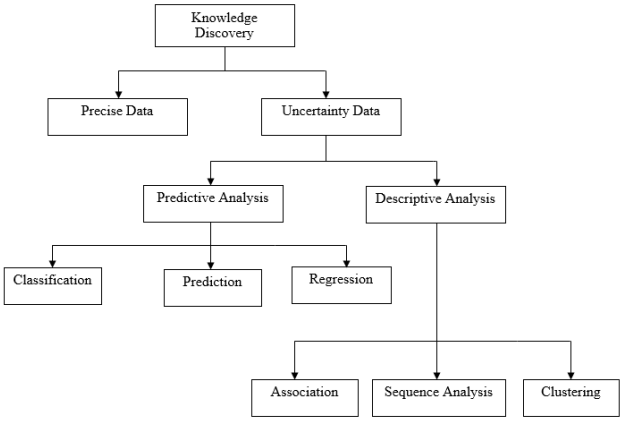

Figure 3: Data Analysis Techniques

Visualization phase of the data mining process facilitates data extraction in an easier manner to be understood by the user. It includes the interpretation process for results and provides results with accurate scalability. Data mining deals with four major techniques or tasks, namely classification, clustering, regression, and association. Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups (clusters). Classification deals with the format of already identified groups for undergoing to the new referential data. Classification can be performed on all types of data, such as categorical, numerical, etc.

1.2 Clustering Technique

Clustering defined as grouping of similar data into clusters is very much useful in sequence analysis in areas like pattern matching and sequence pattern mining. This research work concentrates on five different real-time environmental food samples, through which microorganisms are to be identified. Among these identified microorganisms, the one has its majority in all the samples taken are to be considered for further analysis and the designing processes. Since

microorganisms plays a vital role in the human body for the diseases, the analysis on micro organisms gives some knowledge to help drug discovery.

The microorganisms are studied using a metagenomics sequencing , which gives a detailed report of the microbiomes in the biological data considered. Cell culture is done on the basis of endothelial behaviour during the culture process and the process of growth depends on the electrical charge used for phospholipid surfaces. A study was done on the matches between the big data era and recent data green revolution concentrating on big data technologies for analysis of the new generation data. This work getting an idea on data variations, data size available on the database for collecting the inputs for the prediction process.

This research work uses data mining techniques for the analysis of structured and unstructured proteins. Target sequences from BLAST (Basic Local Alignment Search Technique) are used for the prediction of the protein structure. The protein structures of Macropus Rufus and structured proteolyzed lyzome came into the picture through merge set filter for grouping of data(unsupervised learning). The main problem is the clustering technique selection for protein clustering. The performance of the clustering algorithms namely, SimpleKMeans, Expectation Maximization and MakeDensityBasedClusterer are analyzed. This simulation is carried out on open source data mining tool Bioweka. The efficiency of these algorithms is seen through comparative analysis.

The process described can be used for wide range of applications such as medicine discovery, protein analysis, docking, etc. The information extracted from proteins is a collection of molecules used in the formation of the mass of living beings are considered as inputs. This extracted information is useful for drug discovery which is one of the major applications of data mining. Searching for pattern and retrieving it from database is called retrieval. The process of retrieving the information from two structured proteins is done by using data mining supervised and unsupervised learning techniques.

1.3 Alignment Technique for Motif

The first step in sequence analysis is the collection of biological data irrespective of DNA or protein in the alignment process for matching the target with another one to find an ancestor model. Then, other processes followed by other bioinformatics methodologies such as motifs, patterns mining, etc. Comparative genomics involves analysis of the gene prospect with the warehouse provided and identify their mutual and distinct perceptions in the gene expression on data and results. There are many algorithms used in the computer science field for solving the comparative genomics problem of biological data such as BLAST, FASTA (first FAST search sequencing Algorithm), Needleman Wunsch algorithm, Smith Waterman algorithm, etc. These changes and match identification resulting in the priority bases which pave the way for the reliable results in the cancer cells to pick their mutated part in the gene expression of the unaffected cells. All these mutations are performed with only the changes in the base pairs whether it is particular sequence orientation of genomics or proteomics i.e. prokaryotic or eukaryotic.

1.4 Classification Technique

Classification technique plays a major role in bioinformatics for the coded region analysis. Neural network approach of classification first does the division of finding unknown protein sequences to start processing them. Then it is followed by finding any common similarity between them based on the measures. Available calculations such as MDL (Minimum Description Length) can be used for assessment of the performance of motif finding.

In a neural network, the probabilistic model can also be built taking the input layer as the target, the similarities found commonly are noted as the hidden layer and the process outcome in the output layer. There are many methods available for evaluation of performance such as encoding methods which do a calculation based on the grams available whether it is 2 or 4, etc. The protein sequences can also be classified using the fuzzy approach in which it gives the accurate value with the already existing protein sequences in the databases.

Statistical calculation of the target sequence includes molecular weight, amino acid count, isoelectric point, etc. Once these several features are identified from the sequences, they are given as input to the fuzzy model. Fuzzy model, in turn, results in a matching of the super families or sub families of the provided feature with accuracy taken into maximal consideration. The fuzzy model needs the membership function and CPU (Central Processing Unit) time calculation for effective output. Though the neural network and fuzzy model are accurate, they are not suitable for large databases. They can handle data with fewer features. A set of rules has to be derived with the machine learning concepts named as decision trees for handling a large set of features.

Decision trees maintain the potentiality of the protein sequences. The rules framed for classification needs can be put in the biological databases. Decision trees do not provide the result in the format of families. Despite this, the results show the smaller known sequences from the provided large set of sequences. There are many numbers of public biological databases making use of this rule-based classifier, such as the most common NCBI database also use this kind of approach. In this rule-based classifier, analysis was done on neighborhood also provides the result in an efficient manner.

The problem in the rule-based classifier is that it takes more time compared to the neural network and fuzzy approach. It does not provide any analytical form of the result but only the smaller sequences result. It handles the physical relationship rather than focussing on feature extraction. Whereas, neural network and fuzzy approach do not deal with the physical connection, but take their chemical relationships between the

sequences. Another type of strategy used for classifying protein sequences is the word segmentation method.

NLP(Natural Language Processing) method which deals with the meaningful aspect rather than physic – chemo relationships. Initially, it develops a book containing the content amino acids from all the sequences and then perform the segmentation to identify the matching perceptions in the book. Finally, it considers the results obtained from segmentation, generates the segment feature vectors from the sequences as shown in Figure 4.

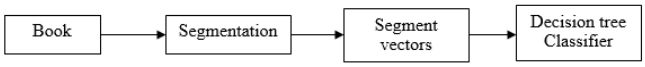

Figure 4: Segment Vector Classifier Model

The meaningful amino acids following the methodology of maximum length available in the sequences are found for framing the book with amino acids and fixing a threshold based on availability. It uses TF-IDF (Term Frequency-Inverse Document Frequency in the training set using the **Equation 1** and **Equation 2** with 'O' mentioning the weight, 'K' refers to the frequency, 'N' regards to the training set size, 'z' is the present sequences and 'N' is the number of sequences.

$$O_t = max(z * N)O_(t, z) \quad (1.1)$$

$$O_(t, z) = K_(t, z)logN/S_t \quad (1.2)$$

HMM(Hidden Markov Model)also proves its efficiency in the computational biology. It uses the methodology of emitting a single monomer from the fixed free sequences. Before starting the HMM, it needs to establish the start codon and stop codon of the sequence.

1.5 Regression Technique

Regression is a technique relating to performance evaluation aspect. The major focus of regression is the error rate calculation for finding the least error rate in modeling useful data without integration aspects. The most familiar regression method is the linear regression, which deals with the straight line equation mentioned in Equation 3**,** where it predicts the z value based on x value.

$$z = cx + d \quad (1.3)$$

The model of multiple regression is also performed with an increased number of input variables, such as the quadratic equation. Association technique is also used in the market basket analysis. It finds the connections among the variables provided with some predefined rules predefined. The architecture of data mining is described in Figure 5. For framing the data mining algorithm, there are four components to be considered namely, structure or pattern, score or rank, a method for finding and finally the strategy used for managing the data provided.

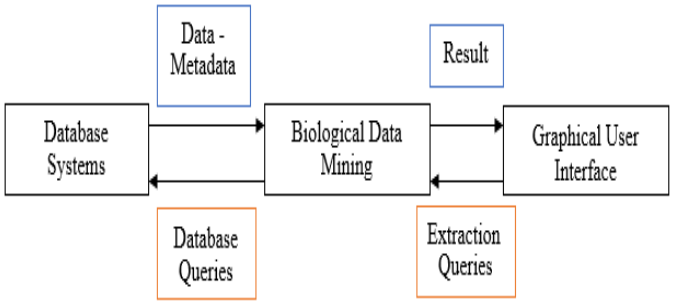

Figure 5: Biological Data Mining Architecture

1.6 DNA and Amino Acids Data

Data mining processes find wide use in education-related management systems with challenges involving quality, privacy and data streaming. Security is also one of the major challenges in data mining. Ford company, Sony company, etc. are examples where it has failed to ensure data are not hacked and accessed by third parties. Though there are many applications for data mining, bioinformatics being an ocean has its major impact on several of its areas. Sequence analysis is the primary action in the computation process of biology. There are three types of biological sequences involved such as DNA, RNA (Ribose Nucleic Acid) and protein. DNA is comprised of four monomer bases such as A (Adenine), T(Thymine), G(Gyanine) and C (Cytosine). RNA is divided into mRNA (messengerRNA), rRNA (ribosomalRNA) and tRNA (transferRNA).

The conversion of DNA sequence to a protein sequence synthesis information done in mRNA. The action of protein synthesis is seen in a rRNA. tRNA is the reverse translation of the protein parameters to the cellular components of ribosomes. DNA is always known for its replication or multiplication in enormous concerned bases. RNA also multiplies itself but there are some proof of its transfer to DNA such as retroviruses. Figure 6 describes the major processes of molecular biology for the gene information handling. RNA is the most identified genetic material in AIDS (Acquired Immune Deficiency Syndrome) causing disease.

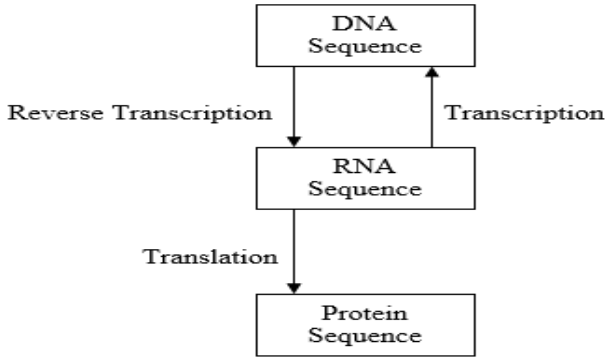

Figure 6: DNA - RNA – Protein

The structure of protein sequences is divided into four types namely, primary, secondary, tertiary and quaternary. Protein sequence acts as an input for the proposed work which is composed of twenty alphabets. Protein sequences are made up of twenty protein monomers known as amino acids. Primary protein structure is an amino acid sequence which is in a form of a chain like structure called polypeptide. Secondary protein structure is also a form of a polypeptide but it includes the data after the folding takes place in protein sequences. Tertiary protein structure is a 3D (Three Dimensional) representation of a polypeptide form. Quaternary protein structure is also a 3D form but includes the instance of more than one polypeptide focussing on the same point. Protein-protein interaction takes place in the quaternary structure. The conversion of DNA to protein sequences through RNA has the standard genetic code to be followed, involving three nucleotides to single amino acid. Since there are twenty amino acids constituting four nucleotides with the possibility of one or two nucleotides in all the available combinations, the codon does not have the ability to satisfy twenty amino acids count. Hence there is a need to take three in the count for nucleotides in the conversion process as shown in Figure 7.

The consideration of DNA and proteins in the biological molecules arise due to their greater size in the computation not as such of other molecules. These two molecules are referred to as the macromolecules. The sequences are said to be the queue of monomers and identified with the symbol of any string with the respect to one monomer each. The term genomics refers to genetic information of organisms providing new biological information. Gene expression analysis is the action to translate the gene's information which is coded into structural and functional groups of cellular components. The structural and functional study of protein monomers on a huge scale is referred to as proteomics. The most identified problem in the bioinformatics is the identification of sequence in the DNA which performs its action with different roles in biology. There are two types of such sequences named as transcription start site and translation start site for the gene information study. Thus, mRNA start producing at this point in the DNA. Microarray experiment fully deals with the mRNA and their relationship activity. The samples at different instances of time in a discrete way provide the correlation in computational biology.

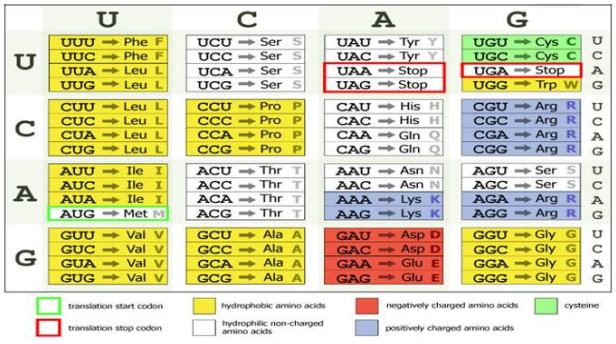

Figure 7: Standard Genetic Code for DNA-RNA-Protein model

2. PROTEIN SEQUENCE ANALYSIS

In the case of DNA, there are sequences which have specific characteristics apart from several extraction techniques for these sequences such as clustering, SVM (Support Vector Machine), neural networks, etc.

Traditional decision designs are not suitable for the specific inbuilt DNA sequences since it requires the existing methodologies to analyze their properties in detail. Thus, genomics cannot be analyzed or designed or not even studied without the use of data mining techniques.

Gene selection can be done in gene expression analysis, in the identification of the respective gene which is connected to the target class or a class provided for finding. In microarray perception, clustering and classification have the ability to predict or analyze the disease or the output related data efficiently with their split up as shown in Figure 8.

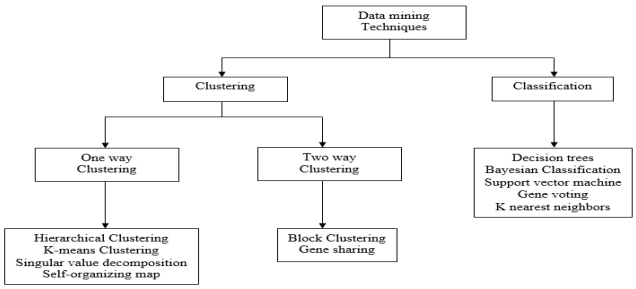

Figure 8: Categories of Data Mining Techniques

In the analysis of proteomics, modification site finding provides detailed answers but without the searching method of data mining such as neural network, etc. its efficiency is not assured. The major theorems of bioinformatics relate to storage, retrieval, and transformation. The simulation processes involved in the biological data are referred to as computational biology.

Data mining in the bioinformatics is used to do the inference of particular need of the structure or analyze the generalization provided in the data such as how two proteins interact, predicting the structure of the protein, classifying what kind of gene it belongs to, etc. The term bioinformatics was first generated by Paulin hogeweg in 1979, for the purpose of analysis, design and visualization aspect of biological data. A concept which is included in this research work for the effective protein sequences. Many problems are found in the biological field with respect to the computer science area. Some of them are interpretation and promoter finding in a biological sequence, motif identification, molecular structure prediction, protein-protein interaction, etc. Many tools have been invented to solve this problem with the machine executable code and algorithms.

There are research tools which follow the traditional way of finding organisms and pin pointing their variations in the gene or finding the diseases into it. Some of the biological computational tools which are useful for small data kind are

mentioned, such as genespring, Affymetrix, Spotfire, etc. In DNA, the nucleotides are ordered in a different manner with respect to the species or organisms referred but all are comprised of single strand double bonded mechanisms. The base pairs in the DNA play an important role in the gene expression. Even a single base pair of either A, T, G, C changes or mutated with one another, the whole genome mechanisms changes and the specific species identification also changes. The process does not provide exact results expected for the input. Most of the biological labs avoid the traditional methods and follow the new concepts of NGS (Next Generation Sequencing).

2.1 Characteristics of the motif in protein sequences

Once a mutation happens, the cells in the genes start the growth resulting in their respective change in the genome. In the protein sequences, motif finding problems are connected with the evolution of convergence. Motifs picking the patterns in repetition can be found and it is accepted with the mutation factor on sequences. Bioinformatics can be divided into two major categories of problem-solving such as prediction and pattern recognition. In both areas, motifs have its own role to provide justification for the solving methodologies. Prediction takes place when there is a sequence given but the structure of the given sequence is unknown. So, the problem-solving methodology prediction comes to predict the structure of the sequence.

2.2 Pattern Recognition

Pattern recognition deals with the set of sequences and target sequences are provided. The structure of the target sequence is known. The problem lies in finding the association with the target and sequences in a warehouse based on selection of the pattern recognition technique. These recognition prospects deal with the amino acids properties including the size, concerned charge and also the alterations made in spite of the existing one since the motifs allow to do it. The residues of the sequences are considered important for predicting their characteristics functions and particular structure. Motifs are highly sensitive to the performance analysis calculation involving sensitivity to include all the true positives. It also does specificity to avoid false positives rather than concentrating on the negative formats as such of other biological methodologies does. Prediction takes place following this. It initially identifies the target features found in the group of sequences provided in the warehouse. These activities which are not available in the target taken into account such as T-cells are most preferred in the steps since it does most of the activity in the protein sequences. The similarities and differences among Protein structures are based on homology and alignment. Alignment provides the positional homology between the given bases.

2.3 Prediction

In prediction, this research work contributes a better resolution functional part of the unknown organism. Healthcare in data mining has its advantages as the life science research or contribution directly points towards the health analysis. It is useful for predicting the age, weight, etc. of the living organisms. In metabolomics, data mining is useful for detecting the disease, overcoming its complexity in the metabolic data through the techniques of feature extraction, dimensionality reduction, etc. Sequence analysis results can also be obtained through efficient vector-valued functions, such as CGR (Chaos Game Representation), etc. The protein structure prediction from real time environmental samples consists of following steps, such as medium preparation, organisms' identification, sequencing, ARCSA algorithm, Protein protein interaction and specific organism structure prediction in real time.

CGR is used to identify letters in the alphabetical order, which has a higher order. It process the protein sequences with higher order amino acids. It takes on all their activity in the binary representations and provides the result in the form of statistical significance with the description of their properties such as standard Euclidean, linear correlation coefficient, etc. The contrasting length small sequences are found using this approach, which is not able to practice on the alignment methods. It takes the genomic signature in their work based on the data already available on oligonucleotide through which it is able to find the two sequences showing the same signature of the one relevant organism.

2.3 Biological data repositories

Biology has turned into rich data science. There is a huge need for biological data for the scientists and availability should also be in the digital form to read. To avoid the waste of time in collecting existing research data, it is easier to get computers in digital form. It is time-consuming and hence databases start evolving. Initially, it was one or two biological databases but now in 2019, it has increased to thousands of public biological databases available with the computational prospects.

Biological databases are divided into two categories, namely sequence databases and structural databases. Sequences databases consist of all types of biological sequences including both DNA and protein, whereas structural databases consist of only proteins. DNA and protein primary sequence repositories are detailed in Figure 9.

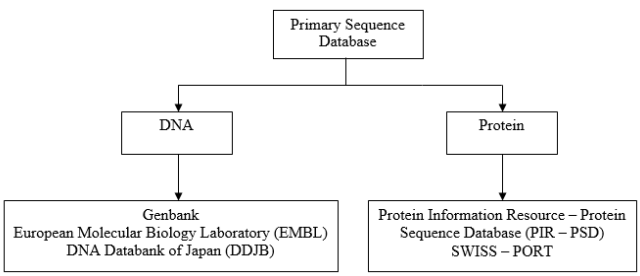

Figure 9: Primary Sequence Division and Databases

Each biological databases has its own identity based on specifications, namely, id, accession numbers, date uploaded, the name of the gene, classification of an organism, the name

of the organism, sequence number, etc. There are specific public databases which deal with the specific domains as their chosen area and all the collections are related to that. For instance, the Medline database consists of only medical related data especially the medicine oriented literature analysis with regard to the molecular changes in it.

There are derived databases containing information relating to the genes from the available public primary repositories including both DNA and protein sequences. These are shown in Figure 10. The derived databases are also divided into two categories such as sequence and structural derived databases. The structure primary - derived databases are shown in Figure 11. The challenges facing all these repositories relate to handling the huge volume of data on their repositories which are increasing tremendously day by day as there are many applications to deal with these stored data. The retrieval needs of these data from databases are human genome project, biological software development, disease analyses, drug discovery, etc.

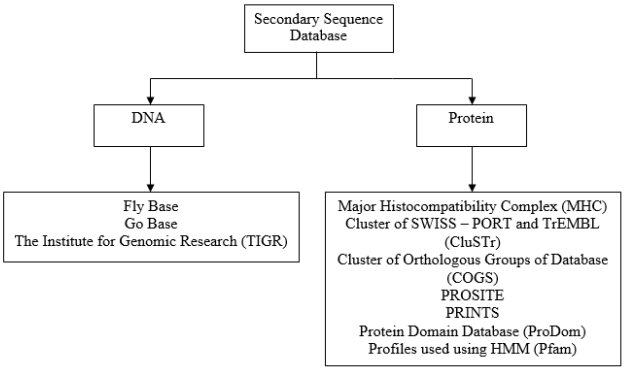

Figure 10: Secondary Sequence Division and Databases

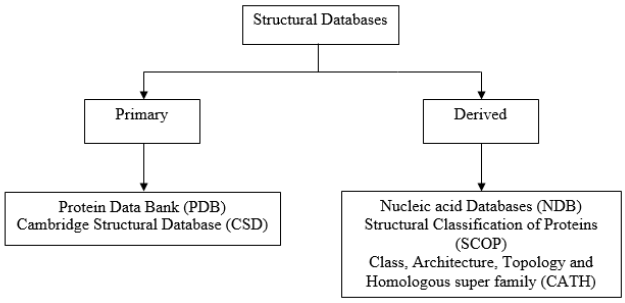

Figure 11: Structural Protein Division and Databases

2.4 PPI (Protein-Protein Interaction)

Proteins plays a major role in the cell, performing biological processes such as, mutations, similarity matching, drug discovery, gene expression etc. The analysis of protein function mainly deals with the unique protein synthesis, but the larger number of protein combines or interacts with the primary structure of other proteins for their functionality performance. The study on proteins should also include the protein-protein interaction models to fully understand the protein functions. So the PPI is considered as major processes.

PPI includes the physical contacts of molecules present in two or more proteins from biochemical actions of electrostatic forces, which also include the hydrophobic effect. These processes are divided into two types, viz. stable effect and transient effect. In these effects, the bonds embedded into them are of both strong and weak. Multi sub unit complexes are taken and made to interact in the protein structure using a stable effect. There are other types of PPI too that includes covalent and non-covalent homo-oligomers and hetero-oligomers. Multiple amino acids are present with a protein, which are linked together by peptide bonds, thereby forming a long chain. The basic principle of the interaction can be explained as, if protein X, protein Y interact, their DNA binding domain and activation domain combine to form a functional transcriptional activator. There are several applications of PPIs like identifying novel PPIs, to know how the mutation affects a protein's interaction with other proteins, manipulating protein-protein interactions and so on.

2.5 NGS

Rapidly evolving high throughput technologies provide biological data on a very large scale. This research work also provides a review of the recent developments in liner time algorithms and tools for big data analysis produced by next-generation sequencing and a comparison of commonly used tools with different algorithms for NGS big data analysis based on their performance, input format and output format, etc. The diversity of contexts in which biological data analysis is performed requires a study of several problems. This action paves a way to find a provably better algorithmic tool to the underlying optimization problem.

The methods to analyze NGS big data includes various algorithms on various tools which differ by their performance, complexity, cost, etc. The research has gathered together or acquired an increase in a number of sample handling, detecting the variants between the genes and analysis of bioinformatics on mentioned tools. NGS resultant data generates short reads that address various biological questions. Next-generation sequencing technologies have grown widely in the field of identifying patterns in epigenetic, genomic and transcriptome levels to predict factors such as disease, age, etc.

This section provides a detailed overview of open source software available for alignment programs, assembly and variant calling methodologies with the challenges to maintain these large giga base pairs big data Next-generation sequencing data analysis methods involves four models.

NGS data has moved biology in a whole including molecular, micro, etc. into the big data era. For example, the European Bioinformatics Institute doubles two petabytes of genomic data every 18 months out of a total storage 20 petabytes and doubles the biological data every 9 months generates voluminous data, running to hundreds of terabytes to petabytes, it is very difficult to maintain this big data with no loss of quality and by maintaining storage capacity and capabilities. Hence, the bioinformatics tools are discovered

and they are refined with new efficient algorithms and technologies for managing the big data. Few public organization, EBI(European Biotechnology Information), NCBI, NIH, etc. has the capacity to overcome the drawbacks seen in storage management and information extraction on big data ease. The better way to process these big biological data in the cloud is by making use of cloud providers such as Amazon etc. NGS data analysis always is preferred a linear time algorithms in their software tools.

2.6 Properties of Protein Prediction

The performance analysis of protein prediction was done using modeler 9.15 analyzer since it has lesser time complexity compared with others. Though there are many methods for converting unstructured proteins to structured proteins. Protein prediction explains the conversion of unstructured protein to structured protein where the unstructured proteins which are inputs to the modeler software are crab datasets taken from NCBI public database, which is comprised of unstructured protein sequences in two dimensional formats and these are given as inputs to the BLAST algorithm for the prediction of protein sequences.

The BLAST algorithm compares the input sequences with the sequences in the database. Another algorithm named FASTA is used for finding the sequences between input and database but it is less preferred as it is less efficient. BLAST uses a statistical method called Altschul-Dembo-Kerlin which calculates how many times a match occurred statistically and selects the sequence(target) based on it. With the help of unstructured target sequences, we can identify structured protein sequences which are written down to identify in a public database PDB. Query sequence structure is used to identify these structures

2.7 Protein Clustering

The examination of organic information mining provides a profound comprehension of how the information mining strategies such as grouping, arrangement, affiliation, expectation, relapse, etc. can be produced. This synthesis of clustering is not that easy in practice as it is based on several constraints requiring the separation of objects into different groups to facilitate data mining and further procedures with the purpose of getting meaningful data from the given bunch of data and making it in a better interface to the user ability for proceeding further.

As humans are very smart and interested in finding solutions for every problem which they come across while traveling through time, several methods have already come up with many new interesting algorithms for clustering problem as well. Again this very special and interesting human nature of dissatisfaction and engineers having a keen interest in innovation things. It motivates to innovate more than one clustering algorithms for clustering which was invented by different engineers and researchers developed into numerous softwares such as, Bioweka, weka, etc.

In the case of each cluster, objects inside are the same between themselves, while they are not the same when compared with other clusters formed earlier. As mentioned before, there are different clustering algorithms to form clusters where bioweka is a tool for comparison. Clustering algorithms are compared in two ways namely, partitioning based and non-partitioning based. Partition based are K-means algorithm, Farthest First algorithm, and Expectation maximization, cobweb algorithm belongs to non-partition based clustering algorithm. While comparing the clustering algorithms, it is based upon two different things i.e. partitioning and non-partitioning. The first partitioning based algorithms and non partitioning based algorithms are followed by a comparison between the partitioning and non partitioning based algorithms.

2.8 Microorganisms culture

Microorganisms are identified on the basis of protein and nucleotide analysis, which includes PCR (Polymerase Chain Reaction) and get electrophoresis. The small region of targeted organisms DNA sequence was sent for phylogenetic analysis for constructing a tree for the resulting taxa. Modeler 9.15 and orthologs modeling approach paved a way to predict a targeted identified organism (e.g. Bacillus cereus) protein structure Figure 12. ARCSA algorithm is an extended version of motif finding algorithms to improve efficiency and accuracy in finding the pattern using the index value of sequences.

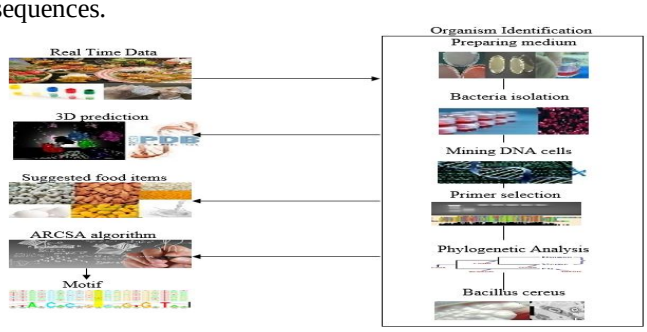

Figure 12: Graphical Representation of Research Model

The protein structure prediction gives a detailed understanding of the characteristics and functions of the considered protein in the three-dimensional model. The prediction concept done here with a learning from the neighbours, such as the targeted sequence has been given as input to the database and from the accuracy of match that takes place there, the first 4 prior known protein structures are retrieved, with the help of the neighbours predicted the present protein structure in three dimensional model. The process of measurement, analysis, and judgment for the neighbour collection are done using BLAST. BLAST provides the accurate matches of a given sequence. For predicting the values other than the nucleotide or amino acid bases, the error rate was Aj - Aj(I), but, in the sequences, the result was considered based on the percentage match done through different algorithms (e.g. BLAST), Needleman Wunsch, Smith-Waterman, etc.) are at the error rate of

IBasevalueindex - BasevalueindexI towards the matches based on general prediction formulae.

In the process of sequencing the large datasets, there are many problems seen in the various types of inputs provided. Some of the input-oriented domains with their specific problems are mentioned. These include GRN (Gene Regulation Network) analysis in an expression is a complicated one by producing correlational possibilities and has the challenge to face an iterative problem. It requires a very large scale data analytics system for a regulatory identification in an affected or diseased network. It is also useful for drug analysis approval to the market.

The process sequence analysis includes searches in the database, which has various dimensions of big data incorporated in it are not quite ease rather than the use of big data technologies. Sequence and PPI, since from 1980, high volumes of data are managed in a database and their updates with new inventions also increase the volume, velocity and varied data formats.
PPI results voluminous data with changes in their nature provides a challenge for an efficient and scalable framework to act in a fast and accurate PPI} pattern generation. Pattern generation plays a major role in the gene analysis to find the promoter sequences related to the ancestor gene. Microarray data cost reduction aspect leads the growth tremendously and it requires big data technologies for speed to construct co-expression regulatory networks using voluminous microarray data.

Bacillus cereus organism was first found out of majority presence in the collected five samples and the functional characteristics of it are predicted. Microorganisms found from each sample divided into groups to suggest the combination of food items in the five samples. In total eight sequences are considered including Bacillus cereus, which is applied to ARCSA algorithm for finding the repeated pattern. The present study provides an insight into the Bacillus cereus structural, functional aspects with other organisms found in priority bases of occurrences and the similarity pattern it embeds into food items suggested through ARCSA algorithm.

3 RESULTS AND DISCUSSION

All activities in the human body are because of chemical reactions, which results in different chemical bonds. These bonds constitute the behavior of internal organs. An abnormal chemical reaction may result in a difference in the behavior of organs, which provides a base for analyzing, monitoring the anomalies and functioning of the human genome. Genome analyses can be performed by two processes, viz.PPI – PPI refers to physical contacts established between two or more proteins as a result of biochemical events and NGS. The main goal of clustering in motif analysis is to identify the similar data with corresponds to a gene, patterns, etc. available in the database. This research work deals with the two types of datasets, viz. first one, the primary input is a real-time environment food samples, the collected bio specimens are analyzed in the wet lab with the required water, direct ventilation and specialized piped utilities for microorganisms identification shown in Table 1**,** then the merge sets filter used on the datasets for PPI factor followed by motif analysis, score calculation and structure prediction. The second one is data downloaded from the biological databases, such as Xenopus laevis, Drosophila busckii, etc. for cross-validation of the proposed system.

Table 1: Found Microorganisms in RTSamples

Samples	Microorganisms Identified	Suggested Food Combinations in Samples	Suggested Microorganisms in Food Items
1	Bacil-lus,Proteus,Shigella	Bos Taurus, Musa, Mangifera Indica	Bacillus, E-coli, etc.
2	Bacillus, Klebsilla	Bos Taurus, Musa, Mangifera Indica, Curcuma Longa	
3	Pseudomonas, Proteus	Bos Taurus, Musa, Mangifera Indica, Oryza Sativa	
4	Bacillus, Pseudomonas, Proteus	Bos Taurus, Musa, Mangifera Indica, Eleusine Coracana, Prunus Dulcis, Anacardium Occidentale	
5	Bacillus, E.Coli, Shigella	Bos Taurus, Musa, Mangifera Indica, Gallus Gallus Domesticus	

4.CONCLUSION

The problem addressed in this research work is to effectively identify the consensus string given any biological data or database and thereby determining the microorganisms in the samples and protein structure available in the sequence produced. Prior to the process of identifying consensus string, the huge volume of amino acids in the dataset got reduced which are highly correlated and redundant in nature. Moreover, the computational complexity of the pattern identification algorithm got reduced..

REFERENCES

[1] Cibulskis K, McKenna A, Fennell T, et al. ContEst: estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* 2011;27:2601–2602. https://doi.org/10.1093/bioinformatics/btr446

[2] U. Vignesh and R. Parvathi. (2018). 3D visualization and cluster analysis of unstructured protein sequences using ARCSA with a file conversion approach. Journal of Supercomputing, https://doi.org/10.1007/s11227-018-2319-4.

[3] U. Vignesh and R. Parvathi. (2017). Clustering on Structured Proteins with Filtering Instances on Bioweka. Journal of Engineering Science and Technology, Volume 12, No.) 3, 820 – 833.

[4] U. Vignesh and R. Parvathi. (2017). Bioweka Classification On Egg And Crab Protein Structured Sequence With Filter Instances. Journal of Advanced Research in Dynamical and Control Systems - Vol. 9. Sp– 17 / 2017.

[5] U. Vignesh and R. Parvathi. (2016). Next Generation Sequencing Data Analysis Software and Methods: A Survey. International Journal of Control Theory and Applications. Volume 9, No. 52, 1 -15.
[6] U. Vignesh and R. Parvathi. (2017). Biological Data Analysis and Visualization: A Survey - Modern Technologies for Big Data Classification and Clustering, DOI: 10.4018/978-1-5225-2805-0.ch010.
[7] Nazir Ahmad, Hatim M. Elhassan Ibrahim Dafallaa, Mohammed Burhanur Rehman, et al. Comparative Study on Load Balancing Algorithm for Multiprocessor Interconnection Networks. IJATCSE 2019;8(3):410-414. https://doi.org/10.30534/ijatcse/2019/13832019
[8] B.Manoj, K.V.K.Sasikanth, M.V.Subbarao, et al. Analysis of Data Science with the use of Big Data. IJATCSE 2018;7(6): 87-90. https://doi.org/10.30534/ijatcse/2018/02762018
[9] Lanka, S., Madhavim, R., Abusahmin, B. S., Puvvada, N., & Lakshminarayana, V. (2017). Predictive data mining techniques for management of high dimensional big-data. Journal of Industrial Pollution Control, 33, 1430-1436.
[10] Madhavi, R., Karri, R. R., Sankar, D. S., Nagesh, P., & Lakshminarayana, V. (2017). Nature inspired techniques to solve complex engineering problems. Journal of Industrial Pollution Control, 33(1), 1304-1311.
[11] Puvvada, N., & Prasad Babu, M. S. (2018). Semantic web based banana expert system. International Journal of Mechanical and Production Engineering Research and Development, 8(3), 364-371.