CHAPTER - 1

INTRODUCTION

1.1 PROLOGUE

Information Technology has made a great revolution in all walks of human life. Such a tremendous growth of information technology has greatly influenced the study of Biology, since Biology is growing with huge amount of information. This information gathering runs with the advent of chemical sequencing expertise. Nowadays Biologists are capable to collect millions and millions of information regarding the sequences of two types of molecules, namely Nucleic acids and Proteins. Biologists analyse biological data to find answers to the origin, function and structure of living organisms. For the interpretation of this massive data gathered by the Biologists, computational tools are required. Computers are being provided in a great way for handling vast quantities of biological data. Such an evolution of information inflow of bio-molecular details led to the birth of new branch of biology which is an intersection of Biology and Informatics. This new branch embracing the fields of Bio-Molecular data and Information Technology is called Bioinformatics [83].

It is a multidisciplinary subject that requires knowledge of several subjects such as Molecular Biology, Biochemistry, Computer Science, Mathematics and Statistics. It is concerned with representation, organization, manipulation, distribution, maintenance and use of information, particularly in digital form. Bioinformatics incorporates two basic things. The first one is the development of database to store and search data and the second one involves the development of statistical tools and algorithms to analyse and determine relationship between biological data sets such as structure of proteins, gene
expressions, sequence of De-ox Y Ribo Nucleic Acid (DNA) and Ribo Nucleic Acid (RNA). Bioinformatics applies algorithm and statistical technique to the interpretation, classification, and understanding of biological datasets [5].

The importance of Bioinformatics is increasing day-by-day. Further the eminence of Bioinformatics could be recognised by three major areas such as [83]

- **Information Management** - Development of database containing all biological information, updating of database and development of better tools for data designing, annotation and mining.

- **Bioinformatics** is a reliable area to understand the relationships between organisms. It provides a great way to the growth of molecular phylogenetics.

- **Drug designing** – It is a recent application in Bioinformatics. This application has come up due to the challenging situations in medicine. Proteins that are more abundant in disease tissue can be identified and new drugs that target these proteins can be developed.

A proteome is the entire complement of proteins expressed in a cell at a given time and it is a branch of functional genomics. It is also an interdisciplinary science that includes Biology, Bioinformatics, and Protein Chemistry. The challenge of understanding the function of each gene in the genome has led to the development of a new aspect of Biology, namely Functional Genomics. The basic research has moved from genes and genomics to proteins and proteomes and their functional analysis. Protein structure prediction and molecular modelling can also be considered as a part of proteomics. Proteomics methods are mainly
used to determine the complement of proteins that are expressed in a cell and how this complement changes under different conditions. Proteomics is an incorporation of the techniques like Protein Data Bank (PDB), Mass Spectrometry, Nuclear Magnetic Resonance, and X-Ray Crystallography for protein identification and characterisation [23].

1.2 DATA MINING IN BIOINFORMATICS

Data mining is the extraction of hidden predictive information from large amount of data. It is also becoming an increasingly significant tool to transform this data into information. Data mining is commonly used in areas such as Marketing, Surveillance, Fraud Detection, and Scientific Discovery. Data mining for Bioinformatics enables researchers to meet the challenge of mining vast amount of bio-molecular data to discover real knowledge [53, 62]. Bioinformatics and Data Mining provide exciting and challenging research and application areas for Computational Sciences.

Data mining commonly involves the following four classes of tasks: [34]

**Classification:** Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome. Secondly, the algorithm is given a set of patterns called prediction set, which contains the same set of attributes, except for the prediction attribute. The algorithm analyses the input and produces a prediction. The prediction accuracy determines the performance of the algorithm [57].
**Clustering:** It is like classification, but the patterns do not have predefined labels. It is an unsupervised learning technique in pattern recognition. Data clustering is applied to many application areas such as Business, Bioinformatics, Medicine, Chemistry, Engineering, etc. Owing to increasing amount of data collected in database, cluster analysis has recently become an active topic in data mining research. Cluster analysis includes two major aspects: Clustering and Cluster validation. Clustering is a process to group similar items together into groups according to certain criteria. These grouped objects are called clusters. The quality of clusters is then evaluated using cluster validation techniques [8, 31, 48].

**Regression:** Regression is a data mining function that predicts a number. A regression task begins with a dataset in which the target values are known. During training process, a regression algorithm estimates the value of the target as a function of predictors for each case in the built data. The relationship between predictors and target are summarized in a model, which can then be applied to a different dataset in which target values are known. It is an attempt to find a function which models the data with least error [34].

**Association rule learning:** It is a popular method for discovering interesting relationship between variables in large databases. Here, hidden relationships are expressed as a collection of association rules and frequent item sets. Frequent item sets are simply a collection of items that frequently occur together and association rule suggests a strong relationship that exists between two items. Association rule mining has become an important data mining technique due to its descriptive and easily understandable nature of rules. It has proved to be useful in many domains, for instance, microarray data analysis, recommender systems and network intrusion detection [54].
1.3 STRUCTURE OF MOLECULAR BIOLOGY

The basic paradigm of Protein Synthesis emerged in the year 1958. The central dogma of Molecular Biology [17, 58] states that DNA is an important genetic material found in most organisms. The gene sequences found in DNA are transcribed into messenger RNA. The messenger RNAs are then translated into the polypeptide chains of proteins by synthesis on ribosomes.

Figure 1.1: The Portrait of the Central Dogma

Figure 1.1 shows the portrait of central dogma [41]. The backbone of molecular biology is represented by the following four major stages [17].

**Replication:** DNA replicates its information in a process that involves many enzymes.

**Transcription:** The DNA codes for the production of messenger RNA (mRNA).
**Processed:** In Eukaryotic cells the mRNA is processed and migrated from the nucleus to the cytoplasm.

**Translation:** Messenger RNA carries coded information to ribosome. The ribosome reads this information and uses it for protein synthesis.

### 1.3.1 Amino Acid and its Role in Protein Structures

Amino acids are the raw materials of proteins. All amino acids contain the elements such as Carbon, Hydrogen, Oxygen, and Nitrogen. There are 20 different amino acids commonly occur in nature and they are put together into a polypeptide chain on the ribosome in a process called protein synthesis. During this process the peptide bond, the covalent bond between two amino acid residues, is formed. Each of 20 amino acids has its specific characteristics, which are mostly defined by the type of the side chain they possess. It is the side chain which makes each of the 20 amino acids unique and provides it with a role to play in a protein structure [23].

Based on the propensity of a side chain to be in contact with polar solvent like water, it may be classified as hydrophobic, polar and charged. The charged amino acid residues include Lysine (+), Arginine (+), Aspartate (-) and Glutamate (-). Polar amino acids include Serine, Threonine, Asparagine, Glutamine, Histidine, and Tyrosine. The hydrophobic amino acids include Alanine, Valine, Leucine, Isoleucine, Proline, Phenylalanine, Tryptophane, Cysteine, and Methionine. The amino acid Glycine does not have a side chain and is hard to assign to one of the above classes. However, Glycine is often found close to or at the surface in loop regions, providing high flexibility to these regions. In contrast, Proline has the opposite effect, providing rigidity to the protein
structure by imposing certain torsion angles on the segment of the polypeptide chain. These two residues are often highly conserved in protein families since they are essential for preserving a particular protein three-dimensional fold [5].

On synthesis, proteins spontaneously fold into 3-D structure. Finally, the shape of the protein determines their biological function [5]. Pictorial representation of the relationship between amino acid sequence and protein structure is shown in Figure 1.2 [83].

![Diagram showing the relationship between amino acid sequence and protein structure.](image)

**Figure 1.2 Relationship between Amino acid Sequence and Protein Structure**
1.4 DESCRIPTION OF PROTEIN STRUCTURES

Proteins are polymers built from amino acids and are most important building blocks. It creates the structure required for their panoply of functions by variations on a common underlying chemical scheme. There are four levels of structures in proteins [5]. Protein structure representation from primary level to quaternary level [40] is shown in figure 1.3.

**Primary Structure:** The primary structure of a biological molecule is a chain of amino acids or exact specification of its atomic composition and the chemical bonds connecting those atoms. The number of amino acids in a protein may range from two to several thousands.

**Secondary Structure:** It occurs when sequence of amino acids are linked by Hydrogen bonds. The most common secondary structure elements are Helices and Sheets. Different sequence of amino acids forms different secondary structure elements. Specific amino acid sequences and secondary structures derived from these sequences confer unique properties and functions. They are referred to as “Motifs” and “Domains”. In some cases, amino acid sequences within motifs and domains are said to be highly conserved. A detailed review about motifs is provided in section 1.5.

**Tertiary Structure:** It occurs when attractions are present between \(\alpha\) – Helices and \(\beta\) – Sheets. It is said to be three dimensional structures of proteins which it assumes by a folding process.

**Quaternary Structure:** Many folded protein molecules are combined to make multi sub-unit complex. This complex structure is called quaternary structure of proteins.
Motif is a consecutive string of amino acids which frequently occur in a given protein sequence whose general character is repeated or conserved in all sequences in a multiple alignment at a particular position. The term motif in biology is used to denote a functional component that is common to a set of proteins [18]. Motifs are over-represented patterns in related proteins. Protein motifs can be classified into sequence motifs and three-dimensional motifs based on the level of structures. For convenience, a protein is usually represented as a sequence or string on an alphabet of twenty characters except B, J, O, U, X and Z [23].
The twenty amino acids form the primary structure of motifs. Protein motifs over the primary structures are called protein sequence motifs. Such motifs can be defined as short conserved segments common to a set of protein sequences which has biological significance. Proteins actually have complex three-dimensional structures. Motifs can also be extracted from their three-dimensional structures; such motifs are called as protein three-dimensional motifs. The objective of motif extraction in proteins is to discover these over-represented patterns-motifs in related proteins. During this procedure, the motifs to be extracted are not known well in advance.

The extraction of motifs in proteins can help to classify protein families and predict protein functions. Moreover, it provides valuable information about the evolution of species. Before the automatic motif extraction algorithms in Computational Biology are developed, motifs are often discovered manually [38, 85]. This procedure is tedious and difficult. Recently, numerous research activities are carried out on the automatic protein motif extraction. The majority of the activities are on protein sequence motif extraction because the protein sequence information is abundant and many existing algorithms and methods on text analysis can be applied to the protein sequence data.

Besides, the three-dimensional structure information is lacking for many proteins. However, the automatic extraction of protein sequence motifs is not an easy and straightforward task. The sequence motifs are usually inexact, containing mismatches and gaps in different sequences. They are often subtle, hard to distinguish from random motifs. Recently, many automatic protein sequence motif extraction algorithms have been proposed.
1.5.1 Motif Patterns

There are many definitions for the protein sequence motif. One of the definitions for the protein sequence motif is given as follows [18]:

Given an integer $h$, a set of $Z$ sequences \{seq$_1$, seq$_2$, seq$_3$, \ldots, seq$_Z$\} on the alphabet \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}, and a threshold $td \leq Z$, a motif is a pattern occurring in at least $x$ sequences of \{seq$_1$, seq$_2$, seq$_3$, \ldots, seq$_Z$\} with at most $h$ errors. The motifs to be extracted are usually a collection of sub sequence that are constrained and represented by certain expressions.

There are basically two different types of sequence motifs based on their representation: Deterministic motifs and Probabilistic motifs [80]. Deterministic motifs will either match or not match a sequence while probabilistic motifs will be given a score of probability of matching against a sequence. The above definition is more suitable for the deterministic motif. For the probabilistic motif, motifs can be considered as the patterns ranking highest when matched against the given set of proteins among all the possible patterns. The algorithms to extract different types of motifs in proteins are very different. Some researchers also work on combining the above two types of motifs [80,102,103]. This thesis focuses the identification of deterministic motifs only.

1.5.2 Related Work on Protein Sequence Motifs

In this section, strength and weakness of several motif discovery methods are studied and compared. The study of relationship between protein structure and its sequence is one of the most important tasks of current Bioinformatics research [51, 102]. Recently, machine learning and data mining technique have largely been used in Bio-Medical data
analysis. The basis of machine learning is to program computers using example data or past experience to detect patterns and regularities. It includes various applications described in section 1.2. In this thesis, Granular computing models based on clustering technique are chosen as the motif discovery method.

Many biochemical tests suggest that a sequence determines conformation completely, because entire information that is necessary to specify protein interaction sites with other molecules which are embedded into its amino acid sequence. Various researches showed that a relatively small number of structurally or functionally conserved sequence regions are available in a large number of protein families. Representation of these conserved regions can range from simple sequence motifs to complex descriptors. These descriptors are profiles, Position Specific Scoring Matrices (PSSM), and Hidden Markov Models (HMM) [91]. Sequence motifs and profiles obtained from biologically significant regions may be used to predict any subsequent reoccurrence of structural or functional areas on other proteins. These functional and structural areas may include enzyme-binding sites, prosthetic group attachment sites or regions involved in binding other small molecules.

Some of the major secondary databases are PROSITE [46], PRINTS [6], BLOCKS [37], and PFAM [90]. PROSITE was the first secondary database discovered in the year 1988. Protein families can be characterised by a single most conserved motif observed in a multiple alignment of known homologues. This information has been pointed out to the development of PROSITE [46, 91]. It is a database of protein families, motifs, and domains. Computational tools help to search such a database which helps to determine to which family of proteins a new sequence belong or which domain or functional site (s) it might contain.
Within PRINTS fingerprint database motifs are encoded as ungapped, unweighted local alignments [6, 91]. It contains motifs extracted from different regions of multiple sequence alignment which in turn helps in predicting existence of similar motifs. BLOCKS database accommodate the most highly conserved motifs present in PROSITE database [34, 91]. PFAM is a database which contains motifs obtained from both highly and less conserved regions, which is the major difference from PROSITE and PRINTS. The sequence motifs obtained from PROSITE, PRINTS, BLOCKS, and PFAM are developed from multiple sequence alignment. These sequence motifs and profiles only inspect conserved elements of sequence alignment from the same protein family and carry little information about conserved sequence regions which transcend protein families.

The most commonly used tools are MEME, Gibbs Sampling [67] and Block Maker [37]. New algorithms include MITRA [28], Profile Branching and generic motif discovery algorithm for sequential data [82]. When using these tools, users are asked to give several protein sequences, presented in FAST Alignment (FASTA) format, as the input data. The size of input data set is limited to the above algorithms, therefore the motifs extracted from the above algorithms carry only little information that transcend across protein families [44].

Some researchers try to obtain sequence motifs that are universally conserved across protein families. To achieve the goal input data size should be large enough to represent all known protein sequences. Han and Baker used an automated approach to identify local sequence motifs that transcend protein families [35]. In their work, K-Means clustering algorithm was used to identify sequence motifs. The algorithm chooses set of initial points for cluster centers in a random
manner. Selecting initial points randomly leads to an unsatisfactory partition because some initial points may lie close to each other. In order to overcome the above mentioned problem, Wei Zhong has proposed Improved K-Means clustering to explore sequence motifs.

Improved K-Means algorithm [119] tries to obtain initial points by using Greedy approach. In that approach, for each run, K-Means clustering algorithm is executed for fixed number of iterations and then initial points which have capacity to form clusters with good structural similarity are selected. The distance of chosen initial points is checked against points already available in the initialization array. If minimum distance of newly selected points is greater than threshold value, these points are added to the initialization array. In this area of research, dataset is said to be huge and selecting initial points using the greedy approach leads to high computational cost. Computational cost is a major problem which is being faced when input dataset is very large.

Hence, Bernard Chen has proposed Granular Computing Model using Fuzzy clustering technique. In his work, segments are first partitioned into small information granules using fuzzy clustering method. Then for each granule, Improved K-Means algorithm is executed. Finally, the clusters formed in each granule are combined to find final sequence motif information [10, 11]. In his another work, Fuzzy Greedy K-Means approach [10, 12] fuzzy granular computing technique is adopted and then initial points chosen are greedier than the Improved K-Means algorithm. In the Greedy K-Means, the best centroids are selected after five runs of K-Means and then K-Means algorithm is executed by considering those centroids.
A dataset containing large volume of features which may be insignificant or even harmful to the process of learning. Such noisy data may cause a worse situation such as confusing the mined information or hiding the impact caused by the true value. Hence, Efficient Super Granular SVM Feature Elimination Model [10, 13] was proposed for small information granules generated by FCM and then for each granule Greedy K-Means clustering technique was applied. Then for each granule ranking SVM was adopted to rank all members in each cluster generated by Greedy K-Means clustering algorithm and then filter out lower ranked members. After feature elimination step, all surviving data points were collected from each information granule to create final protein sequence motif information.

1.6 RESEARCH CHALLENGES

Biological data are being produced at a phenomenal rate. For example as of April 2013, the GeneBank repository of nucleic acid sequences contained 16, 41, 36, 731 entries [43] and SWISS PROT database contains 5, 40, 052 of protein sequences [42]. Enormous quantity and variety of information are being produced with these nucleic acid sequences and proteins. As a result of the surge in data, computers have become indispensable to biological research. The challenges in the area of bioinformatics are three-fold.

First, at its simplest, bioinformatics organises data in a way that allows researchers to access existing information and to submit new entries as they are produced. The amount of information stored in these databases is essentially useless until analysed. Thus the purpose of bioinformatics extends much further. The second aim is to develop high performance computing tools and resources that aid in the analysis of
data and to speed up the searching time. The third one is to deal with some noisy data this may be useless or even harmful to interpret the results in a biological meaningful manner.

1.7 PROBLEM DEFINITION

Protein sequence motifs information is very important to the analysis of biologically significant regions. The conserved regions have potential to determine the conformation, function and activities of the proteins. The main aim of this work is to obtain protein sequence motifs which are universally conserved across protein family boundaries. Therefore, unlike most popular motif discovering algorithms, the input dataset used in this work is extremely large. As a result, an efficient technique is demanded.

Owing to large input dataset considered in this work, an efficient segment selection technique is required to identify important sequence segments. Secondly, considering large volume of data as a single granule requires high execution time. Therefore, in this work, short recurring segments of proteins are explored by utilizing novel granular computing methods.

1.8 RESEARCH CONTRIBUTIONS

There are lots of challenges in the area of data mining and its applications. In this thesis, bioinformatics is chosen as an application area for applying data mining techniques. The research contribution of this thesis is twofold: In this work, the dataset is said to be very large and hence different unsupervised segment selection techniques have been proposed to select significant segments and then different clustering algorithms are applied to identify protein sequence motifs.
Second aspect of the contribution focuses on granular computing techniques. Large complicated problems can be solved efficiently using granular computing models. Hence, crisp and soft granular computing techniques have been proposed and implemented using MATLAB to identify significant protein sequence motifs that transcend across different protein families.

The main contributions of this thesis are mentioned below:

- Unsupervised segment selection techniques such as Shannon-Entropy, Singular Value Decomposition (SVD) Entropy, Information Granule based, and Double Refined K-Means are proposed to select important protein sequence segments from large volume of data.

- A study on benchmark K-Means clustering algorithm is done. An automated approach namely Weighted K-Means clustering algorithm is proposed and implemented. Greedy seed selection techniques are also used to extract protein sequence motifs.

- Crisp granular computing techniques such as K-Means, Weighted K-Means, and K-Harmonic Means are proposed and implemented in order to generate sequence motifs in an efficient manner.

- Some of soft granular computing techniques such as Fuzzy C-Means, Modified Fuzzy C-Means, and Adaptive Fuzzy C-Means are proposed to obtain sequence motifs with high secondary structural similarity.

- DBI measure is used to analyse and evaluate the quality of clusters produced by the proposed algorithms. The
biological significance of motif information is then evaluated using HSSP-BLOSUM62, Z-Score, and INFORMATION-GAIN measures.

1.9 STRUCTURE OF THE THESIS

The framework of the proposed protein sequence motif mining problem is shown in Figure 1.4. This thesis mainly deals with applying data mining techniques to extract protein sequence motifs efficiently. With the aid of domain knowledge and data mining, protein sequence motif mining process is used to efficiently identify significant motif patterns. The results are observed to be satisfactory. The cluster quality and motif quality are evaluated and analysed with statistical indices and domain knowledge. The thesis is organised into seven chapters and the gist of each chapter is detailed here under.

Chapter 1: Introduction

This chapter gives a brief introduction on Molecular Biology and proteins. A note on protein structures and sequence motifs are also given in this chapter to gain knowledge about the existing relationship between secondary structure of proteins and its sequence motifs. The pros and cons of several motif discovery methods which laid a way to overcome the research challenges in protein sequence motif mining problem are given. Basics of data mining and granular computing techniques are also discussed. Research motivations and contributions are given in this chapter.

Chapter 2: Experimental Setup

The second chapter explains the acquisition of protein sequences from database and how to prepare a dataset by converting protein sequences into protein sequence segments using HSSP files. Cluster
validation techniques and the methods to identify significance of extracted motif patterns are also presented.

Figure 1.4: The overall Structure of Protein Sequence Motif Mining
Chapter 3: Segment Selection Techniques

In this chapter, the importance of unsupervised segment selection techniques for protein sequence motif mining is dealt. The main aim of this chapter is to improve the performance of mining process and to obtain better motif patterns. Wrapper and Filter techniques for segment selection process are briefly discussed in this chapter. The expected value of information contained in each protein sequence segment is calculated using Shannon Entropy measure and a novel SVD Entropy measure. Information-Granule based method tries to select segments based on basics of granular computing method. Finally, Double-Reined K-Means segment selection method combines granular and clustering method to obtain significant segments from protein sequences.

Chapter 4: Clustering Techniques

In this chapter, benchmark K-Means algorithm is studied and analysed. Weighted K-Means algorithm is proposed for the first time to identify hidden motif patterns. K-Means and weighted K-Means algorithms suffer from random initialization of centroids which affect final motif patterns. Hence, Greedy initialization technique is adopted to
select the best seeds for clustering process. The results are then evaluated and found to be satisfactory than random initialization technique.

**Chapter 5: Crisp Granular Computing Techniques**

Large and complicated problems are solved efficiently using granular computing techniques. In this work, the dataset is said to be very large and applying clustering techniques on whole dataset is a time consuming process. Hence, in this chapter some of crisp granular computing techniques are proposed to identify significant motif patterns. The techniques such as K-Means, weighted K-Means, and novel K-Harmonic Means granular techniques are proposed. This chapter has given us the knowledge as how to handle large dataset in an efficient way and how to reduce computational time as well.

**Chapter 6: Soft Granular Computing Techniques**

This chapter discusses soft granular computing techniques. Soft computing appears to be good for this task, since it can assign gradual degree of membership to a cluster. Another reason for applying soft computing is to avoid high level of noise present in numerous biological data. The proposed method is based on soft granular techniques such as Fuzzy C-Means, Modified Fuzzy C-Means, and Adaptive Fuzzy C-Means algorithm. A novel bisecting K-Means algorithm is proposed for the first time to select seeds for clustering process. The results obtained using proposed techniques are satisfactory.

**Chapter 7: Conclusion**

This chapter concludes the thesis with a summary of the key findings of the research work carried out with directions for further research.
1.10 SUMMARY

This chapter introduced the basic concepts of Bioinformatics, proteins, relationship between amino acids and proteins, structure of proteins, and sequence motifs. Data mining concepts aid in identifying protein sequence motifs common to a protein family. The outline of the major research challenges and contributions to meet the challenges are also discussed.