CHAPTER I
INTRODUCTION

Proteins are building blocks of cells and tissues; they play a vital role in executing and regulating many biological processes. Amino Acids (AA) are the building blocks of all proteins. There are 20 AA that combine to generate uncountable number of protein sequences. In addition to these, there are two more characters \{B, Z\} that appear in these sequences, which are substitutions/ derivations for some amino acids. In reality, only small subsets of sequences appear [Mount, 2001]. Proteins are categorized into families, classes, subclasses based on functions and structures [Hooman, 2005].

A large number of new proteins are being added to genomic databases by research community as and when they are discovered. Several methods (PROSITE, PRINTS, PFAM, etc.,) are available to classify these proteins into families based on their structural and functional properties. Precisely, there is an acute need for development of faster, more sensitive and accurate methods to meet the present challenges of classification.

PROSITE, PRINTS and PFAM are protein classification methods and they utilize multiple sequence alignments to build Hidden Markov Models. PROSITE is a database consisting of information of significant sites, patterns, and profiles that specify different protein families [Hulo, 2004]. PRINTS is a database of protein fingerprints [Attwood, 2003]. Fingerprints are sets of short sequence motifs (patterns) conserved among members of a protein family. PFAM is a database of alignments and profile-hidden markov models of protein families [Bateman, 2003]. They can predict protein functions only if sequences are sufficiently conserved. Although sequences share some structural similarities, these methods may fail, lest there is an inadequate sequence similarity. In general, they require large amount of time for building models as well as for predicting functions based on them.

The generation of reliable multiple alignments becomes problematic, while dealing with extremely diverged protein sequences. G-Protein Coupled Receptors (GPCR) belongs to highly diverged protein families and they are transmembrane
proteins. They play a significant role in various signal transmission processes, which are directly associated with a variety of human diseases [Predix, 2000].

Protein classification algorithms use multiple sequence alignment, profiles, motifs, short sequences, etc. Profiles and motifs need expert knowledge for attributing structural or functional aspects of a protein. It is possible to derive results through the existing methods based on profiles and motifs with the aid of a good expert system. The simple machine learning methods will not be helpful in the absence of expert intervention.

Machine-learning systems can be developed with

a) Expert interference (Human Computer Interface).

b) Machine learning systems by naive methods.

The naive machine learning methods build rules for decision-making. These rules may not be physically interpretable, but it doesn’t entail expert interference and provides flexibility to build the system automatically.

1.1 PROBLEM SPECIFICATION

This research aims at building rules and ensuing rule based Expert System for protein classification with the following objectives:

1. Protein Classification / Identification.
2. Search Space Reduction and Creation of the Databases.
3. Rule Extraction.

A 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. Motifs motivated the present study to consider the occurrence and frequencies of amino acids as a characteristic of a given protein. These characteristics have been used for introducing novel sets in Sequence Arithmetic (SA) module.

Profiles are position-specific scoring table that encapsulates the sequence information within complete alignments. They define the residues allowed at given positions; which positions are conserved and which degenerate; and which positions,
or regions, can tolerate insertions. The neighborhood characteristics of an amino acid in a protein have been introduced by the motivation from the concepts of profiles. The characteristics based on the neighborhood introduced in this thesis accounts for extraction of location effects, which has not been possible to acquire from Sequence Arithmetic. Therefore, the neighborhood analysis accounts for extracting spatial information.

1.2 METHODOLOGY

The methodology is predominantly based on set theory, relations, classifiers and binarization processes.

A protein is a sequence; it is a collection of characters with some frequencies. It contains character information and positional information. A primitive information system, which is the set of characters that occurs in the sequence, has been built. The absence of character is a potential for grouping sequences. A schematic information building has been proposed, developed and demonstrated, which is named as Sequence Arithmetic.

The family, class, subclass information of the sequence is wrapped to develop a classifier. A novel way of constructing decision tree has been proposed by hybridizing the concepts of Rough Sets and decision tree. This tree is called Reduct based Decision Tree, which differs from the hybrid Reduct based Decision Tree proposed by Minz [Minz, 2005]. The method of construction of the proposed decision tree and hence, extraction of rules is discussed and demonstrated in the present work.

The Sequence Arithmetic is free from arrangement of amino acids in sequence, thus the measures/decisions arrived have been insensitive to the spatial variation. For accounting the spatial aspects (in weak sense), a neighborhood of an amino acid has been introduced, and hence binarizing the neighborhood matrix develops computation of neighbor’s association of a character with other characters. A representation of this derived information is named as Binary Association Matrix (BAM). Class information of the protein is wrapped to this Information System and classifiers are developed by building Reduct based Decision Tree (RDT) in the present thesis.
To reduce the confused set of proteins further the BAM of the class has been subjected to concept lattice and hence rules have been generated. A new decision tree constructed based on Rough Sets hybridization has been validated by considering standard datasets.

A Schematic method of quantification of performance of multilevel classifier has been introduced.

1.3 CONTRIBUTIONS

The following conceptual contribution viz., i) Sequence Arithmetic, ii) Neighborhood Analysis, and iii) Binary Association Matrix have been made in the present research. A novel tool called Reduct based Decision Tree has been developed by hybridizing the Rough Sets and decision tree.

- Different sets: A-Set, D-Set and P-Set as characteristics of sequences of a family have been introduced under Sequence Arithmetic.
- Rule base developed based on the derived information related to P-Set of the sequences, which has been utilized to derive class information.
- Binary Association Matrix construction methodology designed and hence rule base derived for class/ subclass identification.
- The subclass identification rules extracted by adopting concept lattice tools on Binary Association Matrix.

1.4 OVERVIEW OF ROUGH SET PROTEIN CLASSIFIER

The Rough Set Protein Classifier is explained with a block diagram. The demonstration details are given in the sections to follow.

1.4.1 Rough Set Protein Classifier

Rough Set Protein Classifier consists of four modules, viz., i) Sequence Arithmetic module, ii) Reduct based Decision Tree module, iii) Neighborhood Associations/ Analysis module, and 4) Concept Lattice module. The Rough Set Protein Classifier process is shown in Figure.1.1 with unknown sequence ‘y’ as the input and set of sequences < S > as output.
The different modules of the Rough Set Protein Classifier are depicted in Figure 1.2. Given an unknown sequence ‘y’ to the Sequence Arithmetic module, the family information < F > is obtained by mapping with the information in Sequence Arithmetic Knowledge Database.

The family information < F > along with sequence ‘y’ is given to Reduct based Decision Tree module to get class < C > information. Reduct based Decision Tree rules have been used for extracting class information. These rules have been stored in Reduct based Decision Tree Rules Database.

Neighborhood associations have been performed to obtain localized information. This localized information has been used in the construction of Binary Association Matrix in the Neighborhood Associations module.

The Binary Association Matrix has been used in the construction of concept lattice. The node information of the concept lattice assigns the relationship between
the attributes and the objects. Set of sequences $< S >$ within a class closer to unknown sequence ‘y’ is the result of Concept Lattice module.

1.4.2 Demonstration of Rough Set Protein Classifier

G-Protein Coupled Receptors protein sequences and Myoglobin protein sequences are taken for experimentation. G-Protein Coupled Receptors consists of 3896 sequences divided into 5 classes [Brookhaven NL] of which 1041 have been taken for demonstration. Samples of 399 sequences out of available 727 sequences of Myoglobin are considered. These are divided into 36 classes. The derived database of various modules is a one-time process. They have been computed and stored for knowledge extraction (Appendices A, B and C).

The development of multilevel tool, Rough Set Protein Classifier is elucidated in Chapters III, IV, V and VI.

It has been observed that Rough Set Protein Classifier is effective in reducing domain search space.

1.5 ORGANIZATION OF THE THESIS

Chapter II deals with a thorough study of proteins, protein structures and various classification algorithms, followed by a brief overview of the data mining algorithms. The composition of proteins and their importance has been emphasized. A description of two datasets G-Protein Coupled Receptors and Myoglobin proteins that are considered for experimentation are presented. Foundations of set theory, Rough Set Theory, Decision Tree, Concept Lattice are introduced in this chapter.

Chapter III presents arrangement invariant methodology named as Sequence Arithmetic evolved based on the concept of motifs of a sequence and subsequently used for family (families) identification.

Chapter IV demonstrates a typical organization of derived information based on Sequence Arithmetic (confining to P-Sets) and is wrapped with protein decision information like family, class to result in a decision table. The process of extracting the predominant attributes (approximate reducts) is demonstrated. Hybridization of the Rough Sets (approximate reduct) with decision tree is named as Reduct based
Decision Tree (RDT) has been developed and demonstrated. Classification rules have been generated based on RDT and organized in a database called Reduct based Decision Tree Rules Database (RDTRD). These rules assist in attaching a new protein to a class within a family. The performance evaluation of RDT has been carried out on standard datasets.

Chapter V deals with extraction of localized information. To reduce the search space for a given queried protein, one needs to solicit additional information from the protein. Neighborhood metric for other amino acids has been defined and compiled in a matrix for pair of amino acids with one as center and other as neighbor for a predefined neighbor’s distance. A binarization scheme has been evolved to derive a Binary Association Matrix (BAM). A novel way of constructing the decision table has been introduced by considering a row corresponding to selected center from BAM for a protein and appending the class information to it, for a set of proteins of the training set. Rules are compiled by administering RDT on the derived decision table. The rules are organized in a database named as Neighborhood Associations Rules Database (NARD). An illustrative demonstration has been carried out on few test proteins.

Chapter VI deals with a hybridized approach of Neighborhood Associations and Concept Lattice. The Galois concept lattice tool has been implemented on BAM of a class of proteins. The output of the tool has been compiled and stored in a database called Concept lattice Association Rules Database (CARD).

Chapter VII presents the design and analysis issues associated with novel tool RSPC developed in this thesis. Performance Evaluation of RSPC has been demonstrated empirically.

Chapter VIII summarizes the contributions and presents future scope of the work.

The thesis has four Appendices A, B, C and D. Schematic compilations of databases of various levels of RSPC are present in Appendices A, B and C. Appendix D gives the Performance Evaluation of RSPC. A few samples of the rules are appended. The complete lists of rules are available in the enclosed CD.