Chapter 1

INTRODUCTION

Bioinformatics is an interdisciplinary field that integrates computer science and informatics, biology, statistics, applied mathematics, artificial intelligence, etc. to solve the biological problems at the molecular level. Application of advanced statistical and data mining techniques in the area of bioinformatics help to organize, analyze and interpret biological data and thereby prediction of unknown proteins functions. Protein function prediction is a very important and challenging task in bioinformatics and has its major application in drug discovery. The knowledge of the functionality of a protein is very important to develop new approaches in any biological process.

This chapter is organized as follows; section 1.1 presents introduction, section 1.2 presents prediction of protein function and its importance, section 1.3 describes problem description and motivation of the work, section 1.4 describes objective and contributions to the thesis and section 1.5 describes organization of the thesis.

1.1. Introduction

Protein is a large molecule composed of one or more chains of 20 amino acids in a specific order. Proteins are the main building blocks of life and required for the structure, function, and regulation of the body's cells, tissues, and organs. Each protein has unique functions such as enzymes, receptors, hormones and antibodies etc. The central dogma of life clearly describes the flow of information within the living organism. This is shown in Figure-1.1.

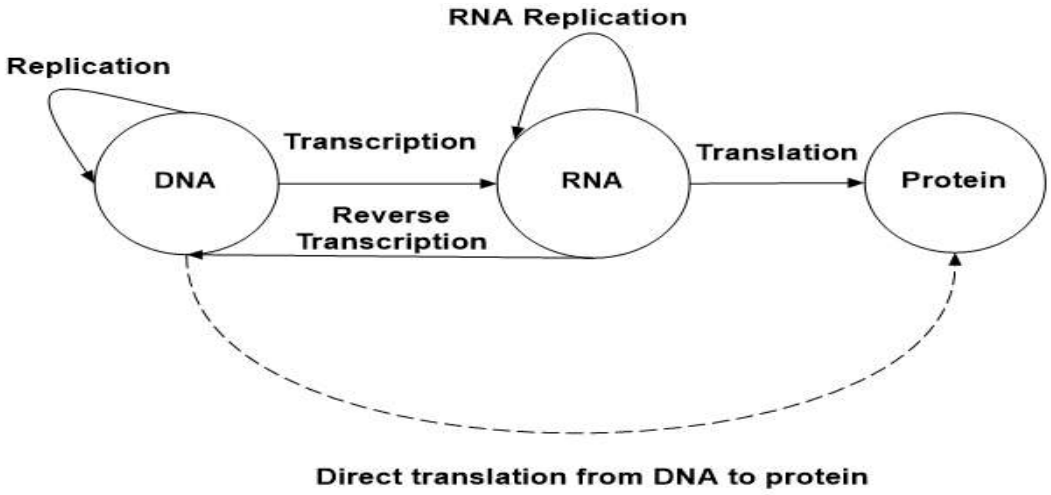

Figure 1.2 State transition diagram of Central Dogma of Life

The central dogma consists of following phases:

1. Replication: Deoxyribonucleic acid (DNA) gets duplicated by a process called replication prior to the occurrence of cell division.

2. Transcription: This is the process of conversion of DNA of chromosome to form Ribonucleic acid (RNA).

3. Translation: In this process the information available in the RNA sequence is decoded to form protein.

4. RNA replication: RNA replication is the copying of one RNA to another.

5. Reverse transcription: Reverse transcription is the transfer of information from RNA to DNA.

6. Direct translation from DNA to protein: Direct translation from DNA to protein has been demonstrated in a cell-free system (i.e. in a test tube).

The proteins are categorized into eight sub-categories according to their functionalities such as hormonal, enzymatic, structural, defensive, storage, transport, and receptors and contractile. In this thesis, three classes of proteins are considered that are ion channels, enzymes and receptors.

Ion channels are membrane proteins that are responsible for electrical signaling by gating the flow of ions across the cell membrane. These are the prominent component of nervous systems. The dysfunction of ion channels play an important role in the development

of various diseases such as hypertension, defective insulin secretion, cardiac arrhythmias, neurological diseases such as epilepsy and even developmental defects such as osteoporosis.

Enzymes are macromolecular biological catalysts that increase the rate of virtually all the chemical reactions within cells and they are indispensable to life. They are responsible for thousands of metabolic processes that sustain life and also responsible for numerous other functions, which include the storage and release of energy, the course of reproduction, the processes of respiration, and vision.

Receptors are a protein molecule usually found embedded within the plasma membrane surface of a cell that receives chemical signals from outside the cell. In this thesis, two types of receptors, nuclear receptors and G-protein coupled receptors are consider for the prediction of their families and subfamilies. Nuclear receptors are responsible for sensing steroid, thyroid hormones and other molecules. Nuclear receptor acts as a ligand activated transcription factor and have the ability to regulate gene expression by integrating with DNA sequence to control the development, metabolism, reproduction and homeostasis thus nuclear receptors are a potential drug target for the various diseases such as diabetes, cancer, osteoporosis and inflammatory diseases. G-protein coupled receptors are responsible for many physiochemical processes such as neurotransmission, metabolism, cellular growth and immune response.

The knowledge of the functionality of a protein is very important to develop new approaches in any biological process. The experimental based protein function prediction required a huge experimental work and human effort to analyze a single gene or protein. So to remove this drawback a number of very high-throughput experimental procedures have been invented to investigate the methods that are used in function prediction. These procedures have generated a large amount of protein sequences these are maintain by the various databases SWISS-PROT (Boeckmann *et al.,* 2003), universal protein resource (UniProt) (Bairoch *et al.*, 2005), national center for biotechnology information (NCBI) (Wheeler *et al.,* 2007) and protein data bank (PDB) (Bernstein, *et al.*,1997). In literature various approaches are exist for protein function prediction include homology based approaches (Söding, 2005; Schwede *et al.,* 2003), similarity based methods (Casari *et al.*, 1995; Skolnick *et al.*,2000; Cai *et al.*, 2003) and feature based approaches (Lin *et al.*, 2006; Watson *et al.,* 2005; Hua *et al.*, 2001; Petrova *et al.*, 2006). In the past, the homology based approaches were used to predict the protein function, but they failed when new proteins were

different from previous one. Therefore, to alleviate the problems associated with homology based traditional approaches, numerous computational intelligence techniques have been proposed in the recent past. The similarity based methods used the structure of a protein and it identifies the protein with most similar structure using structural alignment techniques. The feature based methods find the features of the amino acid sequence and use these features to find the function of the proteins. In the feature based approach the features are derived from the individual protein so these features are more meaningful since they are defined by the knowledge of the protein function and factors which affect a protein function. In addition to the above mentioned approaches various other computational intelligence based approaches were proposed in recent past to deal with the problem of protein function prediction such as support vector machine, Naive Bayes, neural network, k-nearest neighbor classifier, decision trees and random forest etc.

1.2. Prediction of protein functions and its importance

If we are to understand life at a molecular level we must understand how proteins carry out their function. This is also important for understanding the molecular mechanisms of disease, because alterations of protein function are responsible for many diseases. So it is necessary to classify an unknown protein sequence into their families and subfamilies to avoid the expensive experiment at the laboratory. Once a particular sequence *S*, causing a disease *D*, is classified into its family *F*, the researchers can design some new drugs by trying some combination of existing drugs for family F. Thus, this classification problem helps the researchers for treatment of diseases by discovering new drugs. The major application area of the protein function prediction is the drug discovery. If a newly discovered protein, responsible for the cause of a disease gets correctly classified to its families the task of the drug analyst becomes simpler for drug discovery.

1.3. Problem description and motivation of the work

Protein function prediction is the most challenging problem in Bioinformatics due to massive growth of knowledge of unknown proteins with the advancement of high throughput microarray technologies. In the past, the homology based approaches were used to predict the protein function, but they failed when a new protein was different from the previous one. Therefore, to alleviate the problems associated with homology based traditional approaches it

is necessary to design efficient and robust computational intelligence techniques based approaches for the prediction of protein function. There are three main functions of proteins which include catalysis of biological reactions known as enzymes, structuring the organs and membrane proteins that include receptors and ion channels. The correct prediction of enzymes, ion channels and receptors plays a very important role in various application areas such as drug discovery, disease detection etc. Also the prediction of above mentioned protein functions are not easier with available computational intelligence methods, other related algorithms and software tools. Therefore, there is a need for robust and efficient computational intelligence techniques to address the above mentioned problems.

The work presented in this thesis investigates the various available methods in literature for computational intelligence techniques applied to this domain and proposes new efficient and robust approaches for protein function prediction specifically for the prediction of enzymes, receptors and ion channels.

Here, the problem of protein function prediction may be considered as a pattern classification problem. The protein is represented by a chain of amino acids sequences through which various sequence derived properties are extracted. To obtain most important and optimal number of features the various filter, wrapper and hybrid based feature selection approaches are used to improve the prediction performance of protein function.

Proteins are the cause of many diseases so the knowledge of the functionality of a protein is very important to develop new approaches in any biological process. If a newly discovered protein gets correctly classified to its families and their subfamilies, then the task becomes easy for the drug analyst to discover new drugs. So it is very important and challenging to design efficient and robust approaches to predict protein function using sequence derived properties. Here, the main motivations for the work presented in this thesis are as follows:

1. To study the problem of protein function prediction and need for the design and development of efficient and robust computational intelligence based techniques for protein function prediction.

2. To present the literature review of the methodology used for the design and development of the computational intelligence techniques.

3. Identifying the advances and limitations of the existing techniques for the overall design and development to the computational intelligence techniques.

1.3.1. General Framework of the computational intelligence techniques used for protein function prediction

The major four steps that are involved in the design and development of an efficient and robust computational intelligence technique for protein function prediction is shown in Figure1.2.

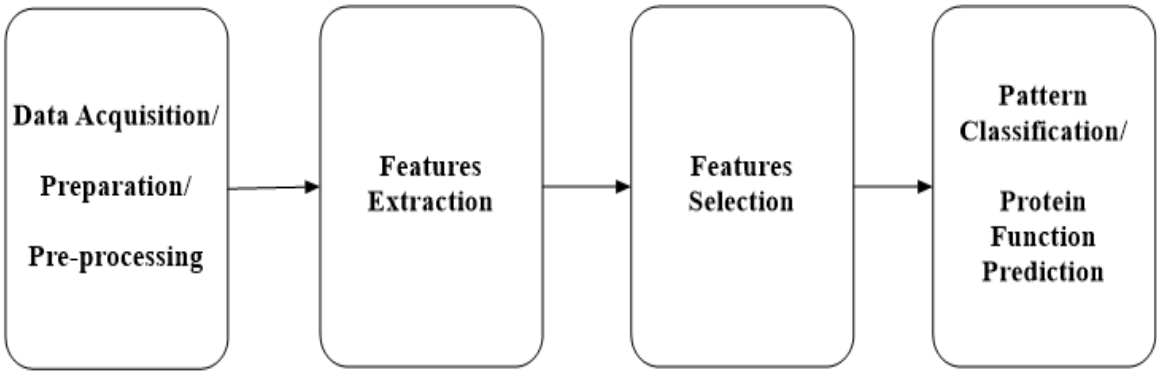

Figure1.2: General framework of protein function prediction

The various steps involved in the design and development of an efficient and robust computational intelligence technique for protein function prediction include: data acquisition and preparation/data pre-processing, feature extraction, feature selection and prediction/classification of protein functions. The brief descriptions of these steps are as follows:

1. **Data acquisition and preparation/data pre-processing**: The first step of protein function prediction is the acquisition of protein sequences from standard repository of protein sequences such as Protein Data Bank (PDB), NCBI, and SWISS-PROT etc. The protein sequences may be incomplete, redundant, noisy and inconsistent. In data preprocessing, the data having missing values are filled, smoothing of noisy data are done and inconsistencies are resolved.

2. **Feature extraction**: In this phase various feature vectors that represent the protein sample, including amino acid composition, dipeptide composition, correlation factors, composition, transition, distribution of physiochemical properties, sequence order descriptors and pseudo amino acid composition etc. are extracted to fully characterize protein sequences.

3. **Feature selection**: In this phase various feature selection methods such as Fisher score based feature selection, ReliefF, fast correlation based filter, minimum redundancy and maximum relevancy, principal component analysis and support vector machine based recursive feature elimination etc. are used to obtain optimal number of features.

4. **Prediction/classification of protein functions**: This phase includes design and development of computational intelligence techniques to predict the protein function. The computational intelligence techniques are train by using appropriate parameters and input data and construct a prediction model to predict the protein function. The validation of the proposed prediction model is performed by using test data for the performance evaluation of the model.

1.4. Objective and contributions to the thesis

The major objective of the present work is to develop efficient and robust computational intelligence techniques for protein function prediction. The success of design and development of efficient and robust computational intelligence techniques relies on the design and development of an appropriate feature extraction, feature selection and pattern classification techniques for the said task.

In this thesis, following associated problems of protein function prediction are investigated

1. Classification of ion channels and their types.
2. Classification of enzymes function.
3. Classification of nuclear receptors and their subfamilies.
4. Classification of G-protein coupled receptors and their subfamilies.

To address the above mentioned problems following computational intelligence techniques are proposed:

1. A multi stage approach for the prediction of ion channels and their subfamilies based on random forest with minimum redundant and maximum relevant sequence derived features.
2. A top down approach to classify enzyme functional classes and subclasses using rotation random forest.

3. An efficient approach for prediction of nuclear receptor and their subfamilies based on fuzzy k-nearest neighbor with maximum relevance minimum redundancy.

4. An efficient and robust approach for the prediction of G-protein coupled receptors and their subfamilies using weighted k-nearest neighbor.

In order to proceed with the task for the prediction for the families and subfamilies of a protein, a detailed investigation was done on various features extracted from amino acid sequence and the various classifiers implemented by earlier researchers.

The major contributions in this thesis are to proposed efficient and robust approaches to predict the functional classes and sub-classes of ion channels, enzymes, nuclear and G-protein coupled receptors using sequence derived properties features vectors such as amino acid composition, dipeptide composition, correlation features, composition, transition, distribution, sequence order descriptors and pseudo amino acid compositions. In this thesis various feature selection methods such as principal component analysis (Wold *et al.*, 1987), Fisher score based feature selection (Guyon *et al.*, 2003), ReliefF (Kira *et al.*, 1992), fast correlation based filter (FCBF) (Yu *et al.,* 2003), minimum redundancy and maximum relevancy (MRMR) (Peng *et al.*, 2005), and support vector machine based recursive feature elimination (SVM-RFE) (Guyon *et al.,* 2002) are also described to obtain a relevant, non-redundant, and robust feature subset.

1.5. Organization of the thesis

Throughout this thesis the main objective is to design and develop efficient and robust computational intelligence based approaches for protein function prediction. The organization of the thesis is as follows:

Chapter1 introduced the basic concepts related to protein function and its importance. The problem description with general framework for protein function prediction and motivation of the work are presented in this chapter. The objective of thesis are described and contributions to the thesis is presented in this chapter. Last section listed the organization of the thesis that describes the coverage of chapters in the thesis.

Chapter 2 presents the theoretical background related to protein function prediction. It presents the literature reviews for the computational intelligence techniques used in

prediction of ion channels, enzymes, nuclear and G-protein coupled receptors. The features extracted from protein sequences that are used in the prediction of protein function are described in this chapter. The basic concepts related to feature selection techniques such as filter, wrapper and hybrid methods and various computational intelligence techniques such as artificial neural network, Naive Bayes classifier, support vector machine, k-nearest-neighbor, decision trees, bagging, boosting, random subspace methods and random forests and are presented. In the last section the performance measures of the classifier are presented.

In chapter 3, a random forest based approach is proposed to predict ion channels families and their subfamilies by using sequence derived features. Here, seven feature vectors are used to represent the protein samples including amino acid composition, dipeptide composition, correlation features, composition, transition, distribution and pseudo amino acid composition. The minimum redundancy and maximum relevance feature selection is used to find the optimal number of features for improving the prediction performance.

In chapter 4, a rotation random forest based ensemble classifier is proposed to predict the functional classes and sub-classes of enzymes by using sequence derived features. Here, seven feature vectors are used to represent the protein sample, including amino acid composition, dipeptide composition, correlation features, composition, transition, distribution and pseudo amino acid composition. The proposed method used 10-fold cross validation to predict the enzymes functional classes and subclasses.

In chapter 5, a fuzzy k-nearest neighbor classifier with minimum redundancy maximum relevance is proposed for the classification of nuclear receptor and their eight subfamilies. The minimum redundancy maximum relevance algorithm is used to select the optimal feature subset. From the obtained results and analysis it is observed that the performance of the proposed approach for the classification of nuclear receptors and their eight subfamilies is very competitive with some other standard methods available in literature.

In chapter 6, a method for the prediction of G-protien coupled receptors is proposed. To address the issues of efficient classification of G-protein coupled receptors and their subfamilies, we propose to use a weighted k-nearest neighbor classifier with UNION of best 50 features selected by Fisher score based feature selection, ReliefF, fast correlation based filter, minimum redundancy maximum relevancy and support vector machine based recursive feature elimination (SVM-RFE) feature selection methods to exploit the advantages of these

feature selection methods. The proposed method used 10-fold cross validation to predict the G-protein coupled receptors and their subfamilies.

Finally chapter 7 presents conclusions and future scope of the work. It describes the usefulness of the efficient and robust approaches for protein function prediction. It also presents the future scope of the work.