# PREFACE

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 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.

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.

Protein is a chain of 20 amino acids in a specific order and performs various functions. 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, ion channels and antibodies etc. The correct prediction of these proteins plays a very important role in various application areas such as drug discovery, disease detection and many more. 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.

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. Hence it is very important and challenging to design efficient and robust approaches to predict protein function using sequence derived properties.

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 ion channels, enzymes and receptors.

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 functional classes and subclasses.
3. Classification of nuclear receptors and their subfamilies.
4. Classification of G-protein coupled receptors and their subfamilies.

To address the above mentioned problem 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 receptors 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.

To examine the efficacy and usefulness of above mentioned proposed approaches an appropriate set of features were extracted including eight feature 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 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 are also described to obtain a relevant, non-redundant, and robust feature subset. The accuracy, Mathew Correlation Coefficient (MCC), precision and ROC area are used to measure the performance of the proposed efficient and robust approaches. The overall thesis is organized into seven chapters. The abstract of each chapter are given 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 is described and contributions to the thesis are presented in this chapter. Last section listed the organization of the thesis that describes the coverage of chapter 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 also 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 method and random forests are presented. In the last section of the chapter 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 sample, 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. The proposed method achieved an overall accuracy of 100%, 98.01%, 91.5%, 93.0%, 92.2%, 78.6%, 95.5%, 84.9%, MCC values of 1.00, 0.92, 0.88, 0.88, 0.90, 0.79, 0.91, 0.81 and ROC area values of 1.00, 0.99, 0.99, 0.99, 0.99, 0.95, 0.99 and 0.96 using 10-fold cross validation to predict the ion channels and non-ion channels, voltage gated ion channels and ligand gated ion channels, four subfamilies (calcium, potassium, sodium and chloride) of voltage gated ion channels, and four subfamilies of ligand gated ion channels and predict subfamilies of voltage gated calcium, potassium, sodium and chloride ion channels respectively.

In chapter 4, an ensemble classifier rotation random forest is proposed to predict the functional classes and subclasses 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 achieved an overall accuracy of 100%, 88.7%, 87.6% , MCC values of 1.00, 0.87, 0.88, ROC area values of 1.00, 0.98, 0.98 and precision of 100%, 89.5% and 88.8% using 10-fold cross validation to predict the enzymes and non-enzymes, functional classes and subclasses of enzymes respectively.

In chapter 5, a fuzzy k- nearest neighbor classifier with minimum redundancy maximum relevance is proposed for the classification of nuclear receptors and their eight subfamilies. The minimum redundancy maximum relevance algorithm is used to select the optimal feature subset and observed that highest accuracy and Matthew's correlation coefficient is obtained with 400 features among 857 features through fuzzy k-NN classifier. The performance of fuzzy k-NN classifier depends on two parameters, number of nearest neighbor (k) and fuzzy coefficient (m) and it is observed that the highest accuracy and MCC is obtained at k=7 and m= 1.25. The overall accuracies of 10-fold cross validation with optimal number of features, k and m are 98.09% and 97.85% and the MCC values of 0.97 and 0.90 for the prediction of nuclear receptors and their subfamilies respectively. 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-protein 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 achieved an overall accuracy of 99.9%, 98.3%, 95.4%, MCC values of 1.00, 0.98, 0.95, ROC area values of 1.00, 0.998, 0.996 and precision of 99.9%, 98.3% and 95.5% using 10-fold cross validation to predict the G-protein coupled receptors and non- G-protein coupled receptors, subfamilies of G-protein coupled receptors and subfamilies of class A G-protein coupled receptors respectively.

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.