Chapter 7

CONCLUSION AND SCOPE FOR FUTURE WORK

This chapter presents the conclusion of the thesis. This chapter is organized in two sections. Section 7.1 presents concluding remarks and section 7.2 discuss the possible scope for future works.

7.1. Concluding remarks

The research contributions and research achievements of this thesis are as follows:

Chapter 1 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 were presented in this chapter. The objective of thesis was described and contributions to the thesis were presented in this chapter. Last section listed the organization of the thesis that described the coverage of chapter in the thesis

Chapter 2 presented the theoretical background related to protein function prediction. It presented the literature reviews for the computational intelligence techniques used in prediction of ion channels, enzymes, nuclear and G-protein coupled receptors and their subfamilies. The features extracted from protein sequences that were used in the prediction of protein function were 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, and decision trees bagging, boosting, random subspace method and random forests were presented. In the last section the performance measures of the classifier were presented.

Chapter 3 proposed a random forest based approach to predict ion channels and their subfamilies by using sequence derived features. The minimum redundancy and maximum relevance feature selection was used to find the optimal number of features for improving the prediction performance. The results showed that the MRMR feature selection algorithm reduces the number of input feature vectors by selecting the important features and improve the overall accuracy and MCC. In the 10-fold cross validation the proposed method had 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 to predict ion channels and non-ion channels, voltage gated ion channels and ligand gated ion channels, four types (calcium, potassium, sodium and chloride) of voltage gated ion channels, ligand gated ion channels and predict subfamilies of voltage gated calcium, potassium, sodium and chloride ion channels respectively. The high accuracies, MCC and ROC area values indicate that the proposed method may be useful for the prediction of ion channels families and their subfamilies.

Chapter 4 proposed an ensemble classifier Rotation random Forest based of rotation forest and random forest to predict the functional classes and sub-classes of enzymes by using sequence derived features. Here, seven feature vectors were used to represent the protein sample, including amino acid composition, dipeptide composition, correlation features, composition, transition, distribution and pseudo amino acid composition. In the 10-fold cross validation the proposed method had 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% to predict the enzymes and non-enzymes, functional classes and subclasses of enzymes respectively. The high accuracies, MCC, ROC area and precision values indicate that the proposed

method is useful for the prediction of functional classes and subclasses of enzymes.

Chapter 5 proposed to use a fuzzy k- nearest neighbor classifier with minimum redundancy maximum relevance for the classification of nuclear receptors and their eight subfamilies. Seven feature vectors are used to represent the protein sample, including amino acid composition, dipeptide composition, correlation, composition, transition, and pseudo amino acid composition. To improve the prediction performance the minimum redundancy maximum relevance algorithm has used to select the optimal feature subset. The values of two parameter k and m have been determined to predict nuclear receptors and their subfamilies 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 MCC values of 0.97 and 0.90 for the prediction of nuclear receptor families and subfamilies respectively. From the obtained results and analysis it was observed that the performance of the proposed approach for the classification of nuclear receptors and their eight subfamilies is performing better than some other standard methods available in literature.

Chapter 6 proposed an optimal feature subset selection method which provides the non-redundant, relevant, and robust feature subset by taking UNION of best 50 features selected by various supervised feature selection methods such as Fisher score based feature selection, ReliefF, FCBF, MRMR and SVM-RFE feature selection methods. In next stage, we proposed to use a weighted k-nearest neighbor classifier to predict the G-protein coupled receptors and their subfamilies. Using the 10-fold cross validation 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% 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. The high accuracies, MCC, ROC area values and precision values indicate that the proposed method is better for the prediction of G-protein coupled receptors families and their subfamilies.

Finally, the overall conclusion of the thesis is being summarized as follows:

In this thesis, following associated problems of protein function prediction were 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.

Further, a detail review was presented and the related problems were investigated. To address the related problems of protein function prediction the following computational intelligence techniques were proposed, implemented and compared with other existing methods in literature.

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.

The obtained experimental results and performance analysis for each of the above mentioned methods and their comparative analysis with other existing methods in literature demonstrate the applicability of these methods. Further,

all above proposed methods are efficient, robust and performing better in comparison to other standard methods in literature

7.2. Scope for Future Works

The research work presented in this thesis can be taken further into different directions. The scope for future works is as follows.

In this thesis, we presented an efficient and robust feature selection and classification approaches for the prediction of functional classes and subclasses of ion channels, enzymes, nuclear receptor and G-protein coupled receptors. These proposed approaches are useful for the researcher and drug analyst to design and discovery of new drugs related to these protein families such as neurotransmitter related , metabolic problems, growth related and brain related diseases. For further enhancing the efficacy and accuracy of the proposed method, in addition to using only sequence derived properties of protein sequences used in this work, one may also investigate the possibility of exploring and using the other features such as structural properties, protein-protein interaction and gene expression data to predict these protein functions. The integration of data with integration of various classifiers may also be useful in protein function prediction.

The other problems related to the field of the study which may further be investigated are as follows:

1. Protein function prediction by using protein-protein interaction data.
2. Analysis of gene expression data by using various feature selection, classification and clustering based approaches.
3. Analysis of pathways by using gene-gene interaction for the disease prediction.