Preface

Knowledge of a protein's 3D structure is essential in complete understanding of its function in atomistic detail. However, in this post-genomics era, the number of proteins with only sequence information has significantly outgrown the number of proteins with both sequence and experimentally determined structural information. This has lead to a greater need for accurate computational prediction of protein structures than ever before. Among the structure prediction approaches protein fold recognition refers to a method of assigning the most probable/compatible structural fold out of the known structural folds for a given protein sequence. Protein fold recognition methods use information collated from known protein structures. Hence, these methods offer powerful and accurate means to detect structural homologs that are otherwise difficult to detect by conventional sequence-based homology search methods. Protein fold recognition which essentially assigns one of the known folds to a new protein sequence derives its strength from the fact that the number of unique folds in nature is limited. It has been estimated that number of naturally occurring protein folds is somewhere around 2000 and currently the experimentally derived protein structures can be collated and clustered to about 1200 folds. Furthermore, it is now believed that what are not known pertain to only the sparsely populous folds more specifically the orphan folds characterized by the presence

of solitary proteins. Therefore it is reasonable to expect a greater chance for a new protein to adopt one of the known protein folds than one of the unknown novel folds.

A number of protein fold recognition methods have been developed so far, however, their prediction accuracies have been far from satisfactory. This could be attributed to the lack of well-characterized data, less exploitation of feature properties and limitations of mathematical/statistical methods used.

The currently available and commonly used methods for assigning folds to the protein sequences can be broadly classified into three categories: (a) Sequence-structure homology recognition methods, (b) Threading methods and (c) Taxonomic methods. Among these approaches, the taxonomic methods achieve highest prediction accuracies but are limited by the low fold-coverage. Therefore, in the present study we have developed a new method of protein fold recognition with high prediction accuracy and high fold-coverage by investigating the fold-discriminatory potential of some new sequence- and structure-based features in Support Vector Machine (SVM)-based set up and further showed its immense use in predicting protein folds. The work carried out has been organized and presented in six chapters preceded by the first introductory review of the related literature and succeeded by the last conclusions chapter.

Chapter 1 gives an introductory review of all the relevant literature pertaining to the work presented in the thesis. It begins with a brief overview of the importance of protein structural information, an overall introduction to protein structure including the

methods of protein structure determination, protein structure prediction and structural classification of proteins. This is followed by a brief overview of the existing fold recognition methods. Finally the chapter introduces the objectives of the present work.

Chapter 2 describes the details of our investigations performed to check the fold-discriminatory potentials of various sequence- and structure-based features and develop a new SVM-based method for protein fold recognition. Of the various features investigated structural information of amino acid residues and amino acid residue pairs viz., secondary structural state frequencies with solvent accessibility state frequencies of amino acids and amino acid pairs gave rise to the best fold-discrimination and therefore they were used as feature vectors for training SVMs. Twofold as well as fivefold cross-validation studies using a standard benchmark test datasets revealed that the new method so developed outperforms all other available methods.

The SVM, which has been used as the classifier, is basically a binary classifier while the protein fold recognition is a multi-class problem. Therefore, to use SVM for protein fold recognition, three multi-class classification methods, namely *one versus all*, *one versus one*, and *Crammer and Singer* method were used and it was found that they yield similar predictions. To the best of our knowledge, this is first instance that an *all-together* multi-class method known as *Crammer and Singer* method has been used for protein fold classification. Further, we also made a performance comparison of SVM with the naïve Bayes classifier which revealed better performance by SVM.

Once the superiority of the features for protein fold discrimination was established using a standard benchmark dataset, the studies were carried out to explore ways to increase the fold-coverage of the new method. Chapter 3 describes the studies pertaining to the ways explored to enhance the fold-coverage of the new method. The fold-coverage can be increased by increasing the threshold for pair-wise sequence identity of proteins to be included into the dataset as well as by minimizing the number of protein domains represented in a fold. By doing so, about 700 folds became available for training and testing. Furthermore, we also investigated the effect of related sequences of varying pair-wise sequence identities on prediction performance which revealed that the prediction accuracies, sensitivities and specificities increase with increase in pair-wise sequence identity cutoff values in the dataset. However, it was also observed that prediction accuracy decreases with the increase in the number of folds for SVM training. This is because of the fact that as the number of folds increases the classification becomes more and more complex. To reduce the effect of large-scale classification, hierarchical approach for protein fold recognition was adopted, in which first structural class of the protein is predicted followed by fold prediction within the predicted structural class.

Hierarchical approach requires a method for protein structural class prediction and thus a new SVM-based method was also developed for protein structural class prediction. The details of the development of structural class prediction are described in Chapter 4. The new method uses the combination of secondary structural content and

secondary structural state frequencies of amino acids as discriminatory features and gives rise to the best protein structural class discrimination. The prediction accuracy obtained is similar to the best accuracy reported in the literature so far.

Chapter 5 describes the details of the studies pertaining to the hierarchical scheme for protein fold classification. Hierarchical approach resulted in better prediction accuracy and helped in increasing the fold coverage. The new SVM-based method named as HPFP (Hierarchical Protein Fold Prediction) outperforms other available methods of protein fold recognition and, therefore, can be used for assignment of 3D folds to the proteins discovered in various genomes and hence can serve as an invaluable annotation tool in structural genomics. Keeping in the view the importance and applicability, HPFP was implemented into a web server and details of the web server implementation are given in Chapter 6.

Having developed the most accurate fold recognition method we resorted to use it for a genome-wide protein fold prediction. For this we chose much studied *Mycobacterium tuberculosis* genome. Chapter 7 describes the details of the structural assignment of proteins encoded by this genome by HPFP. The assignment of 3D folds revealed that all the folds are not represented equally in the genome. The assignments were compared with ModBase assignments which include the proteins structurally characterized by the structural genomics consortium and a substantial level of agreement was observed between the two assignments.

The thesis ends with Conclusions Chapter 8 which gives a concise recollection of the objectives set out during the thesis work along with the important mile-stones crossed, observations made and conclusions reached.

Most of the work presented in this thesis has either been published or written in the form of manuscripts for communication. The list of publications is given on Page. The figures and tables referred throughout this thesis are numbered chapter-wise. The references cited in this thesis have been given collectively at the end and are arranged alphabetically. The secondary structure predictions were done using PSIPRED (McGuffin et al., 2000) and solvent accessibility predictions were done using ACCpro (Cheng et al., 2005). The SVM computations were done using LIBSVM (Chang and Lin, 2001). The typesetting of this thesis including figures and tables was done using Microsoft Office.