(http://www.imtech.res.in/raghava/copid/) to assist the researchers in annotating function of a protein from its composition using whole or partial protein sequence. The COPid ~~have~~ has three modules: search, composition and analysis. The composition module allows calculation of the composition of a sequence and average composition of a group of sequences. The composition module also allows computing composition of various types of amino acids like (e.g. charge, polar, hydrophobic residues). The analysis module provides options like i) comparing composition of two classes of proteins, ii) creating phylogenetic tree based on the composition and iii) generating input patterns for machine learning techniques.

The structure of a protein is responsible for its function, thus determination of protein structure is important to understand its function. The experimental structural elucidation techniques such as X-ray crystallography or NMR are time consuming, costly and not possible for all proteins. This is the reason that only limited unique structures are solved in last fifty years. The problem of protein structure prediction has been approached through three main routes a) computer simulation based on empirical energy calculations; b) knowledge based approaches using information derived from structure-sequence relationships from experimentally determined protein 3D structures; and c) hierarchical methods which involves construction of a model using secondary structure information from amino acid sequence data, which is eventually used to predict the tertiary structure. Till now different computational methods have been developed for predicting secondary structure states using amino acid sequence. But the transition structure states between secondary and tertiary structure, such as super-secondary structure motifs, is not fully explored. Hence in this thesis, a method Bhairpred (http://www.imtech.res.in/raghava/bhairped) developed to predict super-secondary structure motif β-hairpins using two machine learning techniques ANN and SVM. The performance was evaluated in both ideal (experimentally determined attributes) and real life situation (predicted attributes) along with independent evaluation on CASP6 target proteins. On analysis, it was found that quality of Bhairpred predictions depends on the PSIPRED secondary structure prediction.

Over the years a large number of methods have been developed for predicting single or multiple subcellular localization of protein. Though these methods have higher overall accuracy but prediction performance for certain location is poor such as

mitochondrial proteins. Thus in this thesis attempt has been made to developed location specific method for mitochondrial proteins (MitPred; http://www.imtech.res.in/raghava/mitpred). MitPred is a hybrid methods which predicts mitochondrial proteins on the basis of Pfam domain occurrence and SVM model. Here, first time the concept of exclusive domain occurrence in a protein was used. Two types of domain libraries were created. First library contains domains, which are found only in mitochondria. Second library contains Pfam domains that are not found in mitochondria. The prediction was done on the basis of presence or absence of library domain in the query protein. In case no exclusive domain was found then SVM based prediction was used. Similarly using the strategy of MitPred, a method was also developed for another important and abundant class of proteins, nuclear proteins NpPred; http://www.imtech.res.in/raghava/nppred). These two new methods were used to annotate five eukaryotic proteomes (*S. cerevisiae, C.elegans, D. melanogaster,* mouse and human) for prediction of mitochondrial and nuclear proteins.

DNA and RNA interacting proteins play very important role in cellular metabolism along with regulation of gene expression. Most of the DNA-binding proteins prediction methods are developed using PDB protein chain data. Hence, they are not appropriate for high throughput genome annotation. Thus, we developed a method DNAbinder (http://www.imtech.res.in/raghava/dnabinder) that can predict both DNA-interacting protein chains and full length proteins. PSSM based SVM showed the best performance along with a significant positive correlation with the number of homologous sequences used during construction of PSSM. It means that presence of large number of homologous sequences improves the prediction quality. In addition, we also developed method for predicting RNA-binding proteins (RNApred; http://www.imtech.res.in/raghava/rnapred) using similar approaches adopted for DNA-binding protein prediction. On benchmarking both methods showed better performance than other existing methods.

Since all functions of DNA and RNA-binding proteins are mediated though nucleotide-protein interaction. Hence the prediction of DNA and RNA interacting residues is as important as prediction of these proteins. Further, experimental determination of these residues require crystal or NMR structure of protein which itself is a big problem. Hence we also developed sequence based methods for

prediction of DNA- and RNA-binding residues, namely PPRInt (http://www.imtech.res.in/raghava/pprint) and (b) DNAint (http://www.imtech.res.in/raghava/dnaint) respectively. These methods can be useful in identification of DNA- and RNA-binding sites or prediction of potential binding motifs on a protein.

In a nutshell, ~~during~~ the work described in this thesis ~~work~~ attempts ~~have been made~~ to develop methods which may help biologist directly or indirectly in assigning function to newly sequenced proteins. In addition, the proteomes of few important organisms have been annotated using the methods developed in this thesis. Overall methods developed in this thesis will be useful for bioinformaticains and biologist involved in functional annotation of proteins.

An important caveat in automatic prediction of protein function is that the task is a multi-class problem: that is to say an entity can reside in more than one class. For example, a protein can be both nuclear and cytosolic or both a helicase and ligase. In addition, these classes can be of widely different size. These factors can make evaluation of the performance of classification methods difficult. In ontologies things can be complicated by the fact that classes are arranged in a directed acyclic graph rather than a tree. It means that a certain term can be assigned to a protein via two or more different paths through ontology. An example is the GO "mating class", which is under both 'developmental process" and "cell growth and/or maintenance'. It is often very difficult to judge one classifier to be better than another under slight variations of the text data or performance matrix. Thus there is need to develop methods for predicting multiple locations or functions of a protein. The tool developed in this thesis is a small step in right direction. Still, there is need to build comprehensive and accurate tools for predicting all functions of a protein from its sequence.