Mamoon Rashid (2010). *Development of Bioinformatics tools for analysis and prediction of Therapeutic peptides/proteins.* Ph.D. Thesis. CSIR-IMTECH, Chandigarh/ Jawaharlal Nehru University, New Delhi: India.
Supervisor: Dr. G.P.S.Raghava
Acc. No. : Th-241

11 SUMMARY AND FUTURE PROSPECTS

11.1 Summary

The past decade has seen the inflow of voluminous data from various genome sequencing projects across the kingdom of life. Extraction of certain information, relevant to perpetuation of the life processes, from these sequences is one of the daunting challenges in the post-genomic era. Thus, use of these strings of coded information for unraveling the whole biological phenomena needs careful inspection of the genome or proteome of an organism. Since, the amount of data is far beyond the capacity of human inspection, the bioinformatics approaches offer automated, fast, reliable and meaningful methods to handle the situation.

Subcellular localization of protein sequences has been correlated well with the functional and structural properties of the proteins (Andrade et al. 1998). Also, location and function of a protein seem to determine its amino acid composition and folding type (Cedano et al. 1997; Nishikawa and Ooi 1982). Moreover, functions of 48% of the predicted 3995 proteins of *Mycobacterium tuberculosis* H37Rv (Mtb) are yet to be assigned (Camus et al. 2002). In this study we have developed a method for prediction of subcellular location of Mtb proteins using the protein sequences. Various machine learning-based techniques (like SVM, HMM, MEME/MAST) have been employed to classify the location of Mtb proteins. The sequence-based descriptors were meaningful in predicting the subcellular location. Then, this prediction model has been applied to annotate the proteome of Mtb in terms of subcellular location. The results might help in identifying the probable drug or vaccine targets against this dreadful pathogen.

The hormones and neurotransmitters play an important role in cellular signaling and regulatory processes and have been exploited as therapeutics (Leader et al. 2008). The collection and compilation of these molecules with their biological targets would be very useful for the scientific community. There are several databases of their kind available for extracting information about these molecules. But, a single comprehensive platform hosting the hormones and their receptors with their physical and chemical properties was not available till recently. In order to complement existing databases in the field, and to understand hormones and their interaction with receptors, we have developed a database called Hmrbase.

This database provides comprehensive information about hormones and receptors. Various data fields like hormone precursor, subcellular localization, post-translational modification, taxonomy, source organism, function, description, tissue specificity, molecular weight, similarity to other proteins, and mapping of hormone peptide on its corresponding precursor etc. have been included for peptide hormones and their receptors. For non-peptide hormones, the data fields consist of their names, molecular weights and molecular formulae, IUPAC names, canonical and isomeric smile formulae, melting points, LogP values, water solubility, and their corresponding receptors etc. Various co-ordinate files such as PDB, SDF, and MOL files are available for download. Hopefully, Hmrbase will be useful for the researchers working in the area of biomedical sciences, particularly endocrinology.

One of the major limitations that preclude peptides to be used as therapeutics is their susceptibility to proteolysis (Sato et al. 2006). As soon as they enter inside the body they are quickly cleared from the body in general, even more frustrating, are digested rapidly after oral uptake. In this study we have tried to develop a prediction model for half-life estimation of peptides in the complex proteolytic environment. This model is based on the sequence features of amino acid string of corresponding peptide dataset. We used regression model to train the peptide data with its measured half-life from high-throughput assays in complex proteolytic condition. The Pearson correlation coefficient between real and predicted half-life values came to be 0.911 using amino acid compositions.

Since interactome of an organism explains the intricate biology and dissect complex molecular function driving the life processes, it is of utmost importance to identify those interactions. Experimental identification of protein-protein interactions is labor and cost intensive affair. There comes the computational method for expanding and complementing these experimentally determined PPIs to cover the entire interactome space in an organism (Pazos and Valencia 2002; Salwinski and Eisenberg 2003; Shoemaker and Panchenko 2007b; Yellaboina et al. 2007; L. Zhang et al. 2004). In this thesis work, different algorithms have been developed for predicting PPIs. A sequence-based PPI prediction method using protein interaction data of *E. coli, S. cerevisiae,* and *H. pylori* has been proposed. Later on, genome-wide PPIs have been predicted for Mtb using the structural data available in databases like PDB, iPFAM, and 3did. An analysis of Mtb interactome was presented which facilitated retrieval of key aspects of Mtb biology. Further, integration of gene expression and gene

ontology data improved the prediction performance. It is anticipated that careful exploration of this interactome might result into a promising drug target for Mtb.

Protein-protein interactions (PPIs) are modular in nature. Hierarchy exists in this modularity, i.e. we characterize PPIs at the levels of protein, domain, and/or sub-domain. Our understanding and resolution into the mechanism of PPI improve faster as we move down this hierarchy from proteins to domains to the residues contributing in interaction. Most of the methods for predicting DDIs require PPI datasets to predict DDIs. The coverage of protein interaction networks are incomplete (Itzhaki et al. 2006; M. Liu et al. 2009), so the DDI prediction methods founded on these networks may not cover the whole space of interacting domain pairs. Moreover, there exists a marked conservation in DDI across species (Itzhaki et al. 2006). The large amount of crystallographically solved DDI data available in 3did and iPFAM databases demands a robust supervised method to train a model from these examples. This model would be capable of covering those DDIs missed by other methods relying on PPI data. In the present study, we attempted to develop prediction model learning from interacting (positive dataset) and non-interacting (negative dataset) domain pairs. We used a highly curated negative dataset for model development. SVM model has been optimized in a more realistic way using a naïve validation set to check over optimization problem. We also showed that our method performed well on independent dataset. To test the validity and wide coverage of our model, we tried to predict various datasets of predicted DDIs on our model.