Summary and Future Prospects

12.1 Summary

The overall objective of this work has to develop improved and novel prediction methods for identifying potential candidates for subunit vaccine design. The potential vaccine candidates are fragments of antigenic sequences that can trigger immune response. These fragments are known as T cell (CTL or Th) epitopes in case of cell-mediated immunity and are integral part of subunit vaccine design. The generation of CTL epitopes from endogenous antigenic proteins depends on the specificity of many intermediate steps/processes such as production of peptide fragments by proteasome, transport of these peptides to endoplasmic reticulum (ER) through TAP transporter and binding of peptides to MHC class I molecules inside ER. On the other hand, generation of helper T cell epitopes mainly involve degradation of antigenic proteins into fragments by lysosome and subsequent binding of these peptides to MHC class II molecules. The experimental analysis about intracellular processes involved in immune response has increased our understanding of specificity of various events. The experimental analysis of all these intracellular processes to dig out antigenic regions from genomic data is next to impossible as it would involve a lot of time and labour intensive wet experimentation. The management and analysis of immunological data through various computational tools have given birth to a new displine "Immunoinformatics". Immunoinformatics deals with designing of new prediction methods to model various antigen processing events and management of data generated by immunological experiments. The computational approach can provide a good alternative to experimental analysis. On the basis of this analysis, rules can be derived to model these processes and develop methods for assisting subunit vaccine development.

Before designing a new or more accurate method for various exogenous and endogenous antigen-processing events such as MHC binding, proteasomal cleavages or TAP binding and translocation, an ample amount of high quality data is a pre-requisite. Keeping this in mind, we have collected data about peptides involved in cell-mediated immunity that includes MHC binders, non-binders, T cell epitopes and TAP binders from published literature and other available databases. From this collection, a comprehensive database "MHCBN" has been created and implemented on web at www.imtech.res.in/raghava/mhcbn. It is one of the largest databases of its kind with more than 23,000 records. The database has a number of tools for the analysis and extraction of data. The SRS version of this database is hosted at EBI to provide world wide access to data in a standard format. A large and clean dataset is the backbone for development of knowledge based prediction methods because the accuracy of these methods depends on quality and quantity of data. Larger the amount of quantitative data, more reliable will be the prediction method. After collection and compilation of data, our next step is to develop highly accurate prediction methods for various events involved in T cell epitope generation.

The first method has been developed for the prediction of MHC class I binders based on the hybrid approach of quantitative matrices and artificial neural network. The method is available as nHLAPred at http://www.imtech.res.in/raghava/nhlapred for 67 MHC class I alleles with an average accuracy of ~93%. The output display formats can aid in locating promiscuous antigenic region that can bind to several MHC alleles. These regions are capable of triggering an immune response in populations with diverse genetic makeup and therefore can be used as potential vaccine candidate. The proteasome cleavage information has further been incorporated along with MHC prediction to increase the chances MHC binder to be T cell epitope. However, this method is not able to predict the rational modification of the peptide required to increase MHC binding affinity and promiscuousity that may result in increased antigenicity. It has been shown by many experimental groups that rational modification of peptides via point mutagenesis can result in enhancement of binding affinity and promiscuousity of peptides. In order to predict mutated and promiscuous MHC binders, quantitative matrices based method has been developed. The method is available as "MMBPred", at http://www.imtech.res.in/raghava/mmbpred for a large number of MHC class I alleles. Using these two methods, user can predict mutated or natural MHC class I binders or potential T cell epitopes. However, it is not necessary that all MHC binders will stimulate T cells. Thus, the prediction method is bound to get a large number of false positive results. Developing a method based on properties of CTL epitopes instead of MHC binders can reduce this problem of false prediction. Therefore, we have developed CTLPred (http://www.imtech.res.in/raghava/ctlpred), a highly accurate method for CTL epitope prediction using statistical and machine learning techniques such as quantitative matrices (QM), artificial neural network (ANN) and support vector machine (SVM). The

QM, ANN and SVM based methods have achieved an accuracy of 70.0%, 72.4%, and 75.1% respectively. The method also predicts information about MHC restriction of predicted CTL epitopes.

Besides these methods, we have also analyzed the proteasome and immunoproteasome cleavage specificity from *in vitro* or *in vivo* proteasome digestion data. This analysis proves that cleavage specificity is affected not only by the residues at cleavage site but also by the neighboring residues towards N- and C- terminal. On this basis, we have developed Pcleavage (http://www.imtech.res.in/raghava/pcleavage), a method using support vector machine for prediction of proteasome cleavage sites in a given protein sequence. The method can predict standard proteasome and immunoproteasome cleavage sites with 68% and 63.2% accuracy on an independent dataset respectively. The proteasome and immunoproteasome based methods can recognize more than 80% cleavage sites. The prediction of proteasome cleavage can be used in vaccine design in two different ways e.g. for identification of CTL epitopes and for accessing the general cleavability of amino acid sequence. This method can be used in combination with MHC binders prediction method to identify MHC binders possessing proteasomal cleavage sites at C-terminal. Such MHC binders have more chances to be potential T cell epitopes. In addition to proteasome, TAP transporter also plays pivotal role in deciding whether peptide can pass from cytoplasm to endoplasmic reticulum for interacting with MHC or not. Therefore, determination of TAP binding affinity of peptides provides an indirect help in reducing false positive results from MHC binders as well as CTL epitopes prediction. With this aim, we have analyzed TAP binding peptides and observed that residues located at N- & C- terminus affect its binding affinity to TAP. On the basis of this analysis, a highly accurate method using cascade SVM for prediction of TAP binding affinity of a peptide has been developed. The cascade SVM utilizes the sequence and physicochemical properties information of proteins for better prediction. The correlation of 0.88 was achieved between experimentally determined and predicted binding affinity using this method. The method is available as "TAPPred" from http://www.imtech.res.in/raghava/tappred.

On the other hand, a method for MHC II binders prediction is of paramount importance for identifying the antigenic region or helper T cell epitopes from exogenous antigens. We have developed two methods "HLADR4Pred" & "MHC2Pred" for prediction of MHC class II binders. The HLADR4Pred is ANN and SVM based method for predicting the HLA-DRB1*0401 binding regions from antigenic sequence. The observation on HLA-DRB1*0401 allele has been extended to 42 more alleles for assisting in prediction of promiscuous MHC binding regions. The method is available as MHC2Pred from http://www.imtech.res.in/raghava/mhc2pred. The average accuracy of method is ~80% that is nearly 20% higher in comparison to already available most accurate MHC class II binder prediction methods.

After developing individual methods for prediction of binders for MHC class I and Class II alleles, finally, a method for joint prediction of binders for MHC class I and class II alleles has been developed. This method called HLAPred is available for 87 alleles from http://www.imtech.res.in/raghava/hlapred. One of the major features of this server is that it allows mapping of experimentally proven binders (obtained from MHCBN) on an antigenic sequence. In addition to prediction or identification of MHC binders, this server allows post processing of these binders that includes searching of identical peptides in proteins of host/self (eukaryotic) and pathogen/non-self (microbial) proteomes. The predicted antigenic region having similarity to microbial proteins and lacking similarly to higher eukaryotic proteomes have greater chances to be a potential vaccine candidate. The antigenic region with similarity to eukaryotic, specifically human has greater chances of being tolerated or responsible for autoimmune diseases.

However, all the described strategies provide information about antigenic regions in a protein but do not provide any information about protein itself. The selection of a potential target protein is one of the major issues in vaccine/drug design. Therefore, it is important to understand the function of a protein before using it as a potential target. The functional annotation and classification of protein can provide important clues to select the putative vaccine or drug target. In this study, attempts have been made to develop methods for functional annotation and classification of different type of proteins. A method ESLPred (http://www.imtech.res.in/raghava/eslpred) has been developed for subcellular localization of eukaryotic proteins. It is based on global features of proteins & similarity search and can predict different subcellular locations with accuracy of 88%. In addition to this, methods for classification of nuclear receptors and GPCRs have been developed because these receptors are the major drug or vaccine targets. These methods are available as Nrpred (http://www.imtech.res.in/raghava/nrpred), GPCRpred (http://www.imtech.res.in/raghava/gpcrpred). In this series, another method GPCRsclass (http://www.imtech.res.in/raghava/gpcrsclass) has been developed for predicting the amine type of receptors such as acetylcholine, serotonin, dopamine and adrenoceptors.

These are major drug targets for curing nervous disorders such as parkinson's disease. These prediction methods are highly accurate and are based on global features of proteins.

The computational tools developed in this thesis will help an immunologist in analyzing and identifying the most antigenic and favorable vaccine candidates from microbial genome data.

12.2 Future Prospects

The concept of epitope prediction has tremendous biomedical and biotechnological applications. Improved methods for prediction of antigenic regions will accelerate the process of subunit vaccine development. The understating about various intermediates steps of antigen processing will help in formulating more accurate T cell epitope prediction methods. A comprehensive method developed by taking into consideration specificity of MHC molecules, TAP transporter and proteasome cleavage would be a promising tool for mining the T cell epitopes from various genomes. These tools will help in discovering the novel vaccine based on computer predicted antigenic regions. Furthermore, the prediction of promiscuous antigenic peptides would help in designing vaccines that are effective for many genotypes of world.

Together with the ability to perform rational antigen modifications, it can be used to modulate majority of immune reactions. The up regulation of antigen specific immune response by 'agonistic' peptides is desirable in vaccine design against pathogenic invasions and tumors. While the antigen-specific down regulation by the 'antagonistic' peptides would be useful during the inappropriate pathological conditions such as autoimmune disorders. The direct practical applications of this can be visualized in the areas of cellular immunology, transplantations, vaccine design, immunodiagnositcs, affinity purification of proteins and the molecular understanding of susceptibility to diseases.

All these methods would provide an illuminated way to immunologist for designing vaccines against deadly diseases. These methods can help the immunologist to obtain the relevant immunological information from unmanageable amount of data. In summary, all these methods would speed up the discovery of novel vaccines and increase understanding about the intermediate processing events. These methods will act as stepping stones in "*In silco* modeling of immune system and its responses". Within next 10 years, we can expect to see a novel class of vaccine and drug targets explored by immunoinformatics tools.