SUMMARY AND FUTURE PROSPECTS

8.1 Summary

The overall objective of this thesis is to develop improved and novel prediction methods for identifying potential candidate(s) for subunit vaccine based on B-cell epitopes. The present work is emphasized i) on the identification for potential targets which include virulence factors, and ii) development of prediction method(s) of linear B-cell epitope.

The proteins in an organism involve in cellular process, metabolism and in information decryption (transcription and translation), remaining are classified into virulence factors, which allow the germs to establish themselves in the host (adhesions, toxins and hemolytic molecule). Firstly, an attempt was made to predict directly the virulence factors from the pool of proteins or proteome of gram negative bacteria. The traditional strategies for computational annotation of proteins were tried that includes i) similarity search using PSI-BLAST; ii) SVM-based method using amino acid composition and iii) SVM-based method using dipeptide composition which also consider local order of amino acids. It was observed that performance of traditional approaches was very poor in functional classification of proteins. In order to improve the performance, tetrapeptides were used as input features. The approach relies on identifying short signaling patterns and group of patterns for each four broad functional classes that were present in higher number. The performance of method based on tetrapeptides was much higher than traditional methods based on residue composition. The performance was further improved when new and traditional approaches were combined. The *ab-initio* tetrapeptides based module (68.66%) gave better overall accuracy than amino acid (52.39%) or dipeptide composition (47.01 %) based module individually. The hybrid module which used pattern, amino acid and dipeptide composition information, was able to achieve an overall accuracy of 70.75%.

Further, bacterial toxins from gram-positive as well as gram-negative organisms were studied. A systematic attempt was made to collect, compile and analyze the bacterial toxins obtained from Swiss-Prot. The frequency of occurrence of polar uncharged amino acid, asparagine was significantly higher (P value 5.38421E-6) in bacterial toxins than in non-toxins. It was also observed that the composition of bacterial toxins was significantly different than the non-toxins at 0.05 level for the residues alanine, aspartic acid, leucine, methionine, proline, threonine, valine and tyrosine. Phylogenetic analyses were performed with these bacterial toxins in order to understand their co-evolution. Based on the analyses, an attempt was made to develop a method for predicting bacterial toxins, their class (exotoxin or endotoxin) and sub classes of exotoxins. The SVM module based on amino acid composition achieved slightly higher accuracy than dipeptide composition both in discriminating bacterial toxins (96.07% and 92.50% respectively) and type – exotoxins or endotoxins (95.71% and 92.86% respectively). In the functional classification of exotoxins, PSI-BLAST performed better than HMM and combined method predicted with 100% accuracy. The BTXpred server was developed for prediction of bacterial toxins and classifying them based on their release and function. This method, in association with PSI-BLAST, can be used for automated annotation of genomic data. The study also proves that there is direct correlation between the features of the proteins (amino acid, dipeptide composition) and the bacterial toxins. This server also assist preliminary analysis of possible functions of new exotoxins and in designing experiments for functional characterization of newly identified bacterial toxin sequences thereby reducing the number of essential experiments.

Next, neurotoxins sequences were studied, these toxins were collected from Swiss-Prot database, compiled and analyzed. It was observed that the neurotoxin protein sequences contains sufficiently higher amount of cysteine (P value 1.91314E-10) than non-toxin proteins. Similarly, the composition of neurotoxins was significantly different than the non-toxins at 0.01 level for the residues alanine, glutamic acid, isoleucine, lysine, leucine, asparagines, glutamine, methionine, valine and tyrosine. An attempt was

also made to predict neurotoxins from all sources using higher machine learning technique like ANN and SVM, and similarity search tools from primary amino acids sequence. The highest accuracy and MCC achieved for prediction of neurotoxins was of 97.72% and 0.9416 respectively by SVM based on composition. The overall accuracy to classify neurotoxin sequences based on their composition, dipeptide composition and length, dipeptide and length SVM modules were of 78.94%, 88.07%, 84.91% and 87.72%, respectively. In classifying neurotoxins based on source, low accuracy was achieved for cnidaria (30-55%) class, whereas 100% accuracy (amino and dipeptide composition) was achieved in eubacteria, 95% accuracy in arthropoda (amino composition with length) and 90.37% in chordata (dipeptide composition). In classifying neurotoxins based on function, the SVM module based on dipeptide composition with length achieved maximum overall accuracy of 94.88% and the hybrid approach of SVM module (Dipeptide composition + length of the sequence) with PSI-BALST and MEME/MAST achieved an overall accuracy of 95.11% and 96% respectively. The sub-classification of ion-channel blockers was also studied, since these are of particular interest to pharmaceutical companies. The NTXpred server was developed for prediction of neurotoxins and classifying them based on source and function. This method, in association with PSI-BLAST, can be used for automated annotation of genomic data. The study also proves that there is direct correlation between the features of the proteins (amino acid, dipeptide composition and length) and the neurotoxins function.

The potential vaccine target should be non-allergenic to humans. Keeping this in mind, algorithm was developed for predicting allergens (www.imtech.res.in/raghava/algpred) based on the presence of IgE epitopes, MEME/MAST motif, allergen representative peptides (ARPs) BLAST search and SVM based method based on amino acid and dipeptide composition. SVM based module achieved accuracy around 85% using amino acid composition. The IgE epitope based search provides high specificity but have poor sensitivity compared to SVM based approach, as not all allergenic proteins have known IgE epitopes. This IgE epitope based strategy helped in improving the performance of SVM based method, where it increased the sensitivity without loosing the specificity significantly. The motif based method

(MEME/MAST) developed in the study has low specificity. In the prediction of allergens based on ARPs, BLAST was used for similarity searching because it is fast and reliable.

Thereafter, the main emphasis was laid on B-cell epitopes and its prediction. For designing a new or more accurate method for predicting B-cell epitopes, an ample amount of high quality data is a pre-requisite. Keeping this in mind, the B-cell epitopes related information were collected from literature (pubmed and Sciencedirect) and other publicly available database. "Bcipep" database was created and implemented on web at www.imtech.res.in/raghava/bcipep. Bcipep covers information on experimentally determined linear B-cell epitopes of varying immunogenicity and on epitopes from a wide range of pathogens. Latest version of Bcipep contains 3140 entries, where each entry provides detailed description of a B-cell epitope. This database allows keyword search, peptide search and peptide mapping of the query sequences. European Bioinformatics Institute (EBI), UK, hosts Bcipep.

Subsequently, the evaluation of all the existing B-cell prediction methods based on physico-chemical properties was carried out on a large data set of epitopes obtained from Bcipep database. Different combination of physico-chemical properties like flexibility, hydrophilicity, polarity and exposed surfaces were tried and observed that the accuracy of B-cell epitope prediction based on these properties was low (52.92% to 57.53%). It was observed that flexibility based scales as implemented by Karplus and Schulz, 1985, relatively perform better than any other property scale used earlier (57.53%). In order to see the effect of combination of properties, the best parameter, flexibility, was combined with other properties one-by-one. Though combination of flexibility, hydrophilicity, polarity and surface exposed based scales achieved accuracy of 58.70%, sensitivity 56% and specificity 61% but overall performance was not very high. Thereafter artificial neural network was tried for developing algorithm for the prediction of B-cell epitopes. Here, feed forward network and recurrent neural network were implemented. It was observed that RNN has been more successful than FNN in prediction of B-cell epitopes. The accuracy of the method improved significantly (P value

0.01732) when RNN was implemented for training and testing (at 0.02 level). An accuracy of 65.93%, sensitivity 67.14%, specificity 64.71% and MCC 0.3187 was achieved using RNN at threshold 0.5. The length of the peptide is also important in prediction of B-cell epitopes from antigenic sequences. The major problem of using machine-learning technique is that the input window length has to be fixed, whereas B-cell epitopes sequence vary from 5 to 30 as reported in literature. An attempt has been made to develop a method using window length of 10, 12, 14, 16 and 20, where the epitope length is fixed, by introducing or subtracting equal number of residues at both terminals derived from its original antigenic sequence. Based on the study the server ABCpred was developed and is accessible from http://www.imtech.res.in/raghava/abcpred/

Lastly, a case study was performed for the identification of potential targets in *Mycobacterium tuberculosis* H37Rv strain. Twenty one potential vaccine candidates were identified. Although *in silico* results must be confirmed *in vitro* and *in vivo*, the analysis of the *Mycobacterium tuberculosis* genome using the above strategies identified potential vaccine candidates. To conclude, the present thesis undertakes a systematic study of prediction of vaccine candidates based on B-cell epitopes.

8.2 Future prospects

The immunoinformatics analysis enables systematic identification of all the potential antigens of a pathogen, making it theoretically possible to develop a safe and efficacious vaccine against any infectious disease. However, the identified candidate(s) need to be confirmed in real life. The antibody epitope prediction method(s) can play a role in development of monoclonal antibodies, and that antibodies can modify various aspects intracellular infection like HIV and tuberculosis to the benefit of the host.

Currently, the Bcipep database does not cover discontinuous epitopes and it contains peptides having only natural amino acids. In future, this database can be updated with discontinuous epitopes and non-natural amino acids. Beside this, carbohydrate and lipid epitopes can also be added in this database. It is observed that earlier the B-cell