10 SUMMARY AND FUTURE PROSPECTS

This is the era of information technology where technological advancement touched biology too. In the biological field due to ease of national, international collaborations and very much influenced by the philosophy of open source, the scientists, scholars and organizations around the world try to tackle the serious human threats. The research has now globalized and shifting ahead from the traditional individualistic approach. With this philosophy in mind and boom in the technology it is not surprising that ton of data is available freely in a particular field of interest with a single click of a mouse. Although availability of data is boon for a researcher but it comes with few major challenges such as reliability, interpretation, annotation and application. The reliability issue goes hand in hand with the technology, and the accuracy of the instrument used to generate the data, and a researcher has minimal interference in this. However, the annotation and application of this huge amount of data is where a researcher spends most of his/her time and tries to understand the intricacies of life. The data in biology comes in different formats like genome sequences, which are organized codes of ATGC, protein sequences, fluorescent and gel data in the form of images, small molecular data in the form of structure and activity, etc. Bioinformatics or technically known as *in-silico* analysis is no doubt needed in all aspects of data annotation and application or tool generation to further help in the understanding of biology and occupy the base of the pyramid. Experimental approaches, on the other hand, are more specific, costly and used in the confirmation of any hypothesis, which resides at the tip of the pyramid.

Due to the emergence of fields like clinical immunology and other branches of high-throughput tools in immunology, there is a generation of huge amount of data, which led to databases like IEDB, AntiJen, MHCBN, BCIPEP, etc. Unfortunately their focus is on the epitope, and not the antigen. There are many antigens for which specific epitope or MHC binding information is not currently available, yet they are known experimentally to induce either or both innate or adaptive immune responses. Such antigens - or similar pathogenic proteins - might prove useful in vaccine design. These antigens require urgent and rigorous cataloguing. To compensate this, the present work describes the database AntigenDB which is a specialized, value-added database of antigens derived from pathogenic organisms. This resource is intended to be a repository for all experimentally determined antigens, and we started with major organisms and irrespective of whether such an antigen is associated with

the known epitope data. The database is freely accessible through a web browser at http://www.imtech.res.in/raghava/antigendb/.

After antigen compilation in the form of dedicated database, there is issue of epitope prediction. In the present thesis, three types of epitopes have been worked; Linear B-cell epitopes, Conformational B-cell epitope and antigenic/epitopic carbohydrates. Linear B-cell epitope prediction has been a great challenge for the computational immunology field. Since 1983 with the development of different matrices for the prediction of protein secondary structure, researchers have been trying to use similar and derived matrices for the linear B-cell epitopes. They have used very limited dataset with random negative sequences from UniProt and in some cases from same antigen sequences. These matrices and later machine learning algorithms could not raise the prediction bar to a convincing level. In this study first time we have used largest possible epitope dataset in addition to the experimentally proved negative datasets from IEDB in place of random negatives. We have tried to incorporate useful matrices left out in literature in combination of our own derived features to predict linear B-cell epitope. We introduce the concept of cascade algorithm into the B-cell epitope prediction, though cascade had been used in the past for other biological problems. On the newly created IEDB dataset, our dipeptide composition and cascade algorithm performed very well. We used our algorithm in cross-validation mode to compare other existing methods on earlier tested benchmark dataset and found that our method outperformed existing tools. Our models showed a promising result even with the very stringent conditions used. The method is implemented in the form of LBtope web server.

Conformational B-cell epitope is another area of active research where several prediction servers already exist. These are again based on amino acid matrices and used in different techniques starting from simple scale based to sophisticated machine learning algorithms. One major setback of all these models is that most of them require antigen's 3D structure, which is a major bottleneck that requires time, labor and money. Few methods which take antigen sequence as input actually predict the 3D structure and then derive features from it. So in principle there are no methods, which take antigen primary sequence and based on the sequential features, predict the conformational B-cell epitope. In this study first time, we have introduced the concept of composition profile of patterns, which was earlier used for whole protein. After combining with SVM, we could achieve accuracy same as other best structure

based methods. The method was implemented in the form of CBTOPE web server as well as standalone software.

Carbohydrate based vaccine development has got a huge boost due to advancement in the carbohydrate isolation, purification and characterization techniques. This advancement and experimentation led to availability of validated antigenic or epitopic carbohydrates through IEDB database. Like other antigens, antigenic carbohydrate discovery is a tedious job. Contrary to peptide epitopes, there are no *in-silico* models to predict the antigenic carbohydrates. These antigenic and similar non-antigenic carbohydrates are not coded in amino acids, but they are like small chemical molecules. Therefore, we used the Chemoinformatics approach to derive the features or descriptors from each molecule and exploited different machine learning approaches like SMO and Random Forest implemented in Weka. We developed the QSAR based models for the prediction of antigenic carbohydrates with good accuracy.

This thesis also includes challenges from the innate immunity like analysis of the innate immune modulators, NAD^+ and TLR4. Besides having major roles in cellular metabolism, NAD^+ regulates several important immune system pathways. For annotation and better understanding of this molecule, we need to identify its target molecule and binding sites. There are several motif based rules, which define the NAD^+ binding site like presence of popular Rossman fold. However, these rule-based methods could not account of all the NAD^+ binding proteins and their binding sites and mostly structure dependent. In our study, we derived features from the NAD^+ binding proteins obtained from PDB and coupled with PSSM and SVM, developed the sequence based models, which predict the NAD^+ interacting residues with high accuracy.

As the second part of innate immune regulation, we tried to analyze the human TLR4 inhibitors. TLR4 is a major regulatory molecule implicated in the several inflammatory deceases and is a primary receptor for the bacterial endotoxins or LPS. There is major concern to derive inhibitors against this molecule but to our knowledge, there are no *in-silico* models to predict the inhibitors against human TLR4, which could be used in the fast molecular screening protocols. Based on two experimental bioassays available at PubChem database, we obtained several experimentally proved molecules, which were active and inactive against

human TLR4. We calculated the useful descriptors and coupled with regression analysis in Weka and R, we could arrive at a suitable prediction model which could be used had in hand for the large-scale screening of the inhibitors.

Future prospects

The pathogen antigen database (antigenDB) developed in this thesis would be very useful if supplemented with structural information of the corresponding epitopes, i.e. tools for the antigen and epitopes visualization need to be integrated for increasing the understanding. In addition like any other database other important pathogenic organisms should be added periodically.

Linear and Conformational B-cell epitope prediction models (LBtope and CBTOPE) need to be integrated with other structure based tools, and a visualization tool would be an asset to these predictors. It would be interesting to apply the similar algorithm on flexible length linear B-cell epitopes.

It would be interesting to add other antigenic non-peptide models like antigenic hormones or antibiotic to the present antigenic carbohydrate predictor CarboTope. By taking the other side of the human TLR4, it would be really interesting to develop a model for enhancing the human TLR4 signaling, which is contrary to the present model and this can be used as a predictor of human adjuvants for the designing of better vaccines. Conclusively, incorporation of more robust biophysical and structural descriptors might further enhance the models for the prediction of all types of peptide epitopes, and all the theoretical models developed in this thesis are liable to verify experimentally.