7. Summary

In the recent years many peptides with various beneficial therapeutic properties have been reported and are fast gaining the attention of pharmaceutical companies. Though some peptides reach clinical trials, majority of them fail after the first or second phase of the clinical trials because of reasons like short metabolic stability, lack of efficient delivery routes, less distribution and bioavailability due to protease degradation, quick elimination by hepatic and renal pathways and toxicity. In order to assist the scientific community, this thesis mainly focuses on the creation of in-silico resources for optimization of properties which will help in translating peptide drugs into the market.

First, 'PEPlife' (http://webs.iiitd.edu.in/raghava/peplife/), a repository of experimentally verified half-life of 2229 peptides was developed by manual curation of research articles. The main fields in PEPlife include peptide name, sequence, function, origin, modifications, and assay to estimate its half-life. We also predicted the tertiary structure, SMILES and physicochemical properties of the peptides in the database. Modules have been integrated for easy searching, browsing and analysis of data in the database. PEPlife can be used to search for the half-life of 2229 peptides, and study how the modifications affect the half-life and biological activity of a peptide and its analogues. The structures can be used for docking, simulations and QSAR studies.

Next, the information in PEPlife was used to create datasets for predicting the half-life of peptides in blood since most of the therapeutic peptides are parenterally administered. 'PlifePred' prediction web server can predict the half-life of both natural and modified peptides. Models have been developed on 261 peptides containing natural and modified residues, using different chemical descriptors. We used features like amino acid, dipeptide and atom composition and binary pattern of sequences and chemical descriptors for structures. Machine learning methods like SVM, SMOreg, Linear Regression, Gaussian Processes, IBk were employed to develop models. The best model using 43 PaDEL descriptors got a maximum correlation of 0.692 between the predicted and the actual half-life peptides. Secondly, models were developed on 163 natural peptides using amino acid composition feature of peptides and achieved a maximum correlation of 0.643. Thirdly, models were developed on 163 natural peptides using chemical descriptors and attained a maximum correlation of 0.743 using 45 selected PaDEL descriptors. In order to assist researchers in the prediction and designing of half-life of peptides, the models developed

have been integrated into PlifePred web server (http://webs.iiitd.edu.in/raghava/plifepred/). PlifePred can be used to predict the half-life of peptides based on sequence as well as structure and design analogues with desired half-life and physiochemical properties. The limitation of this study is that it has been performed on a small dataset, since more data of experimentally validated half-life of peptides in blood was not available. In future, when such data is available, a more robust method can be developed to predict peptide half-life.

Since peptides are gaining the attention of researchers in the emerging field of nanobiotechnology due to their property to undergo spontaneous assembly to form nanostructures which have various applications ranging from drug delivery to tissue culture scaffolds, we developed a database called 'SAPdb' (http://webs.iiitd.edu.in/raghava/sapdb/). SAPdb has a collection of 637 short peptides which make ordered nanostructures on self-assembly. This comprehensive resource will enable researchers to easily study the properties and experimental conditions which are responsible for the self-assembling trait of dipeptides and tripeptides and the shapes of nanostructures they form. However, we were unable to create a robust method which could predict which peptides could undergo self-assembly based on physiochemical and residue composition. Even previously reported methods based on simulations failed to distinguish between the experimentally verified self-assembling and non-self-assembling peptides with high accuracy. It is hoped that in future, once a larger dataset is available, a better in-silico platform will be developed to successfully differentiate between these two classes of peptides and facilitate the researchers working in the burgeoning discipline of nanobiotechnology.

Most of the peptide drugs in the market today are administered parenterally to patients using needles which reduce patient compliance. Developing non-invasive methods of topical delivery is the best pain-free and infection-free alternative and also increases the in vivo stability of the administered peptide since it bypasses the hepatic first-pass metabolism. To expedite research in this direction, we have developed 'TopicalPdb' (http://webs.iiitd.edu.in/raghava/topicalpdb/). TopicalPdb is a repository with 462, 173 and 22 entries of peptides which can be non-invasively delivered through skin, eye and nose. Using these entries, a prediction method called 'SkinPP' (http://webs.iiitd.edu.in/raghava/skinpp/) was developed to distinguish between skin penetrating and non-skin penetrating peptides. The best model which could successfully differentiate SPPs and random peptides of same length distribution from Swiss-Prot achieved

a MCC of 0.59 and accuracy of 79.63%. While the best model developed using SPPs and non-CPPs reached a MCC of 0.49 with accuracy of 74.55%. SkinPP can be used to predict large libraries of peptides for screening their skin penetrating properties and design their analogs with high efficiency of skin penetration. It has a protein-scan module which can be used to identify SPP like regions within a protein.

The next objective was to create resources for checking the hemotoxicity of peptides in order to improve their therapeutic index. Some peptides are toxic to eukaryotic cells and instead of exerting beneficial effects on the patient, they rupture the erythrocytes. To study the hemolytic potency of the peptides, a database called 'Hemolytik' (http://webs.iiitd.edu.in/raghava/hemolytik/) was developed. Hemolytik harbors 2651 hemolytic and 319 non-hemolytic peptides. It has 237 modified peptides and 387 peptides with a D amino acid in the sequence. Majority of the peptides are tested using human blood as test- sample. Using this compiled information, 'HemoPI' (http://webs.iiitd.edu.in/raghava/hemopi/), which is a web-server to distinguish between hemolytic and non-hemolytic peptide was developed. Since there are no fixed criteria to decide upto which concentration a peptide should be considered non-hemolytic since at very high concentrations most of the peptides were reported to be slightly hemolytic, therefore, we created our own assumptions and created 3 datasets. Models developed using dataset-1 could distinguish between hemolytic and random peptides taken from Swiss-Prot. Models developed on the dataset-2 could differentiate between highly hemolytic peptides and the low hemolytic peptides from Hemolytik database. While dataset-3 had the largest dataset from Hemolytik and DBAASP v.2 databases and was based on more stringent criteria for both highly hemolytic and poor hemolytic peptides. Models were developed using residue composition and motif information as features using various machine learning techniques like SVM, IBK, J48, Logistic, Multilayer Perceptron etc. The best performing models developed were implemented freely accessible, user-friendly HemoPI web-server.

It is anticipated that the in-silico platforms described in this thesis will assist in the progress of peptide-based therapeutics and will be instrumental in expediting their translation into the pharmaceutical market.