Summary and Future Prospects

The overall objective of the thesis is to improve the tertiary structure prediction of small bioactive peptides. The prediction of secondary structure can serve as an important step toward predicting the tertiary structure. One of the most important and abundant secondary structural elements of small peptides is tight turn, especially β-turns, so we have aimed at prediction of β-turns in order to improve the prediction of three-dimensional structure of peptides.

Before improving the β-turn prediction accuracy and developing a new method, there is a need to evaluate the already existing methods in order to know the present status of β-turn prediction accuracy level. Thus to begin the present work, an extensive evaluation of six β-turn prediction methods: Chou-Fasman algorithm, Thornton's algorithm, 1-4 & 2-3 Correlation Model, Sequence Coupled Model, GORBTURNv3.0 and BTPRED has been carried out using old as well as new parameters on a uniform, non-homologous data set of 426 protein chains. The evaluation study has shown that neural network based method BTPRED outperforms all other methods using old parameters. Using new parameters and large data set, statistical methods perform as good as sophisticated neural network based method BTPRED. It has been found that the accuracy of prediction of β-turns can be improved using regular secondary structure information. Further, to assist the researchers working in this field, a web server BTEVAL (http://imtech.res.in/raghava/bteval/) has been developed for online benchmarking of newly developed β-turn prediction methods. Out of six existing β-turn prediction methods, five are statistical and none is available to the public. So, a web server BetaTPred (http://imtech.res.in/raghava/betatpred/) has been developed for predicting β-turns and their types using any of these statistical methods. The server also allows the consensus prediction of β-turns.

It has been shown in the past that accuracy of secondary structure prediction methods can be improved using multiple sequence alignment (MSA). A neural network based method; BetaTPred2 (http://www.imtech.res.in/raghava/betatpred2/) has been developed for predicting β-turns using MSA. In this method, two feed-forward neural networks have been used; i) 'sequence-to-structure' network; and ii) 'structure-to-structure' network. The first 'sequence-to-structure' network is trained with the multiple sequence alignments in the form of PSI-BLAST generated position specific scoring matrices. The initial predictions obtained from the first network and PSIPRED predicted secondary structure have been used as input to the second structure-to-structure network to refine the predictions obtained from the first net. Significant prediction accuracy 75.5% has been achieved when tested using seven-fold cross-validation on a set of 426 non-homologous protein chains. The $Q_{predicted}$, $Q_{observed}$, and MCC values obtained using BetaTPred2 are 49.8%, 72.3% and 0.43, respectively. The performance of this method is better than any other method reported in the literature. This approach has been extended for developing neural network based method, Betaturns (http://www.imtech.res.in/raghava/betaturns/) for predicting β-turn types I, II, IV, VIII and NS in a protein. It has been found that well defined turn types I and II are predicted with a significant accuracy as compared to less numerous and ill-defined type VIII β-turn.

Other tight turns, γ- and α-turns have also been studied in this thesis. Neural network based methods, Gammapred (http://www.imtech.res.in/raghava/gammapred/) for prediction of γ-turns and Alphapred (http://www.imtech.res.in/raghava/alphapred/) for prediction of α-turns has been developed using PSI-BLAST generated PSSM and PSIPRED predicted secondary structure information. In addition to neural network, other machine learning algorithms such as Weka and PEBLS have also been attempted. For γ-and α-turns, the neural network methods have achieved MCC of 0.17 and 0.16 respectively.

To understand the structural and functional aspect of β-turns in proteins, the intra/inter residue interactions have been studied. β-Turns are often stabilized by hydrogen bonding between the carbonyl group of i^{th} residue and NH group of i+3 residue. In addition to conventional hydrogen bonding, the non-conventional, weak

polar interactions such as aromatic-NH (Ar-NH), CH···O and CH···π stabilize β-turns. Thus, these interactions have been analyzed in β-turns.

In β-turns, the most predominant Ar-NH interaction has been found between the aromatic moiety and NH group of the same residue. Among all β-turn types, the type I has the maximum number of Ar-NH interactions. The propensity of different β-turn residues to participate in such interactions varies and shows a clear preference for Met, Ile, Val and Gly residue to act as hydrogen bond donor. On the acceptor side, Trp residue has the highest acceptor efficiency due to its larger aromatic surface as compared to Phe and Tyr residues. In addition to Ar(i)-NH(i) interactions, significant number of Ar(i)-NH(i+1) interactions have also been observed in β-turns which are mostly perceptible at second and third positions of β-turns. In contrast to Ar-NH interactions, there is a very small percentage of C-H···O and C-H···π interactions, nearly 5%. The geometry of these interactions in β-turns has been found to be in consistent with that of regular secondary structures. The propensity of different β-turn residues to participate in such interactions varies and shows a clear preference for Pro and Gly residue as hydrogen bond donor. On the acceptor side, Asp residue is the most frequently residue participating in C-H···O interactions. In C-H···π interactions, Trp is the most prominent acceptor as compared to Phe and Tyr residues. CH···O interactions occur extensively involving side-chains, $C_{aliphatic}$ atoms. Both types of interactions have been found significantly in Type I β-turns and are mostly perceptible at 1 & 3 and 1 & 4 positions of β-turns.

Based on amino acid propensities in these interactions, neural network based methods have been developed for prediction of such interactions in proteins using the same concept of MSA. The method Ar_NHPred (http://www.imtech.res.in/raghava/ar_nhpred) has been developed for prediction of Ar-NH interactions from a given sequence and is found to be 11.4% higher than the random prediction. For prediction C^{α}-H···O and C^{α}-H···π interactions, a recurrent neural network based method CHpredict (http://www.imtech.res.in/raghava/chpredict/) has been developed for predicting such interactions with a separation distance up to 16 residues between donor and acceptor moieties. Using both sequence and secondary structure, nearly 54% of C^{α}-H···O

interactions and 82% of C^{α}-H$\cdots\pi$ interactions at sequential separation of 4 and 3 residues respectively are predicted correctly which is significantly better than random prediction.

Finally, the knowledge of β-turn prediction has been used to predict the complete three-dimensional structure of small bioactive peptides. The analysis of a dataset of 77 bioactive peptides has shown that β-turns are major secondary structure elements in these peptides. So, β-turn information along with regular secondary structures (helices and sheet) information has been used to predict the three-dimensional structure of these peptides. It has been observed that when β-turns information is incorporated along with the helices and sheet information, the average root mean square value decreases by nearly 0.6Å. Based on this study, a web server PEPstr has been developed which allows to predict the tertiary structure of small bioactive peptides using the following steps; i) prediction of regular secondary structure and β-turns using BetaTurns; ii) generation of conformation by assigning dihedral angles corresponding to secondary structure information; iii) placement of side chain angles using Dunbrack backbone dependent rotamer library; and iv) energy minimization using AMBER. The server Pepstr is accessible from http://www.imtech.res.in/raghava/pepstr/.

To conclude, in this thesis a systematic study of prediction of tight turns in proteins and their role in tertiary structure prediction of small bioactive peptides has been carried out.

Future prospects

Accurate prediction of tight turns along with regular secondary structure elements will be valuable in building the three-dimensional structure and improving the prediction accuracy. Even a crude or approximate model can help the experimentalist in guiding his/her experiments. If the secondary structure is known, it is possible to derive a comparatively small number of possible three-dimensional structures using knowledge about the ways that secondary structural elements pack. Such a progress in

protein structure prediction will be useful in drug design, toxicology, medicine and further understanding of protein function.

The prediction of β-turns and other tight turns will be useful in fold recognition studies. Since the present fold recognition methods consider only regular secondary structures, the incorporation of tight turns can further improve the sensitivity of fold recognition methods.

Prediction of β-turns will be useful for detecting higher order structures in proteins such as super-secondary structures. These super-secondary structures are recognized as patterns that make up protein domain. For instance, β-turn is an important component of β-hairpins. Prediction of β-turns would aid in identification of these structures, which will be further helpful in constructing an actual three-dimensional structure.

The analysis of non-conventional interactions and the knowledge of frequency of occurrences of the residue as donors and acceptors in these interactions will be helpful in designing of β-turns. The prediction of these interactions will be particularly helpful in protein folding problem as the protein fold is mainly dictated by inter-residue interactions.

The tertiary structure prediction of small bioactive peptides will be useful in understanding their biological function, designing of peptides and protein-ligand interactions.