

12 Summary

The focus of this thesis is to develop methods for better understanding and prediction of protein structure and function. For the development of bioinformatics methods, quality dataset are needed from recent release of PDB. The protein data bank (PDB) is the principal source of known protein structures. Currently, it holds more than 100,000 structures having 17,000 ligands and 3,000 nucleic acid interacting complexes. Various databases and web servers provide comprehensive information of the structure, protein-ligand and protein-nucleic acid interaction for single or a group of PDB. There are no database and web-server, which can provide function and structure information of multiple PDB, which can be directly used for analysis and development of prediction methods.

To overcome this limitation we developed a platform for creation and compilation of datasets from PDB. The ccPDB (creation and compilation of datasets from PDB) is a database cum web-server, which provides customized structure and function datasets. It allows creation of regular secondary structure datasets based upon the percent composition of helix, sheet and coil region. Similarly, PDB chain having different ratio of irregular secondary structure elements (β -turn, γ -turn, beta bulge, beta-hairpin and psiloop) can be easily searched. The ccPDB also allows searching of DNA, RNA and specific ligand interacting PDB chains. These PDB chains are further annotated and the region having helix, sheet, coil or turns are marked in PDB chains. The ccPDB data and server is freely available on <http://crdd.osdd.net/raghava/ccpdb/> for public.

The ccPDB platform stores the secondary structure elements information and DNA/RNA, ligand interacting information at residue level. Our next objectives are to utilize the ccPDB data for annotation, analysis and development of prediction methods. The best method for protein annotation is Blast2GO, which assign GO terms for unknown proteins (Conesa et al., 2005). We advanced the concept of protein level annotation to residue level annotation and developed StarPDB (Structural Annotation of Residues using PDB). BLAST is used to find similar PDB chain in the recent release of PDB, and query sequence is annotated at residue level using ccPDB data. StarPDB provides regular secondary structure annotation, irregular secondary structure annotation (turns and loops), DNA/RNA and ligand interacting residues in query protein

Chapter 12

Summary and Future Prospects

sequence. StarPDB is the first method, which can annotate all possible ligand (~17,000) interacting residues in the query sequence. StarPDB has rich visualization and provides online selection, editing, theme change and export of results. StarPDB is freely available for the scientific community at <http://crdd.osdd.net/raghava/starpdb/>.

For better understanding of ligand interaction amino acids, we developed a web-based tool LPIcom. Based upon the propensity, amino acid composition and physicochemical properties score of interacting residue the ligands can cluster in different groups. LPIcom also provides percentage composition of interacting and non-interacting residues of a specific ligand. For a case study we analyzed ATP interacting residues and the result matched with the literature (Chauhan et al., 2009b). LPIcom also generates the sample logo, two-sample logos and possible motif. The last module predicts the possible interacting sites for a particular ligand using propensity score. LPIcom is freely available on <http://crdd.osdd.net/raghava/lpicom/> for public.

The next part of the thesis deals with analysis and prediction of secondary structure and tertiary structure of peptides and proteins. In the past decade, many methods had developed for the secondary structure prediction of proteins, but none for peptide secondary structure prediction. It was observed that similar sequence segments adopt a different structure in peptides and proteins due to a different environment and degrees of freedom. We developed a hybrid method for peptide secondary structure prediction with 81.12%, 71.1%, 84.22% accuracy for helix, sheet, and coil respectively. In comparison PSIPRED, achieved 82.7%, 54.9%, 78.4% accuracy for helix, sheet, and coil respectively. Based on a hybrid method, a web server PEP2D (<http://crdd.osdd.net/raghava/pep2d>) has been developed. The PEP2D has unique module Mutant Peptides, which generate all possible mutant and predict the secondary structure of all mutants.

After regular secondary structure elements, β -turns are the most important and abundant structural element. The β -turn helps in making protein structure compact by forming the turn in proteins. Initially β -turn prediction method was based upon propensity scores derived from PDB, but these methods have limited accuracy. We updated the old propensity scores on a large dataset of 20,000 PDB and calculated all possible amino acid pair wise propensity scores. It was observed that amino acid glycine, proline, asparagine and aspartic acid are favored, and leucine, isoleucine, methionine and valine are not favored in β -turns formation. Position wise analyses

Cha

suggest
3rd and p
position
was mo
most fa
and thi
position
disturbi
simple
MCC f

In the
achiev
predict
MCC.
compa
dataset
is
(http://
design

Dihec
impro
but in
meth
predi
supe
not a
a lit
Euc
with

suggest that aspartic acid is favored at 4th position; glycine is favored at 3rd and 4th, asparagine at 3rd and proline at 2nd position. In the case of pair wise analyses P1-2, proline is favored at second position; asparagine and glutamate were favored at third position in P1-3 pair. The pair of CC was most favored at position 1-4, possibly due to disulphide bond formation and glycine was most favored at position 4. Similarly, for tripeptide P1-2-3 proline, glycine is favored at second and third position and for P2-3-4 glycine, aspartic acid are more favored at third and fourth position. β -turn forming amino acids and pairs are more favored in β -turn formation, and β -turn disturbing amino acid and their pair disfavor the formation of β -turns. We also developed a simple propensity based method using various positional propensity score and achieved 0.36 MCC for complete β -turn method.

In the past decade many residue level β -turn prediction methods had been developed, which achieve maximum MCC of 0.50. We developed a turn level method (BetaTPred3) for β -turn prediction in proteins using PSSM, secondary structure and propensity score and achieved 0.51 MCC. In the case of β -turn types, BetaTPred3 achieved higher MCC for all turn types as compared to previous methods. Next we updated the β -turn prediction method with a latest dataset of 6376 PDB chains, in order to enhance the realistic prediction of β -turns. A web server is developed to predict β -turns in proteins and is available at (<http://crdd.osdd.net/raghava/betatpred3>). For the first time, we developed a module for designing of β -turns and to identify best mutations either to induce or break β -turns in proteins.

Dihedral angles are used for accurately defining the local ordering/confirmation in proteins to improve the quality of structure. Many methods are developed for prediction of dihedral angles, but in the lack of proper benchmark, it is difficult to state which method is better than another method. To establish the best dihedral prediction method, we benchmarked the three latest prediction methods: SPINEX, ANGLOR and TANGLE on three datasets and demonstrate the superiority of SPINEX over other methods. The prediction method has a limitation that they do not account for a relationship between phi psi angles. To overcome these limitations, we created a library of representative dihedral angles from 100% non-redundant PDB. Based upon the Euclidean distance between the similar patterns, we selected the pattern having the least distance with other similar patterns. Using this approach, we created representative library of pattern 7, 5

Chapter 12

Summary and Future Prospects

and 3 amino acids length. Next, we combined the representative library of pattern 7, 5 and 3 and developed RBP7531 and SRBP7531 approach. Both approaches are implemented in a web server FiSiPred (<http://crdd.osdd.net/raghava/fisipred/>).

The last objective of the thesis is to utilize the information of secondary structure elements and dihedral angles for prediction of tertiary structure of peptides and proteins. Previously our group has developed PEPstr, a method for prediction of peptides based upon secondary structure and β -turns. The next version PEPstr2 is based upon peptide secondary structure (PEP2D), instead of protein secondary structure method (PSIPRED). We also replaced the ideal value of dihedral angles of helix, sheet and coil with representative dihedral angles using SRBP 7531 approach. Instead of amber molecular dynamics, Modeller is used for performing molecular dynamics. We also performed an evaluation of large-scale dataset of 1273 peptides and achieve 4.41 Å RMSD. On a test dataset of 24 peptides, PEPstr2 achieved 2.3 Å RMSD as compared to 3.4 Å of PEP-FOLD. PEPstr2 better predicts the helix and coil region and for few cases, PEP-FOLD had better predicted the sheet region. For the first time, we also implemented homology-based prediction of peptides and details of homologous peptides. The PEPstr2 is freely available to scientific community at <http://crdd.osdd.net/raghava/pepstr2>.

Next, we extended the secondary structure and representative dihedral angles based structure prediction approach to proteins. We developed a method TSSPRED for homology and *ab-initio* based prediction of protein tertiary structure. The *ab-initio* method predicts tertiary structure using representative dihedral angles, SRBP7531 approach and molecular dynamics using Modeller. TSSPRED employ composite method *i.e.* combining the best of homology and *ab-initio* method for tertiary structure prediction. The homology of the unknown protein is predicted using HHsuite 2.0 package and Modeller using 70% non-redundant PDB database. The regions having no homology or weak homology were not predicted using HHsuite. Instead, we predicted these regions using the *ab-initio* prediction method. Thus, a composite approach utilizes the best of homology and *ab-initio* approaches for the final prediction of tertiary structure of proteins. User can predict the tertiary structure of proteins using five different approached of TSSPRED at <http://crdd.osdd.net/raghava/tsppred/>.

Chap

To conclu
tertiary st
creation o
developec
Using the
proteins i
secondary
prediction

The last
represent
(PEPstr2
and *ab-in*
used the
provide
available
public fr
and pred
the relati

12.1 Fu

The stud
other res
methods
develop
structur
houses :
update
perform
StarPD
structur

To conclude, we have developed a systematic approach for prediction of peptide and protein tertiary structure prediction methods. The first step is to develop a platform (ccPDB) for the creation of quality dataset of protein structure and function from centre lease of PDB. Next, we developed a web-based tool (LPIcom) for analysis of protein-ligand interacting amino acids. Using the data of ccPDB a web-based annotation tool (StarPDB) is developed for annotation of proteins using rich visualization tools. The next section of the thesis deals with prediction of secondary structure elements including peptide secondary structure (PEP2D) and β -turn prediction (BetaTPred3) and development of representative dihedral angle library.

The last section of the thesis utilizes the peptide secondary structure prediction and representative dihedral angles library for *de-novo* prediction of peptide tertiary structure (PEPstr2). A composite approach (TSPPRED) is developed which combines the best homology and *ab-initio* based prediction for protein tertiary structure prediction. To serve the public, we used the power of World Wide Web (WWW), which is the most powerful and easiest way to provide services and software to public. All the methods developed in this thesis are freely available for the scientific community and the data used in this thesis is freely available for public from their respective website and ccPDB web server. We hope the ccPDB data; analysis and prediction methods will be helpful for the scientific community in better understanding of the relationship between sequence and structure of peptide and proteins.

12.1 Future prospects

The study conducted in this thesis is based upon the work of years of research performed by other researchers around the globe. We advanced and improved the idea of previous research and methods. The work done in this thesis can be further improved, and better algorithm can be developed. A number of methods have been developed in the past for quality prediction of structure and function, but most of these methods never updated to the latest data. The ccPDB houses structure and function information in readily usable format, which can be used for easy update of structure and function prediction methods. The LPIcom tool can be further upgraded to perform docking of the ligand in proteins and modulate the ligand-protein interaction. The StarPDB tool can also be further improved to annotate the structure of protein and display the structure using visualization methods. For the first time, we developed the turn level prediction

roach; there is a lot of scope for improvement of prediction of various turns and turn level prediction approach. The concept of representative dihedral angle can be used for generation of structural alphabets of whole PDB, which can be used for the *ab-initio* prediction of peptides and proteins. We hope that the use of structural alphabets will resolve the sheet prediction problem of peptide and protein structure prediction. The structure alphabets can be used for custom design of peptide and proteins and bring improvement in *ab-initio* prediction approaches.