10.1 Summary

At present the number of known protein structures have increased enormously due to rapid advancement in the structural genomics. Protein Data Bank is a representative database of biomolecular structures contains about 86000 experimentally determined protein structures including different types of ligands. The gap between the number of reported sequences and experimental structures is continuously increasing. Identifying and predicting small molecule binding residues in a protein structures are important for understanding the function of these proteins. The knowledge of small molecule protein interactions helps in the annotation of protein functions and play important role in drug design. Target based drug designing is important in identification of new drug target and binding of drug molecules. The conventional way of target identification and analyzing physiochemical properties of binding drug molecules is a tedious process but computer aided drug design is a skilled approach to overcome the time and expensive process of conventional method. Despite advances in drug discovery process, the development a new drug is still lengthy, expensive exercise with very low success rate. In this respect, this thesis focused on developing and implementing several techniques and algorithms for the identification of ligand binding sites, predict inhibitory activity and design inhibitor analogs.

Proteins are one of the fundamental functional units responsible for many biological mechanisms in living organisms. Hence, the systematic analysis of protein structures and protein-ligand interaction provides ample reasons to understand the functions of proteins. Many important drug targets are protein, hence identifying a protein's interacting residues with ligand is extremely important to drug design process. Identifying small molecules (e.g. ATP, GTP, ADP, GDP, FAD, NAD, UTP etc.) that bind to target proteins may help in understanding their biological function and physiochemical properties. Currently number of methods have been developed to identify the structurally and functionally important residues such as small molecules interaction protein residues, protein function sites, metal binding sites, DNA/RNA binding sites and so on. The dataset used in these methods are scattered in the literature and there is need a platform to compile all these datasets. To meet these challenges computational biology helpful in addressing these issues. In order to support the scientific community, we have developed a web portal, ccPDB for providing resources for functional annotation of proteins. This portal has three major modules: i) compilation of datasets, ii) PDB tools, iii) web servers. We have developed several interfaces to create users specific datasets; compiled datasets form recent release of PDB, compilation of datasets from literature. Compilation of datasets is major module of ccPDB developed for creating standard datasets for different categories such as secondary structure (e.g. helix, beta sheet, beta-turns, gamma-turns) nucleotides (e.g., ATP, GTP, NAD, FAD) interacting, metal (e.g., Mg, Ca, Zn) interacting, DNA/RNA interacting residues. In module PDB tools, we have developed various web interfaces that include i) analysis of PDB for analyzing PDB file from more than 40 servers; ii) searching in PDB using keyword; iii) extraction of PDB chains. Web servers module allow users to perform various task that includes; i) prediction servers for identification of nucleotide interaction residues, ii) generate patterns for creating patterns, iii) BLAST search for similarity search and iv) benchmarking option for evaluating methods. ccPDB is a unique resource to obtain all necessary information regarding the structure and function of proteins at amino acid level.

The analysis and identification of interactions between proteins and small molecules is a crucial step to understanding many biochemical processes, so it plays a critical role in drug discovery. Many important drug targets are protein, hence identifying a protein's interacting residues with ligand is extremely important to drug design. Identifying interacting residues or binding site between small molecules and target proteins may help in analysis their biological function and physiochemical properties. Thus finding and predicting small molecule-binding residues in protein structures are important subjects in drug discovery field. One of the fundamental questions is to understand how these ligand molecules interact with specifically recognized proteins and where the particular residue interacts or binds. During last few years, many small molecules binding proteins has been discovered due to progress of large number of sequencing projects. But till now annotation and identification is a challenging task for computational biology. We can answer these questions only when we have a better understanding of interaction between protein and ligand using machine learning techniques.

One of the major challenges in post-genomic era is to provide functional annotations for large number of proteins arising from genome sequencing projects. The function of many proteins depends on their interaction with small molecules or ligands. ATP and GTP are such important ligands that plays critical role as a coenzyme in the functionality of many proteins. There is a need to develop method for identifying ATP and GTP interacting residues in proteins, in order to understand mechanism of protein-ligands interaction.

Biological activities of small natural and drug molecules are directly or indirectly depends on its physiochemical properties. A number of prediction methods have been developed in the past for the prediction of physiochemical properties of small molecules. Most of these methods are specific for particular group of compound and their performance was not significantly high. *In silico* prediction of physiochemical properties of small molecules are very important, because experimentally calculation of these properties is very difficult and time consuming. So, highly accurate computational/Chemoinformatics method required to calculate the physiochemical properties of small molecules. In this study we have first developed QSAR model using selected chemical descriptors to predicting inhibitory activity of EGFR inhibitor. We have also used structure-based approach (docking) and calculate docking energy and developed docking energy based prediction model. Here, we have concluded that ligand based approach (QSAR) is much better than docking approach. Next, we have integrated both QSAR and docking approaches by using docking generated energy-based scores as descriptors for QSAR modeling.

An analog based drug designing is a new field in drug discovery. The structure of an existing molecules or inhibitors is the basic requirement for analog designing. The structure of an existing molecules or inhibitors is obtained from Drug bank, PubChem or any others drug databases and finally select the best drug compounds to generating their analogs. Some database contains compounds in Mol or SD file format but some direct structure. Compounds, which contain only structure, its structure converted in file format with the help of Chemoinformatics tools. Finally, performed docking of analog molecules with target protein for optimization. For optimization we have used AutoDock, AutoDock Vina and DOCK softwares. Finally, selected few top scored analogs and calculate their ADMET properties. Finally these top scored analogs can be experimentally validated.

10.2 Future prospects

The ccPDB is a important tool for Bioinformaticians to use in the annotation the structure and function of proteins. In ccPDB, we created and maintained important data sets from various releases of PDB. In addition, we also developed a series of web-based tools for creating new data sets. Therefore ccPDB would be a good encyclopedia for better understanding of ligand-binding protein sequences and structure-functional annotation, which can be used in many drug discovery applications.

The identification of novel drug targets and their inhibitors is a major challenge in the field of drug designing and development. The ATPint and GTPbinder methods will be helpful in the identification of ATP and GTP interacting protein residues. These tools also helpful in understanding and analysis of ATP and GTP binding protein sequence. So these tools would helpful in analysis of large number of ligand binding protein and explore their binding sites. These tools would also provide information regarding diversity and nature of ATP binding sites. ATPint method would helpful in identification of different types of kinase targets. GTPbinder tool would helpful in identification and exploration of G-protein coupled receptors protein families. The identification and targeting of interacting residues of proteins can be used for designing of drug molecules against diseases. Thus, this protein-small molecule interaction knowledge would prove useful in understanding various biological processes and also would help in therapeutic implication.

The selection of important chemical descriptors and docking energy based QSAR models for predicting EGFR and their mutant inhibitor would be profitably employed to design new inhibitors with better inhibitory activity values. We feel that the performance of these models would certainly improved from additional quantitative data that need to be generated from biochemical experiments, producing better models for the prediction of inhibitory activity values. The selected few highly significant and important chemical descriptors would help the medicinal chemist to analyze physiochemical properties and biological activities. The webserver “ntEGFR” developed in this thesis will shed light on the way to design potent EGFR inhibitors. Briefly we can say that such research will generate the knowledge required for the rational design of novel EGFR inhibitors. The analogs and database screened EGFR inhibitors would represent novel scaffolds/candidates represent promising starting points as lead compounds and certainly aid the experimental designing of potent EGFR inhibitor. Moreover, the present work would certainly enhance the outcome of experimentalist working in field of protein.