SYNOPSIS

Biological processes are governed by highly specific macromolecular interactions. Understanding the precise mechanism of ligand recognition in proteins is essential for deriving features responsible for such recognition capabilities. Although protein sequences give first-hand information about their function, their three-dimensional structures, which are the evolutionary units, convey the function better. Three-dimensional structures of many proteins determined through X-ray crystallography and/or NMR are available in the Protein Data Bank, a public repository. This resource has stimulated the development of computational techniques to read and analyse the wealth of structural data. Structural bioinformatics is an area that provides a means to transform information in the protein structures into functional insights and enable addressing a variety of questions about what defines and dictates ligand recognition. Large-scale analyses of several protein-ligand complexes have indicated that both one-to-many and many-to-one relationships of protein-folds and ligand-types are widely seen in the PDB. This means that a given ligand can be recognized by diverse proteins and a given protein can recognize different types of ligands at the same location, ligands referring to endogenous ligands, natural metabolites as well as small molecule inhibitors, and drugs. Given this, it is important to understand the determinants of recognition of a given ligand. This becomes important for applications in drug discovery that includes, lead design and lead optimization, assessment of draggability of a target, identification of off-target effects, polypharmacological targets and drug repurposing. The present work utilizes the information present at the functional sites, rationalizing many examples of ligand binding and deriving useful patterns that can be used for genome-wide function annotation and drug discovery applications.

A large-scale analysis of the binding of two important classes of ligands, a sugar and nucleotides was carried out by analysing the sub-structures at their binding sites by matching, aligning and clustering. The two ligands studied are sialic acid, and nucleoside mon/di/tri phosphates (nucleotides or NTPs), for their binding to many proteins reported in the PDB.
Sialic acid was found to be recognized by 170 different proteins representing 17 unique sequence families. Our approach deciphered a unified understanding of the basis of recognition of this ligand and showed six structural motifs, which contained different combinations of one or more key structural features, over a common scaffold. The site features refer to certain residues in the binding site that are seen to most frequently occur at their respective topological positions, a result that was evident upon binding site comparisons and 3-D alignment of sites in the different proteins.

In the case of nucleotide ligands, 4,677 structures of protein-nucleotide complexes from PDB, belonging to 145 different structural folds and 394 sequence families were analysed, and our results indicated that the sites from diverse proteins group into 27 site-types and further into nine super-types having a structural motif for each site-type. The identified motifs were highly specific when scanned against the entire PDB. A computational alanine mutagenesis study indicated that residues identified to be highly conserved in the motifs also contribute most to binding. Alternate orientations of the ligand in several site-types were observed and rationalized, indicating the possibility of some residues serving as anchors for NTP recognition. Many examples of convergent evolution were identified through this analysis.

Next, in order to find the entire set of all binding sites in the proteins that have known 3-D structures, a large-scale analysis of 30457 representative protein structures in PDB was carried out by detecting additional binding sites in them, followed by a site-comparison to recognize unexpected similarities. The pocket-space currently defined in PDB is incomplete, as binding-sites remain uncharacterized in many proteins despite the availability of their structures. To bridge this gap, we computationally detect pockets in the entire PDB and present a comprehensive resource of an 'augmented pocketome' consisting of 249,096 pockets (*http://proline.biochem.iisc.ernet.in/PocketDB*), which shows a 2-fold increase in fold coverage and a 7-fold increase in pocket-space when compared to what is currently known. Possible ligand associations for about 56% of these pockets were deduced. A clustering of the pocketome led to the identification of 2,161 site-types, and the associated ligands into 1,037 ligand-types, which together provided fold-to-site-type-to-ligand-type associations. The resource facilitates a structure-based function annotation, delineation of a structural basis of ligand recognition, and provides functional clues for Domain of Unknown Function (DUFs), allosteric proteins and druggable pockets.

Next, it was of interest to study the entire binding site-space in a proteome of a given organism. *Mycobacterium tuberculosis (Mtb)* is a deadly pathogen being studied from multiple perspectives in the laboratory. Structural models of about 70% of the proteome and a set of 13,858 small molecule binding pockets were already available from a previous study in the laboratory. To systematically identify druggable targets in this category, a genome-wide screen was carried out to comprehensively identify all proteins binding to NTP ligands. Being equipped on one hand, with the knowledge of binding site motifs, and on the other, with the structural models of *Mtb* proteins at a genome-scale, this exercise was made possible. A total of 1,768 proteins in *Mtb* (about 43% of the proteome) were predicted to bind NTP ligands, which constitute the NTPome. Using an experimental proteomics approach involving dye-ligand affinity chromatography, 47 different proteins were validated, of which four are hypothetical proteins. This analysis also provides the precise list of binding site residues in each case, and the probable ligand binding pose. As the list includes several known and potential drug targets, the identification of NTP binding can directly facilitate structure-based drug design of these targets. Further, 305 proteins in the NTPome were identified as important drug targets.

A successful drug discovery program stands on the first principles of good target identification. Previous studies carried out in the laboratory combined with contributions from this thesis work, a promising target, nucleoside diphosphate kinase (NDK), that possesses many of the desirable properties of an ideal target, was identified. Using the NTP motifs previously derived, the nucleotide pocket in NDK was studied for the most important residues that confer binding, thus prioritizing the pocket residues for mediating crucial interactions with the ligand. Next, a screen of the nucleotide pocket of NDK against the drug-binding sites known in PDB resulted in identifying a handful of potential screening candidates that contained several elements of an ideal ligand for the NDK pocket. A thorough biochemical characterization using protein purification and subsequent quantification by a spectrophotometric coupled assay, led to the experimental testing of the identified candidates against NDK enzyme activity. This resulted in identifying enalapril and cloxacillin as NDK inhibitors. Of the two compounds, enalapril was efficient in 100% inhibition of NDK activity, while cloxacillin showed a maximum inhibition of 78%. A test for inhibition of *Mtb* H37Rv growth resulted in determining cell inhibition for both compounds, of which cloxacillin was known previously. Thus, by using the principles of structure-based drug design, two lead molecules were identified, which are reported for the first time for *Mtb* NDK.

Thus, using a structural bioinformatics approach, a large-scale characterization of small-molecule binding sites has been carried out in this thesis work. The work starts from identifying ligand recognition principles of proteins and understanding the molecular capabilities of ligand binding in diverse proteins, which forms the crux of protein-ligand interaction profiling. Using the sub-structures for recognizing function in uncharacterized proteins, the work carried out here presents a pipeline for function annotation. One goal of understanding the functional capabilities of macromolecules is to design and develop drugs, which have also been dealt with in this thesis work, housing the principles of target identification, druggability and finally, lead identification