Structural Bioinformatics: Transforming protein structures into biological insights

Yeturu Kalidas AND Nagasuma Chandra

Abstract | Structural bioinformatics is an area that has emerged to comprehend and interpret large amounts of structural data, and promises to provide a high resolution understanding of biology. Protein structures can be compared, analyzed and mined in various ways, which allows us to understand the functions of these molecules and reason precisely how and why such capabilities have emerged in them. The main advantages these methods have over simpler sequence based methods are that besides helping in associating a molecule with a function, they also provide ultimate insights into the mechanisms by which various biological events take place. This report provides an overview of structural bioinformatics, various advances in the recent years and the range and scope of data driven protein structural analysis. In particular, current trends in structure prediction, structure alignments, deriving sub-structures and structural motifs, understanding features critical for molecular recognition as well as using these for understanding function of protein molecules are presented. Application of structural knowledge in drug discovery for lead identification as well as novel ways of understanding drug adverse effects and drug resistance are also discussed. Finally prospects for structure based vaccine design are also outlined. The various aspects of structural bioinformatics discussed here, show how biological insights can be obtained from protein structures.

1. Introduction

The advent of the 'omics' era that began with high throughput gene sequencing has led to deciphering complete genome sequences of thousands of organisms (Kyrpides 1999; Liolios *et al.*, 2008, for example). The obvious next step in comprehending this huge pool of data and for their application across all life-science disciplines would be to understand the function of each of the gene products. The quest for understanding protein function has resulted in a rapid accumulation of structural data, with more than 50,000 entries in the Protein Data Bank (Berman *et al.*, 2000). Needless to say, the highest resolution of information of the function of the protein is obtained through the understanding of their three-dimensional structures. While sequence data has become easier to derive, with millions of proteins now sequenced, structural data that comes from X-ray crystallography and from the NMR experiments, numbers only about 52,103 (http://www.rcsb.org/) today. The available genome sequences in fact indicate that the sequence space sampled by these is already nearly complete, up to the level of tetra-peptides (Poddar *et al.*, 2007), with most of the 160000 different tetra-peptides, 8000 different tri-peptides, 400 di-peptides and certainly the 20 different amino acids occurring as segments in some protein or the other. To bridge the wide gap between sequence and structure, various computational methods have emerged that can predict the structure of a protein molecule with high confidence in many cases (Pillardy *et al.*, 2001;

Bioinformatics Centre, Supercomputer Education and Research centre, Indian Institute of Science, Bangalore
nchandra@serc.iisc.ernet.in

Sanchez and Sali 1997; Unger 2004; Jones *et al.*, 1992; Sun 1993). Applying such molecular models have led to a significant increase in structural data, which now approaches millions. The need to navigate and comprehend this large resource of experimental and theoretical structural data, has automatically led to the genesis of a new discipline called *structural bioinformatics* (Burley 2000; Bourne and Weissig, 2003), which has become well established in the last decade. *Structural Bioinformatics* is probably best thought of as the discipline, which, rationalizes, and classifies information contained in the three dimensional structures of molecules, in terms of their functional capabilities, thus ultimately helping in understanding at atomic level detail, how biological organisms encode, make use of, and pass on information. The main advantages these methods have over simpler sequence based methods are that besides helping in associating a molecule with a function, they also provide ultimate insights into the mechanisms by which various biological events take place.

In principle, the term could encompass all biological macromolecules, but is used predominantly in the context of protein molecules. Comparing proteins, deriving structural patterns, correlating with function and ultimately utilizing such patterns for prediction are all integral components of structural bioinformatics. Given the complexities involved in solving new crystal or NMR structures of protein molecules, it might often feel like a successful end to a long struggle, but in reality a structure is just the beginning of a journey to understand the function of the protein molecule. Structural bioinformatics is an important area that serves as a bridge in transforming protein structures into biological insights. This report provides an overview of the advances in the discipline, with a focus on the computational methods that have been developed over the years along with a glimpse of the applications that emerge out of such capabilities, often using examples from work carried out by our group. An overview of various aspects of structural bioinformatics is illustrated in Figure 1.

2. Understanding function through protein structure

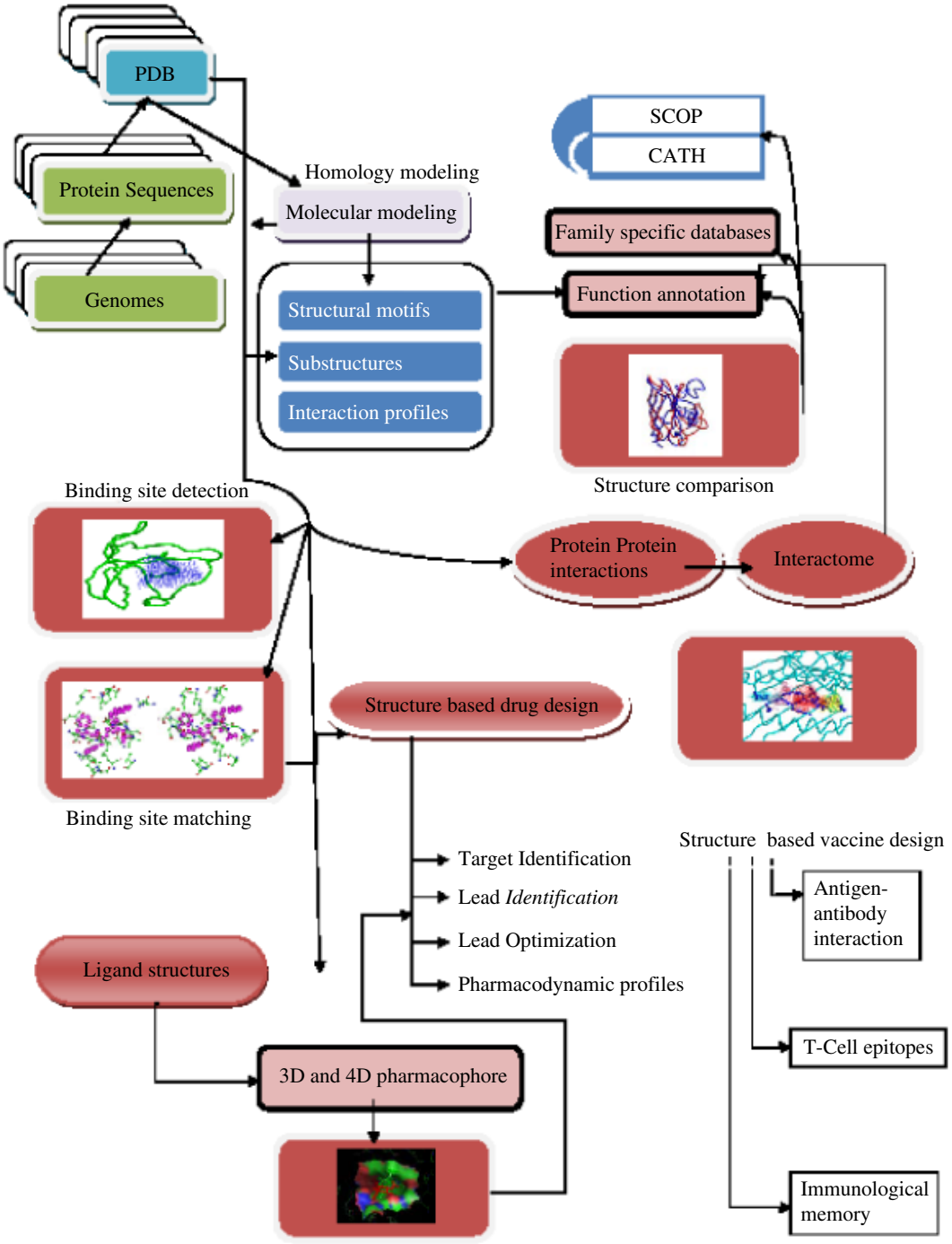
Ever since classic experiments by Christian Anfinsen (Anfinsen 1959), sequences have been known to contain the information required to fold the protein molecule and dictate its function. Since sequence data is more readily available, the 'sequence-function' paradigm is heavily used, in which, the sequence of a particular protein is compared with other related sequences in databases to infer their functional role(s). Derivation of sequence-structure-function relationships in protein molecules is a fundamental objective in bioinformatics. In fact, much of present day biology is dependent on the knowledge of such relationships. Biology, in its present practice, is effectively a relational science, decisions made with one system being heavily influenced by the knowledge obtained from other systems. It is quite understandable therefore, why recognizing similarities and deriving relationships are crucial for all further knowledge. This need is not only heightened, but is also rendered feasible with the large numbers of genomes being sequenced in the last few years. Where available, protein structures provide much better functional insight, than their sequences alone.

2.1. *Molecular modelling*

A number of methods have been developed in the last couple of decades to model the structures of protein molecules. Of these, homology modelling seeks to predict the structure of a protein by using a structural template of a homologous sequence, in cases where such a template is available. This is based on the premise that two sequences who are homologous also share the same structural fold. An analysis of the protein fold space indeed reveals that the available protein structures cluster only into certain regions of the entire space (Holm and Sander, 1996), indicating that several protein families share common structural folds even where they do not share sufficient sequence similarities. Energy minimisation that uses molecular mechanics based force fields and in some cases also molecular dynamics simulations, are then used to refine the initial models obtained by using the templates. This methodology is well established now and is beginning to be used in a high-throughput manner (Pieper *et al.*, 2004) {http://salilab.org/modbase/} to model entire proteomes (Peitsch 1997). The different methods vary mainly in terms of positioning of side chains, loop building, treatment of neighbourhoods, force-field parameters and in model refinement techniques (Sanchez and Sali 1997). The success seen at the popular CASP experiments conducted once every two years stand testimony to the advances in this area and to the confidence one can have in models thus built (Moult *et al.*, 1995; 2007).

One of the first structural bioinformatics analyses to be carried out, although not called by that name at that time, is the computation of the Ramachandran map (Ramachandran *et al.*, 1963), which provides a rational basis for describing stereochemically possible structures of polypeptides. In this, the 'structure space' of protein chains is reduced to two-dimensions, by representing a structure in terms of the torsion angles of the

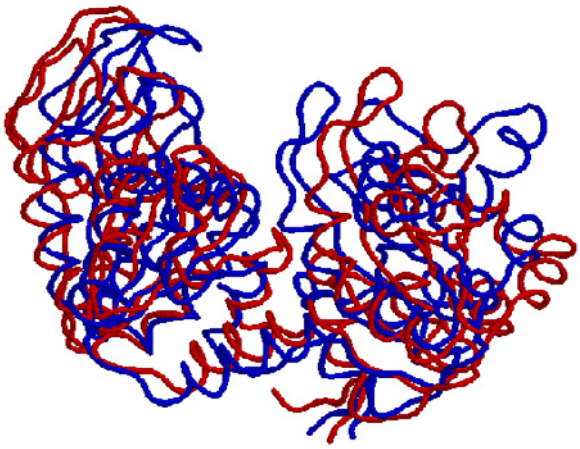
Figure 1: An overview of different aspects of structural bioinformatics.



protein backbone. Today, this map is used as an integral part of structure determination, in order to estimate the quality of protein structures. As a conceptual extension to this analysis, analysis of side chain conformations in proteins (Bhat *et al.*, 1979), design of rotamer libraries for use in molecular modelling (Dunbrack and Karplus, 1994) and structure validation and several other analyses, that are used quite routinely in crystal structure refinement and for quality estimation (Laskowski *et al.*, 1993).

Hundreds of models of protein structures have been built and used for obtaining a variety of biological insights. An example of the use of

Figure 2: Comparison of the inactive 'open' state (red), with the active 'closed' state (blue) of phosphoglycerate kinase. A large domain motion occurs upon binding of the two substrates.



this technique, is the molecular modelling of the closed conformation of a ternary complex of phosphoglycerate kinase (Chandra *et al.*, 1998, Figure 2), which indicated that upon substrate binding a large conformational change would be essential to facilitate catalysis, a prediction that was validated by a crystal structure of the closed form of the enzyme from *T. brucei* (Bernstein *et al.*, 1998). Another example is the molecular model of the assembly of the chromatosome particle, which has lead to the understanding of the nature of interaction of the globular domain and the functional role of the C-terminal domain of the linker histone, providing clues to certain important factors in chromatin formation (Bharat *et al.*, 2003). Yet another example is the homology modeling studies of δ-Aminolevulinic acid dehydrase from *Plasmodium falciparum,* which provided a rationale to explain unique properties of this enzyme, such as alkaline pH optima and apicoplast localization, making it resemble the plant enzyme, while at the same time manifesting certain unique properties such as Mg2+-independent and EDTA resistant catalytic activity (Shanmugham *et al.*, 2004). There are also a number of examples in literature where molecular models have been used in drug discovery, either at the lead design or at the lead optimisation stage (Tanrikulu and Schneider 2008). An early notable example of that is the design of 'captopril', an anti-hypertensive drug that inhibits angiotensin converting enzyme (ACE), based on structural clues obtained from functionally analogous carboxypeptidase (Ondetti *et al.*, 1977). With the complete sequencing of several genomes, as comparative genomics becomes feasible, direct clues about sets of proteins are obtained, leading to rational target identification and rational design of lead compounds, both critical steps in drug discovery (Raman *et al.*, 2007) .

Some relationships among proteins at the fold level are readily identified due to the sequence similarities among them. However, in many cases the sequence similarities can be very low and thus relationships not obvious. It is now well accepted that conservation at the structure level can be much higher and thus more detectable than at the sequence level (For example, 1B3A and 1TVX have low sequence identity but high structural similarity (Lo *et al.*, 2007)). In a different context, this issue is hotly debated to determine whether such molecules are the result of convergent evolution or actually products of divergent evolution but the divergence is so high that they cannot be recognised. Nevertheless, many more structures can be predicted by recognizing with which of the known folds a given sequence is most compatible; a technique popularly known as threading (Jones *et al.*, 1992; Rost *et al.*, 1997) or by comparing profiles of sequences that contain clues about their preferred neighbourhoods. Threading works by winding the query sequence on to the fold of a template backbone from a database of folds and evaluating the feasibility of the threaded structure in terms of geometric and chemical compatibility, through the measurement of all pair-wise interaction potentials of the individual residues (Jones *et al.*, 1992). The profile based methods simultaneously compare multiple features of a protein that captures structural environment of each residue, using dynamic programming methods

and are complementary to the threading methods. They have been applied in a variety of cases, which have led to understanding aspects such as the functional family, a protein belongs to or the ligand the given protein is most likely to recognize. A last category of modelling is that of ab-initio structure prediction, which can be achieved without any structural templates, since it is based on the premise that the native conformation of the protein will have a global minimal energy and hence appropriate computational methods should be able to find that conformation through a thorough search (Pillardy *et al.*, 2001). This approach however has many practical difficulties due to the combinatorial nature of the possible arrangements of each residue with respect to the next, in three dimension as well as the difficulty in discriminating real structures from the decoys and thus cannot as yet be used as a routine technique.

2.2. *Function annotation*

Once a structure is available, its function can be obtained in some cases by comparison of its fold to that of another protein whose function is already determined experimentally and transfer of that functional knowledge to the new protein. Function itself can be defined at different interdependent levels, the two most important of them being (a) the level of molecular function which includes binding of a particular ligand and catalysis of a particular reaction, and (b) the level of the biological process, which refers to the larger function of the protein. For example, the function of the RecA protein could be described as ATP binding and DNA binding at the first level and as a component of homologous recombination and DNA repair at the second level. 'Fold to function' models have been the basis for functional annotation of proteins in some cases. When two proteins exhibit high structural similarity along their entire polypeptide chains, they are likely to have similar functions, both at the molecular function level as well as at the biological process level. When two proteins exhibit only a part similarity in their structures, their functions are not necessarily the same and more detailed studies would be required to infer function, as described later. Part similarity can exist in two broad ways; (i) medium-to-high similarity in a portion of the polypeptide chain, indicating the presence of a common domain in the two proteins or (ii) low-to-medium level similarity in most part of the polypeptide chain. For the first category, inferring molecular level function would be possible for the conserved region in many cases, but inferring biological process level function would not be possible. For the second category, functional inference at either level would not be meaningful since fold level similarity does not necessarily imply conservation at the functional regions of the molecule and hence does not also imply conservation in function, especially at the level of the biological process. Thus, structure to function models work best when there is high conservation in the entire protein, application of which are described several times in the literature. An interesting example is the annotation of function of Rv3214 from *Mycobacterium tuberculosis* as a broad-spectrum phosphatase, important for mycobacterial phosphate metabolism in vivo (Watkins and Baker 2006). This protein was originally annotated as EntD through sequence similarity with the *Escherichia coli* EntD, a 4′-phosphopantetheinyl transferase implicated in siderophore biosynthesis. After solving its crystal structure as part of a structural genomics initiative, closer comparisons of structure and sequence indicated the protein to be a phosphatase belonging to the dPGM superfamily, later confirmed by biochemical experiments. Another example of obtaining biological insights through structure is that of Rv1347c, a putative antibiotic resistance protein from *Mycobacterium tuberculosis*, which revealed a GCN5-related fold, suggested an alternative function in Siderophore Biosynthesis, rather than its annotation as a putative aminoglycoside *N*-acetyltransferase (Card *et al.*, 2005).

3. Structure comparison and classification

3.1. *Algorithms*

An essential pre-requisite for inferring function from structures is to compare them and use appropriate metrics to describe structural similarity. While comparing protein molecules through their sequences has now become a well established routine task in most cases, structural comparison of protein molecules still remains a challenge. Matching 3D objects in any field is a non-trivial matter. For proteins, additional complexity arises from the need to compare molecules of different sizes, need to consider insertions and deletions, commonly known as 'indels' as well as non-topological similarities. Many protein structure comparison algorithms have been proposed for estimating the extent of similarity between two proteins. A majority of them consider backbones corresponding to each of the proteins and align them by defining a set of equivalences between pairs of atoms between the two proteins. Equivalences between methods can be derived at by any of the strategies - dynamic programming, distance matrices, fragment matching, geometric hashing, maximal common subgraph detection or local geometry matching.

DALI (Holm and Sander 1993) for example uses distance matrices, CE (Shindyalov and Bourne 1998) uses combinatorial extension of alignment path, the method by Taylor and Orengo (1989) uses dynamic programming, that by Szustakowski and Weng (2000) uses genetic algorithms, that by Zhu and Weng (2005) uses maximal common subgraphs between proteins represented as graphs and that by (Krissinel and Henrick 2004) aligns matching of secondary structural elements followed by local refinement to align Cα atoms. DALI represents a protein structure as a 2D distance matrix that considers distances between all pairs of Cα atoms. The matrix hence formed becomes a frame invariant representation, containing sufficient information for reconstruction of the 3D object except for possible loss of chirality. An elegant scoring function is used to score pairs of fragments with matching distances, to finally obtain a score indicating the extent of similarity. Commonly used metrics for comparing structures are root mean squared deviation, Z- scores that indicate quality of alignment that overcomes some of the drawbacks of the RMSD metric. The dynamic programming by Taylor and Orengo 1989 is similar to that of Needleman and Wunsch for sequence alignment (Needleman and Wunsch 1970), but has the drawback of requiring huge computational resources-time and memory. The maximal common subgraph detection by Zhu and Weng (Zhu and Weng 2005) involves incremental construction of the graph between pairs of Cα atoms and uses local geometric properties to arrive at pairs of nodes, assigns edges by directionality based scoring scheme, iteratively prunes the bad vertices and finally uses dynamic programming to arrive at final alignment on this simplified graph. Unfortunately, the formulations have turned out to be NP-Hard (Zhu and Weng 2005), leading to the development of many heuristics. Two main issues about protein structure comparison algorithms are, to what extent are *indels* tolerated and whether *non-topological* similarities are detected. MatchProt, a new fast algorithm developed addresses some of these issues (Bhattacharya *et al.*, 2006). The formulation involves a novel method of characterization of the residues of a protein in the context of its overall structure by projecting them on the real line in a neighborhood preserving way. This characterization is used to define a similarity function between the residues of two proteins and find the optimal equivalences. Non-topological similarities in a set of circularly permuted proteins are identified between sets of proteins efficiently, resulting in a more realistic estimation of their extents of similarity than many other algorithms available for that purpose. Various algorithms available for structural comparison and other analyses are indicated in Table 1.

3.2. *Classification*

As many protein structures became available, Murzin and co-workers (Murzin *et al.*, 1995; Andreeva *et al.*, 2008) made an insightful classification of about 3000 protein structures available at that time, by visual comparison guided by their intuition and organized protein structures into a database called SCOP. A hierarchical organization was used, consisting of four levels: the structural class, super-family, family and fold. Each protein is described at these levels. About 405 unique folds were observed at that time from about 6500 structures, which has grown today into more than 1086 folds, 1777 super-families and 3486 families (SCOP-1.73 release). Subsequently Thornton and co-workers developed a classification scheme and a resulting database called CATH (Orengo *et al.*, 1997). In this, structures are also described based on a hierarchical organization, but are compared with each other by using structural comparison algorithms. These databases are most useful resources for understanding a protein structure and are heavily used by structural biologists and bioinformaticians. Various databases of protein structures and their derived features are indicated in Table 2. PALI, a database of phylogeny and alignment of members of SCOP families, (Gowri *et al.*, 2003), SMotif- a database of structural motifs in proteins (Pugalenthi *et al.*, 2007), CAMPASS, a database of structural superfamilies (Sowdhamini *et al.*, 1998), are examples of databases resulting from structural bioinformatics analyses. Several tools to extract various structural features and probe their roles in stabilizing the structure or imparting function, have also been developed that enable such analysis over the internet at great ease (Ananthalakshmi *et al.*, 2005). Classification of protein sequences and structures into families is a fundamental task in biology. Some relationships are detected by the similarities in their sequences, many more by the similarities in their structures.

3.3. *Family specific databases*

Specific classification schemes and databases for many protein families have also been developed which present structure-function relationships in great detail. One such example from our work is the lectin knowledge base (Chandra *et al.*, 2006; http://proline.physics.iisc.ernet.in/lectindb/ which classifies 941 unique plant lectins from 241 different plants into 7 fold types. Multiple alignments within each fold class have been carried out, followed by phylogenetic analyses, which are useful to understand the extent of divergence in detail and hence subtle but definite functional differences/adaptations, within each fold. A database

Table 1: Examples of Algorithms for various structural bioinformatics operations

Operation	Description and References
Structure Prediction	Threading (Jones*et al.*, 1992) Optimization of potential energy function (Pillardy *et al.*, 2001) Threading (Rost *et al.*, 1997) Genetic algorithm(Unger 2004)
Structure Comparison	Projection of residue vectors (Bhattacharya *et al.*, 2006) Graph theoretic comparison (Harrison *et al.*, 2003) Alignment of distance matrices (Holm and Sander 1993) Matching secondary structural elements (Krissinel and Henrick 2004) Combinatorial extension of alignment path (Shindyalov and Bourne 1998) Genetic algorithm (Szustakowski and Weng 2000) Dynamic programming (Taylor and Orengo 1989) Flexible structure alignment (Ye and Godzik 2003) Graph theoretic comparison (Zhu and Weng 2005)
Substructure Retrieval	Spatial motifs (Kleywegt 1999) Matching secondary structural matrices (Shi *et al.*, 2007) Depth first search (Stark and Russel 2003) Hydrogen bond signatures (Prasad *et al.*, 2004)
Binding Site Detection	Energy based clustering of probe grid cells (An *et al.*, 2005) Scanning of 3D grid (Hendlich *et al.*, 1997) Scanning of 3D grid and residue conservation information (Huang and Schroeder 2006) Depth based clustering of grid cells (Kalidas and Chandra 2008a) Delaunay triangulation (Kleywegt and Jones 1994)
Estimating ligand recognition	Support vector regression based (Bock and Gough 2002) Spherical harmonics matching of shape complementarity (Cai *et al.*, 2002)
Protein-Protein Interaction	Neural network based (Fariselli *et al.*, 2002) Surface patch analysis (Jones and Thornton 1997)
Docking	Genetic algorithm "Autodock" (Morris*et al.*, 1999; Khodade*et al.*, 2007) Geometric hashing for docking of ligands "FlexX" (Rarey*et al.*, 1996) Fragmentation based docking of ligand "LUDI" (Bohm 1992) Protein-protein docking based on shape complementarity (FTDOCK) (Gabb*et al.*, 1997)

Table 2: Examples of primary and derived structural databases available

Database	URL	Reference
Protein Data Bank (PDB)	http://www.pdb.org	Berman *et al.,* 2000
The Macromolecular Structure Database (MSD)	http://www.ebi.ac.uk/msd/	Henrick *et al.,* 2003
Fold Classification based on structure-structure assignments (FSSP)	http://ekhidna.biocenter.helsinki.fi/dali$_{s}erver/$	Holm and Sander 1993
Structural classification of proteins (SCOP)	http://scop.mrc-lmb.cam.ac.uk/scop/	Murzin *et al.,* 1995
Class architecture topology and hierarchical classification of proteins (CATH)	http://www.cathdb.info	Orengo *et al.,* 1997
Protein Function Prediction ProFunc	http://www.ebi.ac.uk/thornton-srv/databases/ProFunc/	Laskowski *et al.,* 2005a,b
PDB-Ligand	http://www.idrtech.com/PDB-Ligand/	Shin and Cho 2005
PDB ligand search	http://www.ebi.ac.uk/msd-srv/msdmotif/chem	
Database of biologically relevant ligand sites (LigASite)	http://www.bigre.ulb.ac.be/Users/benoit/LigASite/	Dessailly *et al.,* 2008
Structural motif databases (MALISAM)	http://prodata.swmed.edu/malisam/	Cheng *et al.,* 2008
Patterns in non-homologous tertiary structures (PINTS)	http://www.russell.embl.de/pints/	Stark and Russell 2003
Database of phylogeny and alignment of members of SCOP families, (PALI)	http://pauling.mbu.iisc.ernet.in/ pali/	Gowri et al., 2003
Database of structural motifs in proteins (SMotif)	http://caps.ncbs.res.in/SMotif/index.html	Pugalenthi et al., 2007
Database for prediction of protein-protein interactions (POINT)	http://insilico.csie.ntu.edu.tw:9999/point/	Huang *et al.,* 2004
Database of surface patches (SURFACE)	http://cbm.bio.uniroma2.it/surface/	Ferre *et al.,* 2004
Database of conformational angles in protein structures CADB-3.0	http://cluster.physics.iisc.ernet.in/cadb/	Gopalakrishnan et al. 2007

analysis of b-prism-I or jacalin-like lectins shows that hyper-variability in the binding site loops generates carbohydrate recognition diversity. The study, also predicted that a spectacular variety of quaternary associations would be possible in this family of lectins that have implications for glycan recognition, a new type of association subsequently proved to exist in the crystal structure of banana lectin (Singh *et al.*, 2004). The study also helped in identifying that BL3, a loop housing a conserved aspartic acid in the binding site of this family of proteins, with its invariant conformation to be a determinant of lectin activity, by generating carbohydrate recognition capability, whereas a multitude of factors such as sequence and lengths of other loop regions in the vicinity dictate specificity (Raval *et al.*, 2004).

3.4. *Machine learning algorithms for classification and structure prediction*

Machine learning approaches help in identifying the features whose selection leads to higher accuracies of prediction and hence providing biological insights. Support vector machine (SVM) is an optimization formulation strategy of classifying a given set of points in feature space into two classes. SVM based classifiers have been used in a variety of aspects of structural biology, for example fold level classification of proteins (Shamim *et al.*, 2007), prediction of protein–protein interaction regions on the surface of protein (Bradford and Westhead 2005), prediction of MHC binding peptides (Donnes and Elofsson 2002), turn prediction of amino acid sequences (Meissner *et al.*, 2008). Neural networks are based on arriving at a relationship between input features to result in the required output. The relationship between features often tends to be non-obvious. (Holley and Karplus 1989) present a neural network based approach for determination of secondary structures in proteins. (Bock and Gough 2002) use support vector regression for estimating the free energy of binding for a ligand for use in virtual screening. There have been several methods developed to derive various structural features from protein sequences as well, examples of which are prediction of secondary structures (Raghava *et al.*, 2002) and prediction of trans-membrane helices in proteins (Ganapathiraju *et al.*, 2008).

4. Function annotation through Sub-structure derivation and comparison

What ultimately matters for a biological system to perform its role is the ability for a given protein to do a particular task, but not whether a given protein has a particular sequence or structure. There are a number of examples in the literature, which illustrate that structures convey the 'meaning', more efficiently than sequences, here 'meaning' referring to the 'function' of the protein. On the other hand, there are also a number of instances, which illustrate that a particular 'function' is achieved by proteins whose sequences and structures are dissimilar. For example, at least three different proteins with different folds and architectures recognize mannose and exhibit mannose-mediated physiology (Ramachandraiah and Chandra, 2000). In other words, structures also fail to convey the 'meaning' in many cases. We do not yet know if this failure is because of our inability to recognize any similarities in such seemingly dissimilar proteins or it is simply because no similarities actually exist among them.

Obviously, the success in deriving various relationships is dependent on the methods used. There are a number of sequence-based methods such as BLAST and FASTA, which are used routinely today for identifying sequence homologues. Newer ways of comparing molecules and recognizing similarities at various levels has been an area of intense research, resulting in progress in many fronts, such as the evolution of pattern recognition methods applied to sequences (e.g., PSI-BLAST, PRINTS, development of various substitution matrices for use with database searching and alignment protocols (BLOSUM), as well as in the emergence of various fold-recognition (Genthreader (Guffin *et al.*, 2000), 3D-PSSM(Kelley *et al.*, 2000) etc) and structure comparison methods (DALI, VAST (Madej *et al.*, 1995)). Most of the sequence alignment methods are based on recognizing common sequence patterns where as the structural alignment methods are based on recognition of common topological arrangement of sub-structures (such as the secondary structural elements).

4.1. *Interaction fingerprints*

Based on the premise that the precise 3-dimensional disposition of key residues in a protein molecule is what matters for its function, or what conveys the 'meaning' for a biological system, but not what means it uses to achieve this (Prasad *et al.*, 2003; 2004), the concept of comparing two molecules through their intra-molecular interaction networks was explored, since these networks dictate the disposition of amino acids in a protein structure. For this, a method (HBPRINT) of comparing molecules through their interaction patterns was developed. Signature patterns, or fingerprints, of interaction networks in pre-classified protein structural families are computed using an approach to find structural

equivalences and consensus hydrogen bonds. Such fingerprints, when used as features to compare protein molecules, have resulted in the identification of new unexpected similarities. Structures of garlic lectin and Charcot Leyden crystal protein belong to two different folds, do not share any significant sequence similarity, yet show similarities in their interaction patterns, but interestingly have similar function- that of carbohydrate binding (Prasad *et al.*, 2004). The problem however at this point of time, in using this on a larger scale, lies in finding suitable search algorithms that will compare fingerprints which are discrete point sets, yet allowing for 'gaps'.

4.2. *Identification of functionally important regions in the protein*

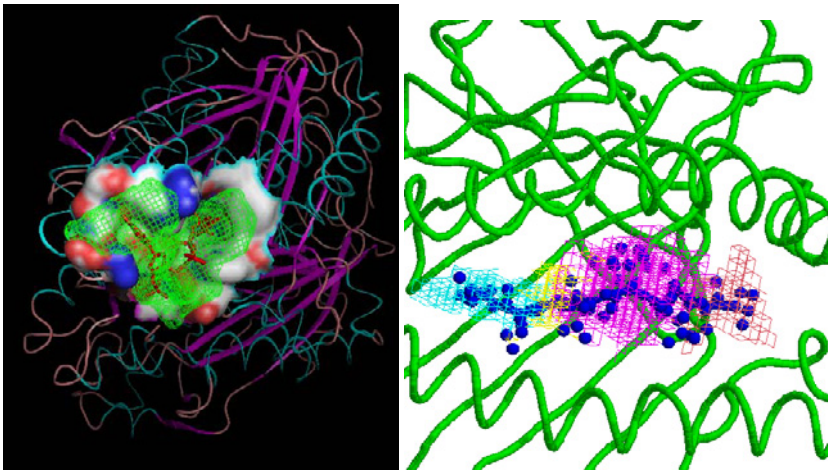
A different level of understanding protein function is to extract functionally important regions in them and compare and classify them with an aim of associating them with particular function(s). It has long been recognized that understanding ligand binding to a protein molecule holds the key to understanding function of the molecule. Even when protein structures are determined crystallographically as a complex with a ligand, a complete description of their binding sites is not always obtained because they may not be complexed with all the ligands required for the function of the molecule or because the complexed ligands are often substitutes of the natural ligands. A key step in the process of gaining functional insights from protein structures is therefore identification of all relevant binding sites in protein molecules. A further requirement for accurate identification of binding sites comes from the observation of moonlighting of protein molecules (Jeffery 1999; 2003), where many protein molecules have been found to have more than one function, quite often through different binding sites or even different binding modes at overlapping sites on the same protein. Even where crystal structures are available, they are rarely available as complexes with different ligands that may be required for moonlighting, hence making prediction by computational methods very important. A number of methods have emerged in the last decade for the task of locating binding sites in proteins (An *et al.*, 2005; Bhinge A 2004; Brady and Stouten 2000; Brylinski *et al.*, 2007; Chakrabarti and Lanczycki 2007; Coleman and Sharp 2006; Coleman *et al.*, 2006; Goodford 1985; Hendlich *et al.*, 1997; Huang and Schroeder 2006; Kalidas and Chandra 2008a; Peters *et al.*, 1996; Kleywegt and Jones 1994; Landon *et al.*, 2007; Levitt and Banaszak 1992; Liang *et al.*, 1998; Soga *et al.*, 2007; Tong *et al.*, 2008; Venkatachalam *et al.*, 2003; Glaser *et al.*, 2006). They can be broadly classified into (a) geometry based and (b) energy based methods. The geometry based methods are generally known to be faster while the energy based methods score better in terms of high accuracy of the sub-pockets predicted. Some examples of the geometry based methods are LigsiteCSC (Huang and Schroeder, 2006), CASTP (Liang *et al.*, 1998), PASS (Brady and Stouten, 2000), LigandFit (Venkatachalam *et al.*, 2003), VOIDOO (Kleywegt and Jones, 1994), APROPOS (Peters *et al.*, 1996), LIGSITE (Hendlich *et al.*, 1997), SURFNET(Glaser *et al.*, 2006), while examples of energy based methods are GRID (Goodford, 1985) Pocket finder (An *et al.*, 2005), Q-SiteFinder (Laurie and Jackson, 2005), desolvation based free-energy models (Coleman *et al.*, 2006) and solvent mapping models (Landon *et al.*, 2007). Roterman and co-workers have also reported identification of active sites based on the characteristics of the spatial distribution of hydrophobicity in a protein molecule, using a fuzzy-oil-drop model (Brylinski *et al.*, 2007). The different methods focus on different properties such as size, hydrophobicity, energy potential, solvent accessibility, desolvation energy or residue propensity for representing and hence analyzing the pockets. The chosen descriptor directly influences the quality of prediction. Hence it is important to explore use of different features to represent protein molecules and subsequently predict binding sites.

Recently, we developed a new geometry based method PocketDepth (Kalidas and Chandra 2008a; http://proline.physics.iisc.ernet.in/pocketdepth/) that divides a putative pocket into subspaces in a grid and computes their depths within the pocket, which is subsequently used to retain and cluster only the high-depth subspaces, thus utilizing the information of the neighbourhood of the relevant atoms in putative sites. By considering depth in the subspaces available in a pocket as defined and implemented in our method, we gain to understand how central and not merely how deep a given space is to a pocket. These centrally located cells with high-depth counts must be taken into account while designing ligands for a pocket. High prediction accuracies were obtained using this algorithm both in terms of the number of correct predictions as well as the extent of correctness of each prediction. An example of a prediction using PocketDepth is shown in Figure 3.

5. Site comparison

A given function could be conserved simply by having similarities in some elements of the structure, such as the binding site residues. A classic example is the large family of serine proteases which are classified into different sequence and structural families, but all come under the functional class of

Figure 3: Examples of binding site detection using PocketDepth in FabH from *Mycobacterium tuberculosis* and a human HLA allele, from their protein structures. The sites are shown as meshes. The ligand in the original crystal structures is also shown in both cases.



serine proteases due to the presence of the catalytic triad (Carter and Wells., 1987). Comparison of binding sites differ from comparison of whole structures for two main reasons (a) the binding sites are small containing only a few residues and (b) these residues need not be contiguous in sequence. Alignment of two sites containing discrete sets of atoms involves evaluation of a huge number of mappings. This makes it important to have efficient algorithms with low time and space complexities that are capable of identifying and ranking different extents of similarities appropriately. With several structural genomics projects as well as advances in computational methods for structure prediction, the structural databases are growing at a rapid pace, providing experimental structures of thousands and confident homology models of millions of protein molecules. The need for large scale comparisons of binding sites is hence accentuated. There are some intuitive tools already available for such a purpose. (An *et al.*, 2005, Bindowski *et al.*, 2004, Binkowski *et al.*, 2003; Bock *et al.*, 2007, Brakoulias and Jackson 2004, Bron and Kerbosch 1973, Campbell *et al.*, 2003, Chaumette 2004, Gold and Jackson 2006a;b, Kleywegt 1999, Kuhn *et al.*, 2007; Minai *et al.*, 2008, Morris *et al.*, 2005, Park and Kim 2006, Powers *et al.*, 2006, Stark and Russell 2003, Kalidas and Chandra, 2008b). SitesBase (Gold and Jackson 2006a;b) and PINTS (Stark and Russell 2003) use a method based on Geometric Hashing which involves selection of triads of points representing atomic types and positions in each site and comparing the triangles formed by triads. Graph based methods in which binding sites are represented as graphs identify maximal common sub-graphs between a pair of sites to find similarities between them. Alternately a depth first traversal strategy is adopted to find common set of nodes between a pair of graphs connected by similar pattern of edges. *Spherical Harmonic* (Morris et al., 2005) representation of a binding site captures distribution of points representing the site in terms of coefficients of characteristic frequencies. Very recently, we have developed a new algorithm called PocketMatch for representing a binding site in a frame invariant manner and comparison of pairs of sites based on alignment of sorted sequences of distances between pairs of points representing sites (Kalidas and Chandra 2008b). In this, each binding site is represented by 90 lists of sorted distances of pairs of atoms, flagged with residue type information, thus capturing both the shape and chemical nature of amino acid residues in the site. The sorted arrays are then aligned using a greedy incremental alignment strategy and scored to finally obtain PM*Score* values for each pair of sites. Perturbation analysis in which a portion of the points representing the sites were perturbed randomly both in their positions and in their chemical types, showed that chance similarities were virtually non-existent. An all versus all comparison of about 1000 binding sites in the PDBbind database using this algorithm also demonstrated that shape information alone is insufficient to discriminate between diverse binding sites, which however can be overcome by combining it with chemical nature of amino acids. An example of finding similarities in unrelated proteins is shown in Figure 4. ATP binding proteins (Stockwell

and Thornton, 2006) and sugar binding proteins (Ramachandraiah and Chandra, 2000) are some more examples of proteins containing different folds but have a common function in terms of the ligand they recognize due to similarities in their binding sites (Figure 5).

6. Protein–ligand interactions

One of the important aspects of present-day molecular biology is to gain an understanding of the molecular basis of the recognition phenomenon, so as to understand how proteins are capable of specific and reversible interactions with ligands. This can be achieved by studying the inter-relationship between protein structure, internal molecular dynamics, guided by its intra-molecular forces, influenced by other substances such as allosteric factors and function in terms of ligand binding, the inter-molecular forces involved, driven by the thermodynamic components. The measurement of thermodynamic parameters is important because all reversible biomolecular interactions involve a redistribution of non-covalent forces. The most experimentally accessible of the thermodynamic quantities occurring on a protein going from the free unbound state to the bound state is the uptake or release of heat or enthalpy. A wide variety of experimental methods are used for direct or indirect determination of thermodynamic quantities and hence the ligand binding strengths. These involve the calculation of thermodynamic quantities from theoretical relationships. For example, the enthalpy changes can be determined from the temperature dependence of the equilibrium binding or dissociation constant. High sensitivity calorimetric measurements on the other hand allow precise and direct determination of the change in enthalpy values. Computationally, the binding strengths can be measured by analysing their extents of interaction judged by their structures. Commonly used metrics such as interaction energies, buried surface area upon complexation, shape complementarity values (Cai *et al.*, 2002) or by simply analysing the number and nature of the hydrogen bonds involved in interaction.

6.1. *Deriving determinants of ligand recognition*

We often encounter examples of proteins of the same structural family but with subtle differences in their binding sites, recognizing completely different ligands and hence having completely different functions. Good examples in this category are the proteins adopting the TIM barrel fold. Proteins in this fold have diverse functions which include examples such as triose phosphate isomerase exhibiting an isomerase function, whereas tryptophan synthase that binds with a synthetase function. On the other hand, we also encounter examples of proteins from entirely different families, yet capable of binding to the same ligand and as consequence, have similar functions. The question that these factors bring about is what makes a protein capable of binding to the given ligand. Structural bioinformatics has been on several occasions used to derive critical determinants of recognition of a given ligand by its cognate proteins.

An example of such analysis shown in Figure 5 illustrates similarities in the sites and conformation of the phosphates of ATP in the P-loop containing proteins. Another example is the study of several carbohydrate binding proteins to identify common minimum principles required for the recognition of mannose, glucose and galactose, which indeed form much of the basis for recognition of higher sugars (Prabu *et al.*, 2006). In each of the three cases, proteins are indeed quite diverse, belonging to different structural and functional families and without any significant sequence similarities among them. Yet, they all share a common feature of the capability to recognize the same sugar- mannose, glucose or galactose. To understand the recognition principles in each case, the binding sites of each of the structures in each dataset were compared. Viewing them by aligning the individual sugars did not indicate presence of any patterns in the various sites. However, a search for similarities in the sites led to an alignment in which the ligands, along with their sites, were rotated about the axis of the ligand such that an O2 or an O3 hydroxyl could align with an O4 hydroxyl of the sugar from another protein. Analysis of the structures, overlaid in the re-oriented frameworks in all the three cases, lead to the derivation of common patterns in the occurrences of amino acid residues as well as in their relative spatial distributions within each data set (Figure 6). The study identifies an aspartic acid – O4 sugar hydroxyl interaction to be highly conserved, which appears to be crucial for recognition of all three sugars. Other interactions are specific to particular sugars, leading to individual fingerprints. Knowledge of these determinants will be useful in functional annotation of newer proteins as well as in lead design and protein engineering.

Another example is the identification of determinants of histamine recognition (Konkimalla and Chandra 2003). Towards understanding how histamine, a vital neurotransmitter, can perform multiple physiological tasks, an analysis of the different proteins that bind histamine was carried out. Their structural comparison reveals conformational rigidity of histamine. Yet, flexibility in the modes of histamine binding was observed, which appears to suit specific biological roles of the proteins. These results will be helpful in developing specific antihistamines.

Figure 4: Similarities detected in the binding sites of HIV Protease and Chaperon regulator protease containing Indinavir (PDB Ligand code - MK1) by using PocketMatch.

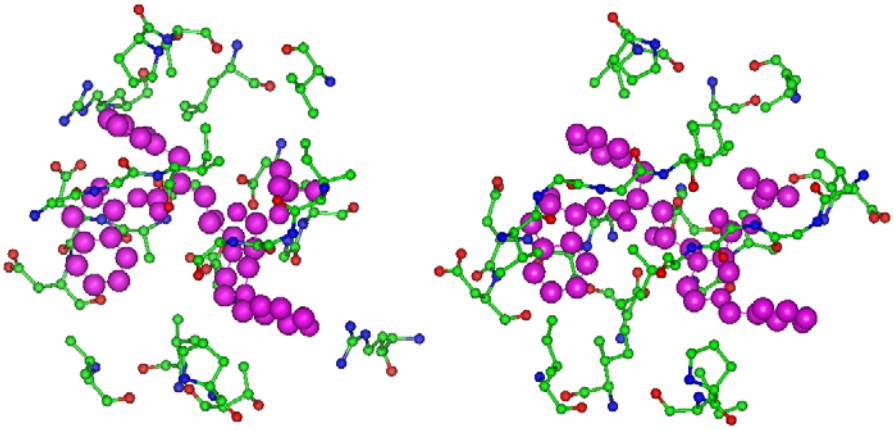
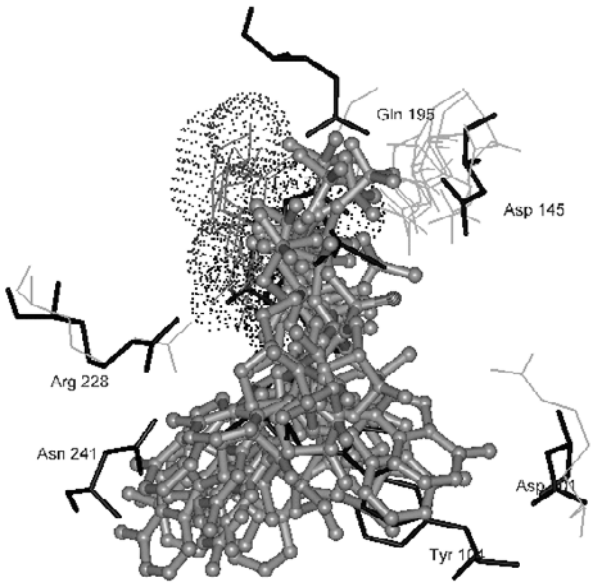


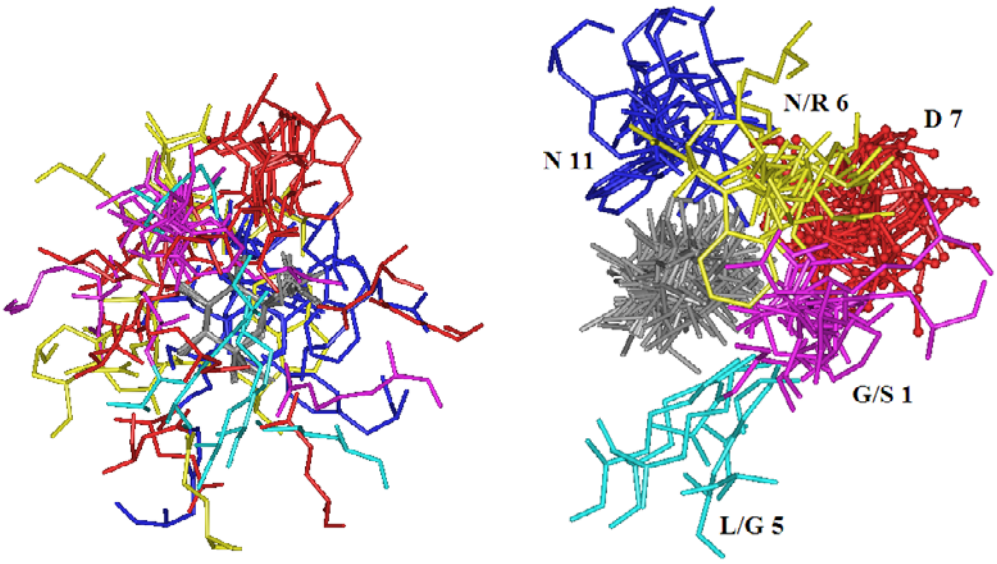
Figure 5: Superposition of ATP molecules in a variety of P-loop containing protein structures, indicating high conservation in their phosphates, but more flexibility in the base and sugar conformations. Conserved binding site residues in the proteins are indicated[42].



6.2. *Protein-Protein interactions*

It is well understood that proteins do not work in isolation and but require to interact with other proteins and sometimes nucleic acids to maintain normal physiology, since most biological processes are carried out by macromolecular assemblies and regulated through a complex network of protein- protein interactions. Interactions are of different types, the most important of them being complex formation leading to large protein-protein assemblies. An example of this category would be a ribosome or a RuvABC complex required for DNA recombination. Interactions can also be mediated through sugar molecules present as part of glycans that ride on proteins. Some interactions can also be in the form of influences where a given protein influences the function of another through increase or decrease in the levels of the associated metabolite, leading to feed-forward or feed-back regulations. Understanding protein-

Figure 6: A study showing glucose molecules in several different protein complexes after superposition of glucose molecules (left) and after finding binding site similarities (right). The figure on the right indicate that when re-orientation of sugar along with its site in one structure as compared to another is permitted, similarities in the sites can be discerned, addressing the issue of ligand presentation.



protein interactions would pertain mainly to the first category of interactions. A number of complexed structures are being determined experimentally and the current release of PDB contains several protein-protein complexes, providing a wealth of information on the nature of the interfaces and the types of interactions that stabilize protein-protein complexes. Experimental approaches studying protein-protein interactions have certain limitations and need to be complemented by computational methods. Different types of interaction prediction methods have emerged in the recent years, that involve one or more of the methods that involve consideration of gene neighbourhoods (Dandekar *et al.*, 1998) or phylogenetic profiles (Snel *et al.*, 2000) or detection of gene fusion (Enright and Ouzounis 2001) in another organism. These methods are all based on sequence information and provide quick information about possible protein-protein linkages. They do not however tell us if the two proteins can form a structural complex and where they do, there is no information on the mode of interaction or which segments of the two proteins may be involved. Structure based methods (Jones and Thornton 1997) are required to address these issues, which are becoming increasingly more feasible. Some of the recently developed algorithms are FTDOCK (Gabb *et al.*, 1997) which involves rigid-body docking on two biomolecules in order to predict their correct binding geometry. Protein-protein interfaces are generally larger, less conserved and often involve a fair amount of hydrophobic residues, making it difficult to detect as compared to that of protein-small molecular recognition. Some other methods depend upon identification and comparison of surface patches (Jones and Thornton, 1997) on protein surfaces, but methods in this category are in general still in their infancy with a lot of scope for improvement.

7. Structure based drug discovery

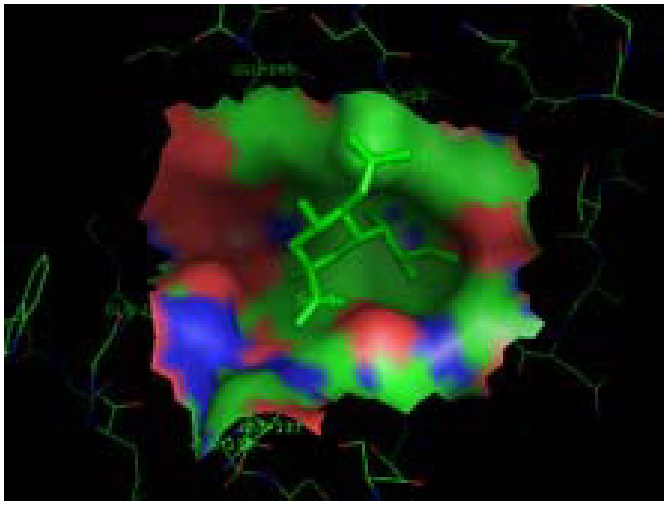
7.1. *Target identification*

Knowledge of the structure of the target macromolecule helps us in estimating the feasibility of the protein as a target and also facilitates computational docking of the ligand molecule into its binding site. Some examples of drugs designed by structure-based methods are Zanamivir and Oseltamivir (Alymova *et al.*, 2005) (against influenza neuraminidase), Nelfinavir, Amprenavir and Lopinavir (targeting HIV protease (Nair *et al.*, 2002)). Prior to docking, it is important to identify the binding site in the target protein, information for which is available many times through the structures of the complexes of the protein with its natural substrate. Chemical modification or site-directed mutagenesis data of the target protein can also provide clues about the binding site residues, where structures of complexes are not known.

7.2. *Lead identification and optimization*

A 'lead' can be viewed as a representative of a compound series with sufficient potential to

Figure 7: Docking of sialic acid to the binding site of influenza virus haemagglutinin using AutoDock. The binding site is shown as a surface and the docked pose of the ligand is shown in sticks.



progress into a full drug development programme. From the times of total reliance on the intuition of the medicinal chemist, the methods available for lead identification have come a long way. Computational methods used widely for this purpose cater to two broad scenarios, first, where the target macromolecular structure is known either through experiment or through prediction and second, where either such structural information about a target is not available or the target molecule is not even identified. For the first class, structure-based design algorithms exist, which utilise the detailed information about the precise geometry and chemistry that exist at binding sites of protein molecules, while for the second class, the methods are predominantly based on statistical correlations between the properties of a series of ligand molecules with a testable biological activity. Structure based drug design (SBDD) has emerged to be a well established field for designing appropriate these lead compounds or small molecules to enhance or inhibit the activity of the protein in question, when the structure of the target protein and the binding site details are known. Lead identification can be achieved in many ways. The most common and the most successful of them is to dock a given compound into an approximately defined binding site of the protein, which will establish if the ligand can bind to the protein in a geometrically and energetically feasible manner. When a library of compounds are available, virtual screening techniques can be employed, which range from gross one-dimensional property based screening of ligands to high resolution docking based screening. Very often a combination of physicochemical properties of the ligands are used in preliminary screening to select a small subset of the library, for which detailed docking is performed. Another method of identifying a lead compound is to build it up in the binding site of the protein from the basic building blocks or simple fragments. Each approach has its own strengths and limitations, but newer methods for more accurate lead identification are being developed to overcome the limitations.

7.2.1. *Docking*

Docking refers to the optimal positioning of a ligand molecule with respect to the binding site of a target structure. Many methods have been developed to perform ligand docking. The simplest is the rigid-body docking (Kuntz *et al.*, 1982), which represents internal volume of the ligand and void volume of the site by set of points and evaluates all superposable substructures between the two sets of points. Heuristic clique based searches (Ewing and Kuntz 1997; DOCK). Rarey and co-workers (1996) developed FlexX where the base fragment of the ligand is placed into the binding site considering complimentary interactions with atoms of site using geometric hashing followed by incremental addition of fragments to base fragment to arrive at the structure of the given ligand. Many possible energetically favourable conformations of the ligand are generated and later clustered by pose clustering based on root mean squared deviation (Linnainmaa *et al.*, 1988). Other methods available for this purpose are based on molecular dynamics simulations, stochastic search techniques such as simulated annealing and Monte Carlo simulations, and evolutionary algorithms (e.g. AUTODOCK (Morris *et al.*, 1999). An example of docking sialic acid to influenza virus haemagglutinin is shown in Figure 7. The strength of binding of the ligand to the target is usually determined by considering the intermolecular energies contributed by the interaction forces arising from electrostatic, hydrogen-bond, van der Waals and hydrophobic interactions (Muegge and Martin 1999; Sobolev *et al.*, 1999). The contribution of the solvent in ligand binding can also be explicitly considered. Quantum chemical models for evaluating interaction potential are also available (Xiang *et al.*, 2004; Zoete *et al.*, 2003). There are numerous examples in literature that report the use of docking in structure based lead identification. In some cases, they also provide a basis to rationalize relative affinities of a series of ligands, determined experimentally. Some examples of that from our work are designing appropriate small molecules to inhibit the activity of epidermal

growth factor receptor (Konkimalla *et al.*, 2007), rationalization of the binding affinities of a series of conformationally locked thiosugars as potent α-mannosidase inhibitors (Sivapriya *et al.*, 2007) and identifying molecular determinants of substrate and inhibitor specificity of polyphenol oxidase (Kanade *et al.*, 2007).

7.2.2. *Virtual high-throughput screening*

As the promise of structure-based drug design begins to be realised (Congreve *et al.*, 2005), the need for expanding to a larger scale is becoming more acute. A common need in present-day drug discovery therefore is to carry out a database search to find probable ligands, also referred to as 'virtual screening', so as to enrich biologically active compounds during 'lead' identification. A good example of this approach is the identification of the lead compounds to replace the anti-cancer drug Gleevec by overcoming the problem of drug resistance. The structure of the ABL tyrosine kinase, the target of Gleevec has been used to identify two promising lead compounds, which exhibited significant inhibitions in ABL tyrosine phosphorylation assays (Peng *et al.*, 2003).

On the computational front, development of high performance methods for computationally intense tasks such as docking, could lead to use of structure-based methods in virtual screening of millions of compounds for lead design. Towards this goal, AutoDock, a widely used, genetic algorithm based docking tool has been parallelized (Khodade 2007) enabling virtual screening, which would otherwise have been prohibitive on a routine scale due to the large computing times involved in docking.

7.2.3. *Ab-initio design*

Having to design a lead compound when just the structure of the protein molecule is available, would be analogous to having a lock and finding a key that fits into the keyhole. Ligand design however, is much more complex since both the ligand and the binding site can change their shapes to some extents upon binding. One way of finding the key would be to intuitively pick a few and test them or to screen against a set of keys, which are similar in concept to docking and virtual screening respectively. Another way of finding the key would be to simply take the components of the key and assemble it within the key hole of the lock so as to get the right fit, an approach that would be similar in concept to *ab-initio* design. This method has the greatest advantage of not being dependent on prior knowledge of a set of ligands that would bind to the protein and can be carried out with any protein whose structure is known and the binding site identified. A number of methods have been reported in literature, one of the most popular methods being LUDI (Bohm 1992), which uses a fragment-based approach. It suggests how suitable small fragments can be positioned into clefts of protein structures such as in an active site of an enzyme, in such a way that hydrogen bonds can be formed with the enzyme and hydrophobic pockets are filled with hydrophobic groups. The fragments are then scored in terms of their interaction energies using an empirical scoring function. Combinatorial chemistry has also been used to create a large library of structures with sufficient diversity, which is subsequently used for screening. *De novo* molecular design methods have been used to design new structures by sequentially adding molecular fragments to a growing structure, by adding functionality to an appropriately sized molecular scaffold, or by adding fragments building toward the center of a molecule starting from distant sites thought to interact with the target. While in principle, these approaches have the advantage of generating diverse molecular structures, in practice, only a few successes are reported, making *ab initio* design more a goal and not as yet a reality.

7.2.3.1. Guided ab-initio design. Although, there has been tremendous progress in the development of algorithms to address various aspects of structure-based drug design in the recent years (Anderson, 2003), there is still a need for improvement in methodology. For example, methods developed for fragment-based ligand design methods to bind at a given protein binding site (Honma, 2003), has enabled *ab-initio* ligand design, but do not have the intrinsic ability to pick out and suitably weight the crucial interactions. As a result, ligands designed with this approach do not always exhibit the intended pharmacological profiles. An effort made to address this issue, uses recognition fingerprints derived from a structural bioinformatics study, to carry out critical interaction guided fragment-based design of lead compounds targeted at the binding sites. An example with mannose, galactose and glucose binding proteins has been reported earlier (Prabu *et al.*, 2006). Results obtained from this approach have been shown to be superior to those from standard *ab-initio* design protocols. Guided design would be been useful in significantly improving the definition of the interaction search space and also in improving the probabilities of identifying more native-like ligands, which has implications for identifying leads with better affinities and specificities in a drug-design exercise. Identifying leads from a guided search would also be useful in a database search by narrowing down on the pharmacophore space as well as the interaction space to be searched.

7.3. *Pharmacodynamic profiling*

Most drugs in current clinical practice have been designed without the advantage of detailed knowledge of the interactions with various molecules in the cell. Hence, it is not surprising that almost every drug exhibits unwanted effects. The reasons for these adverse effects are not well understood in the majority of cases. Although, not well explored in the literature, structural bioinformatics has the potential to address several issues in understanding the mechanism of drug action and in designing improved drugs. An effort made in that direction pertains to the analysis of H2 antihistamines, which studies the cause for the paradoxical side effect they exert. Based on an understanding of histamine physiology, a systems landscape consisting of several proteins was identified that would be relevant to study the pharmacodynamics of anti-histamines. Docking and analysis of clinically used antihistamines into each of the components of the identified system, resulted in identifying histamine N-methyl transferase (HNMT) as a potential unintended target for H2-antihistamines. By unintentional inhibition of HNMT, a protein that removes excess levels of histamine by N-methylation, the drug leads to an accumulation of large levels of histamine, leading to the observed side effects (Figure 8: Vinod *et al.*, 2006). The study provides guidelines for the design of safer H2-antihistamines. The method also has the potential for application as a general strategy in understanding drug effects.

7.4. *Understanding drug resistance*

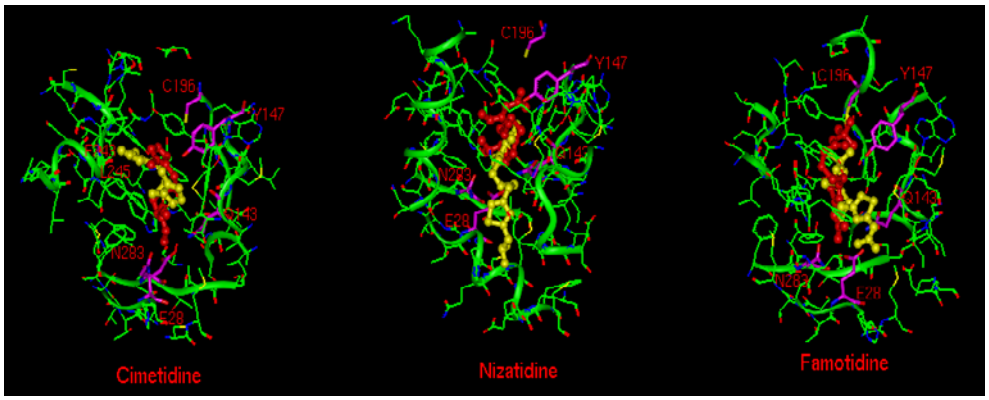
Resistance to drugs used clinically is a major problem that renders many drugs ineffective, and has been increasingly on the rise. One of the common ways by which resistance emerges is the mutation of the target at the binding site, reducing the affinity of the drug for the target. Understanding the structural changes that take place due to these mutations will help enormously in modifying the structure of drugs such that they overcome of the problem of that particular mode of drug resistance. Some examples in this direction are the computational study of resistance patterns of mutant HIV-1 aspartic proteases towards ritonavir and other antivirals and design of new derivatives to overcome that (Altman *et al.*, 2008, Nair *et al.*, 2002). Structural studies of HIV reverse transcriptase complexes with non-nucleoside inhibitors (Stammers, 2008) have contributed to the design of newer generation inhibitors and identified a number of features which may contribute to their much improved resistance profiles. However, at this stage, there are only a few successful examples in the literature, which utilize structural level studies to understand drug resistance. As more and more structural data become available, it can be expected that classification and hence prediction of mutations that can give rise to resistance to a class of drugs in particular protein families would become more common. In the future, it would also become possible to map the various single nucleotide polymorphisms (SNPs) in the human population in drug targets with their estimated binding affinities of the drugs binding to them, which would then provide a basis to understand differences in phenotypic response between different individuals for a given drug therapy.

8. Immunoinformatics and Structure based vaccine design

The huge diversity in the components that form our immune systems together with the complexity in their interactions and regulation make computational modelling and simulation very important. With advances in high-throughput technologies, experimental data of its various aspects are being gathered at high rates, leading to the genesis of the new discipline immunoinformatics (Flower, 2007). At present, it primarily refers to the management and analysis of immunological data. However its larger role would be in facilitating conversion of an immunological problem to a computationally tractable modelling problem, whose simulation and analysis yields biologically meaningful and interesting answers. Major developments in the area include development of several immunological databases (Lefranc *et al.*, 2008), analysis of the sequences of immunologically relevant molecules, modelling and analysing their structures, mathematical modelling of the immune systems and machine learning approaches to recognize patterns at various levels in the immunome (Petrovsky and Brusic, 2002). Immunoinformatics can be expected to aid in deciphering the immunome in the genome consisting of immunoglobulins, T-cell receptors and other relevant molecules, but more importantly will provide a framework to understand the ways by which the immune system to carry out is various tasks.

Structural level knowledge has been obtained for the antigen antibody interactions, and for many MHC or HLA molecules. Structural bases for antigen-antibody recognition has been discussed in detail in a previous issue (Bhowmick *et al.*, 2007), which also provides molecular insights into functional mimicry. Structural bioinformatics methods are still not commonly used in either antibody design or T-cell epitope design. Different

Figure 8: An example of docking of drugs to H2 antihistamines to Histamine N-methyl transferase, a study that explains their adverse effects.



approaches are being attempted though, that will lead to ready incorporation of structural knowledge in the various computational methods used for vaccine design.

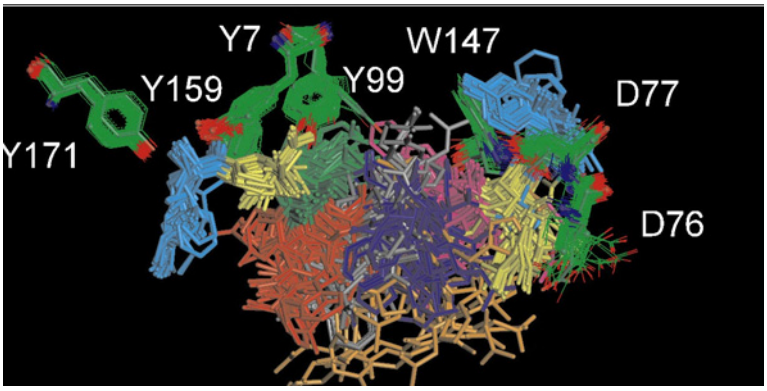
Drugs and Vaccines are entirely different in terms of their application, mode of working or the end uses. Yet, a structural view of the underlying molecular mechanisms provides a unifying theme for the design of drugs and vaccines. The term 'structure-based design' is now popular for drugs, but has not been widely used as yet for vaccines. It is however only logical for the vaccine design or indeed any molecular design to be based on three-dimensional atomic level structural information. Vaccines have conventionally been designed without the advantage of structural information, in many cases without even the advantage of any molecular level mechanistic information. The last few years are witnessing a paradigm shift in vaccine design, due to a surge in various 'omics' data as well as the development of many computational methods to analyse such data. Data mining from the genome sequences of hundreds of pathogenic microbes using novel algorithms are increasingly leading to derivation and annotation of the potential 'immunome' in these microbes (Korber *et al.*, 2006). Vaccines can be broadly classified into two categories- those that modulate the B-cell responses such as antibodies or antigens that generate antibodies in vivo and those that modulate T-cell responses, which include whole proteins, subunits, genes coding for specific subunits, specific peptides as well as non-peptide T-cell antigens.

8.1. *T-cell epitope identification*

T-cell responses are known to play a vital role in host immune responses upon exposure to pathogens such as *Mycobacterium tuberculosis* and the human immunodeficiency virus (HIV) in normal physiological conditions. One of the key requirements by an ideal vaccine candidate therefore is to trigger T-cell responses, so as to checkmate the pathogens. These are usually stimulated by short peptides commonly referred to as epitopes, which are derived from specific antigens from the pathogen and recognized by major histocompatibility complex (MHC) molecules or human leukocyte antigens (HLA), in humans. The non-self peptide- MHC complexes in turn recruit the T-cell receptors upon which further reactions are triggered. It is thus crucial to have a detailed understanding of the recognition of the epitopes by various MHC molecules. More than a hundred crystal structures for several MHC and HLA alleles covering all the three major loci have become available in the recent years. A structural bioinformatics study of peptide-HLA complexes to derive features that generate recognition specificity, useful for guiding the design process has been reported earlier (Dash *et al.*, 2007, Figure 9). Knowledge of the determinants of such recognition will also be useful in reverse engineering allele specific epitopes, which is likely to become an important process in rational vaccine design. Combination of statistical models and sequence derived patterns with structural patterns have been shown to be useful in identifying T-cell epitopes in different protein families from the genome of *Mycobacterium tuberculosis* (Vani *et al.*, 2006, 2007; Chaitra *et al.*, 2005, 2008) as well as the H5N1 influenza virus (Parida *et al.*, 2007).

In addition to B-cell antigen and T-cell epitope predictions, structural bioinformatics can also e expected to be of great value in understanding several fundamental issues relevant

Figure 9: Peptide binding space in the binding groove of HLA molecules. A structural bioinformatics study showing peptides in different HLA-peptide complexes, after structural superposition of the individual HLA molecules. The first residue (left most) and the ninth residue (right) of the nonameric peptides are shown in cyan; second and eighth in yellow; third in green; fourth in red; fifth in grey, sixth in blue; seventh in pink. Orange indicates those residues which in some cases present at the fifth and sixth positions of the peptides do not align with any of the nine positions, but appear more like insertions to the nonameric framework. HLA residues making conserved hydrogen bonds with the peptides in several alleles studied here are shown in atom colour and labelled (Dash *et al.*, 2007).



to rational vaccine design, such as recognition of peptidomimetics, proteasomal cleavage, trimming, transportation and presentation of peptides, self versus non-self discrimination and T-cell receptor recruitment. Structures of the protein molecules involved in many of these processes are already known, providing a framework to understand the basis for various molecular events and subsequently for higher confidence prediction of the antigens.

8.2. *Perpetuation of Immunological Memory*

Understanding the molecular mechanisms of immunological memory assumes importance in vaccine design. A mechanism for the maintenance of immunological memory through the operation of a network of idiotypic and anti-idiotypic antibodies (Ab2) has been proposed earlier by Nayak and co-workers (2001). Peptides derived from an internal image carrying anti-idiotypic antibody are hypothesized to facilitate the perpetuation of antigen specific T cell memory through similarity in peptide-MHC binding as that of the antigenic peptide. Using a structural bioinformatics study, the existence of such peptidomimics of the antigen in the Ab2 variable region and their similarity of MHC-I binding were identified (Gangadhar *et al.*, 2007). The analysis indicated that peptidomimics from Ab2 variable regions have structurally similar MHC-I binding patterns as compared to antigenic peptides, indicating a structural basis for memory perpetuation. Similar insights were obtained from the study of anti-idiotypic antibodies specific to rinderpest virus haemagglutinin (Vani *et al.*, 2007).

9. Future perspectives

With rapid accumulation of structural data and developments in the methods enabling their comprehension, structural bioinformatics is likely to make a significant impact across life sciences disciplines. Various application areas such as drug discovery and molecular design are likely to automatically benefit from this in a much more integral manner, than what we are witnessing today. Despite the advances seen in many aspects of this discipline, several questions still remain open, warranting further research in the area. One main requirement is the development of high performance algorithms and tools to speed up structural bioinformatics research and to increase the levels of sensitivity in recognizing various structural patterns that imply function. As seen in the trend of the periodical CASP experiments, advances in structure prediction are leading to the generation of protein structures with higher levels of confidence. With the development of better and more efficient structure comparison methods at fold and sub-structure levels, various structural motifs and sub-structures that relate to a particular function are likely to be identified, that would be used in a more routine manner in annotation of the function of a given protein. The coming years are likely to see significant advances in understanding and predicting protein-protein interactions, for which development of newer, efficient and more sensitive algorithms will be required.

The major advances in the area are likely to happen when the 'omics' level systems models

get integrated with detailed structural models of individual molecules. Comprehension of large volumes of complex information and its application in a higher-order understanding of the biological systems has necessitated the use of systematic mathematical analyses. When these whole systems can be understood at the level of the structures of the individual molecules and their inter-molecular interactions, whole new avenues will open up for modeling systems and simulating them, that will help us seek answers for a variety of questions. In drug discovery too, advances in this area will increasingly lead to the shift from ligand-driven statistical models that have been in vogue in the last decade or so, to the latest target-enriched structural and simulation models.

Acknowledgements

We are grateful to Prof. M. Vijayan for encouragement and useful discussions at various times. Financial support from DBT is gratefully acknowledged. Use of facilities at the Super Computer Education & Research Centre, Bioinformatics Centre, and Interactive Graphics facility supported by DBT is also acknowledged.

Received 10 July 2008; revised 8 September 2008.

References

1. M. D. Altman, E. A. Nalivaika, M. Prabu-Jeyabalan, C. A. Schiffer, and B. Tidor. Computational design and experimental study of tighter binding peptides to an inactivated mutant of HIV-1 protease. Proteins, 70(3):678–694, 2008.
2. I. Alymova, G. Taylor, and A. Portner. Neuraminidase inhibitors as antiviral agents. Curr Drug Targets Infect Disord, 5:401–9, 2005.
3. J. An, M. Totrov, and R. Abagyan. Pocketome via comprehensive identification and classification of ligand binding envelopes. Mol Cell Proteomics, 4(6):752–761, 2005.
4. P. Ananthalakshmi, K. Samayamohan, C. Chokalingam, C. Mayilarasi, and K. Sekar. Psst-2.0: Protein data bank sequence search tool. Applied Bioinformatics, 4:141–145, 2005.
5. A. C. Anderson. The process of structure-based drug design. Chemistry & Biology, 10:787–797, 2003.
6. A. Andreeva, D. Howorth, J. Chandonia, S. Brenner, T. Hubbard, C. Chothia, and A. Murzin. Data growth and its impact on the scop database: new developments. Nucleic Acids Research, 36:D419–D425, 2008.
7. C. B. Anfinsen, The Molecular Basis of Evolution John Wiley & Sons, Inc., New York, 1959.
8. H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne. The protein data bank. Nucleic Acids Research, 28:235–242, 2000.
9. B. E. Bernstein, D. M. Williams, J. C. Bressi, P. Kuhn, M. H. Gelb, G. M. Blackburn, and W. G. Hol. A bisubstrate analog induces unexpected conformational changes in phosphoglycerate kinase from trypanosoma brucei. J Mol Biol, 279(5):1137–1148, 1998.
10. M. M. S. Bharath, N. R. Chandra, and M. R. S. Rao. Molecular modeling of the chromatosome particle. Nucleic Acids Research, 31:4264–4274, 2003.
11. T. Bhat, V. Sasisekharan, and M. Vijayan. An analysis of side chain conformation in proteins. International Journal of Peptide and Protein Research, 13:170–84, 1979.
12. S. Bhattacharya, C. Bhattacharyya, and N. R. Chandra. Projections for fast protein structure retrieval. BMC Bioinformatics, 7 Suppl 5, 2006.
13. A. Bhinge, P. Chakrabarti, K. Uthanumallian, K. Bajaj, K. Chakraborty, and Varadarajan, R. Accurate detection of protein ligand binding sites using molecular dynamics simulations. Structure, 12, 2004.
14. A. Bhowmick, L. Krishnan, and D. Salunke. Structural immunology: Mechanisms of antigen recognition. J Ind Inst Sci, 87:35–41, 2007.
15. T. A. Bindowski, P. Freeman, and J. Liang. PvSoar: Detecting similar surface patterns of pocket and void surface of amino acid residues on proteins. Nucleic Acids Research, 32:555–558, 2004.
16. J. R. Bock and D. A. Gough. A new method to estimate ligand-receptor energetics. Mol Cell Proteomics, 1(11):904–910, 2002.
17. M. E. Bock, C. Garutti, and C. Guerra. Effective labeling of molecular surface points for cavity detection and location of putative binding sites. Comput Syst Bioinformatics Conf, 6:263–274, 2007.
18. H.-J. Böhm. On the use of LUDI to search the fine chemicals directory for ligands of proteins of known three-dimensional structure. Journal of Computer-Aided Molecular Design, 8:623–632, 1994.
19. P. E. Bourne and H. Weissig, Editors: L. Bordoli and T. Schwede Structural Bioinformatics Wiley InterScience, 2008
20. J. R. Bradford and D. R. Westhead. Improved prediction of protein-protein binding sites using a support vector machines approach. Bioinformatics, 21(8):1487–1494, 2005.
21. G. P. Brady and P. F. Stouten. Fast prediction and visualization of protein binding pockets with pass. J Comput Aided Mol Des, 14(4):383–401, 2000.
22. A. Brakoulias, R. M. Jackson. Towards a structural classification of phosphate binding sites in protein-nucleotide complexes: An automated all-against-all structural comparison using geometric matching. Proteins: Structure, Function and Bioinformatics, 56:250–260, 2004.
23. C. Bron and J. Kerbosch. Finding all cliques of an undirected graph [h]. Communications of the ACM, 16, 1973.
24. M. Brylinski, M. Kochanczyk, E. Broniatowska, and I. Roterman. Localization of ligand binding site in proteins identified in silico. J Mol Model, 13(6-7):665–675, 2007.
25. S. Burley. An overview of structural genomics. Nature Structure and Molecular Biology, 7:932–934, 2000.
26. H.-J. Bohm. The computer program LUDI: A new method for the de novo design of enzyme inhibitors. Journal of Computer-Aided Molecular Design, 6:61–78, 1992.
27. W. Cai, X. Shao, and B. Maigret. Protein-ligand recognition using spherical harmonic molecular surfaces: towards a fast and efficient filter for large virtual throughput screening. J Mol Graph Model, 20(4):313–328, 2002.
28. S. J. Campbell, N. D. Gold, R. M. Jackson, and D. R.Westhead. Ligand binding: functional site location, similarity and docking. Current Opinion in Structural Biology, 13:389–395, 2003.
29. G. L. Card, N. A. Peterson, C. A. Smith, B. Rupp, B. M. Schick, and E. N. Baker. The crystal structure of rv1347c, a putative antibiotic resistance protein from mycobacterium tuberculosis, reveals a gcn5-related fold and suggests an alternative function in siderophore biosynthesis. J Biol Chem, 280(14):13978–13986, 2005.
30. P. Carter and J. Wells. Engineering enzyme specificity by "substrate-assisted catalysis". Science, 237:394–399, 1987.
31. M. Chaitra, S. Hariharaputran, N. Chandra, M. Shaila, and R. Nayak. Defining putative t cell epitopes from pe and ppe families of proteins of mycobacterium tuberculosis with

vaccine potential. Vaccine, 23:1265–1272, 2005.
32. M. G. Chaitra, M. S. Shaila, N. R. Chandra, and R. Nayak. Hla-a*0201-restricted cytotoxic t-cell epitopes in three pe/ppe family proteins of mycobacterium tuberculosis. Scandinavian Journal of Immunology, 67:411–417, 2008.
33. S. Chakrabarti and C. J. Lanczycki. Analysis and prediction of functionally important sites in proteins. Protein Sci, 16(1):4–13, 2007.
34. N. Chandra, H. Muirhead, J. Holbrook, B. Bernstein, W. Hol, and R. Sessions. A general method of domain closure is applied to phosphoglycerate kinase and the result compared with the crystal structure of a closed conformation of the enzyme. Proteins, 30:372–80, 1998.
35. N. Chandra, N. Kumar, J. Jeyakani, D. Singh, S. Gowda, and M. Prathima. Lectindb: a plant lectin database. Glycobiology, 16:938–46, 2006.
36. F. Chaumette. Image moments: A general and useful set of features for visual servoing. IEEE Transactions on Robotics, 20, 2004.
37. H. Cheng, B. H. Kim, and N. V. Grishin. Malisam: a database of structurally analogous motifs in proteins. Nucleic Acids Res, 36(Database issue):211–217, 2008.
38. R. G. Coleman and K. A. Sharp. Travel depth, a new shape descriptor for macromolecules: application to ligand binding. Journal of Molecular Biology, 362:441– 458, 2006.
39. R. G. Coleman, A. C. Salzberg, and A. C. Cheng. Structure based identification of small molecule binding sites using free energy model. J. Chem. Inf. Model., 46: 2631–2637, 2006.
40. M. Congreve, C. Murray, and T. Blundell. Structural biology and drug discovery. Drug Discovery Today, 10:895–907, 2005.
41. T. Dandekar, B. Snel, M. Huynen, and P. Bork. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci, 23(9):324–328, 1998.
42. S. Datta, R. Krishna, N. Ganesh, N. R. Chandra, K. Muniyappa, and M. Vijayan. Crystal structures of mycobacterium smegmatis reca and its nucleotide complexes. Journal of Bacteriology, 185:4280–4284, 2003.
43. B. H. Dessailly, M. F. Lensink, C. A. Orengo, and S. J. Wodak. Ligasite–a database of biologically relevant binding sites in proteins with known apo-structures. Nucleic Acids Res, 36(Database issue):667–673, 2008.
44. P. Dönnes and A. Elofsson. Prediction of MHC class I binding peptides, using SVMHC. BMC Bioinformatics, 3:25–25, 2002.
45. R. L. Dunbrack and M. Karplus. Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. Nat Struct Biol, 1(5):334–340, 1994.
46. A. J. Enright and C. A. Ouzounis. Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. Genome Biol, 2(9), 2001.
47. T. J. A. Ewing and I. D. Kuntz. Critical evaluation of search algorithms for automated molecular docking and database screening. Journal of Computational Chemistry, 18:1175–1189, 1998.
48. P. Fariselli, F. Pazos, A. Valencia, and R. Casadio. Prediction of protein–protein interaction sites in hetero-complexes with neural networks. Eur J Biochem, 269(5): 1356–1361, 2002.
49. F. Ferrè, G. Ausiello, A. Zanzoni, and M. Helmer-Citterich. SURFACE: a database of protein surface regions for functional annotation. Nucleic Acids Res, 32:240-244, 2004.
50. D. R. Flower. Immunoinformatics and the *in silico* prediction of immunogenicity. An introduction. Methods Mol Biol, 409:1–15, 2007.
51. H. A. Gabb, R. M. Jackson, and M. J. Sternberg. Modelling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol, 272 (1):106–120, 1997.
52. M. Ganapathiraju, N. Balakrishnan, R. Reddy, and J. Klein-Seetharaman. Trans-membrane helix prediction using amino acid property features and latent semantic analysis. BMC Bioinformatics, 9 Suppl 1, 2008.
53. V. Gangadhar, J. Jeyakani, M. Shaila, R. Nayak, and N. Chandra. Perpetuation of immunological memory through common mhc-i binding modes of peptidomimic and antigenic peptides. Biochem Biophys Res Commun., 364:308–12, 2007.
54. F. Glaser, R. J. Morris, R. J. Najmanovich, R. A. Laskowski, and J. M. Thornton. A method for localizing ligand binding pockets in protein structures. Proteins, 62(2): 479–488, 2006.
55. N. D. Gold and R. M. Jackson. Fold independent structural comparisons of proteinligand binding sites for exploring functional relationships. Journal of Molecular Biology, 355:1112–1124, 2006a.
56. N. D. Gold and R. M. Jackson. Sitesbase: a database for structur-based protein-ligand binding site comparisons. Nucleic Acids Research, 34:D231–D234, 2006b.
57. P. J. Goodford. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J Med Chem, 28(7):849–857, 1985.
58. V. S. Gowri, S. B. Pandit, P. S. Karthik, N. Srinivasan, and S. Balaji. Integration of related sequences with protein three-dimensional structural families in an updated version of pali database. Nucleic Acids Res, 31:486–488, 2003.
59. A. Harrison, F. Pearl, I. Sillitoe, T. Slidel, R. Mott, J. Thornton and C. Orengo. Recognizing the fold of a protein structure. Bioinformatics, 19:1748-1759, 2003.
60. M. Hendlich, F. Rippmann, and G. Barnickel. Ligsite: automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph Model, 15(6): 359–363, 1997.
61. L. H. Holley and M. Karplus. Protein secondary structure prediction with a neural network. Proc Natl Acad Sci U S A, 86:152–156, 1989.
62. L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. J Mol Biol, 233(1):123–138, 1993.
63. L. Holm and C. Sander. Mapping the protein universe. Science, 273(5275):595–603, 1996.
64. T. Honma. Recent advances in de novo design strategy for practical lead identification. Med Res Rev, 23:606–32, 2003.
65. B. Huang and M. Schroeder. Ligsitecsc: predicting ligand binding sites using the connolly surface and degree of conservation. BMC Struct Biol, 6:19–19, 2006.
66. T. W. Huang, A. C. Tien, W. S. Huang, Y. C. Lee, C. L. Peng, H. H. Tseng, C. Y. Kao, and C. Y. Huang. Point: a database for the prediction of protein-protein interactions based on the orthologous interactome. Bioinformatics, 20(17):3273–3276, 2004.
67. C. J. Jeffery. Moonlighting proteins. Trends in Biosciences, 24, 1999.
68. C. J. Jeffery. Moonlighting proteins: old proteins learning new tricks. Trends in Genetics, 19:415–417, 2003.
69. D. T. Jones, W. R. Taylor, and J. M. Thornton. A new approach to protein fold recognition. Nature, 358(6381):86–89, 1992.
70. S. Jones and J. M. Thornton. Prediction of protein-protein interaction sites using patch analysis. J Mol Biol, 272(1):133–143, 1997.
71. Y. Kalidas and N. Chandra. PocketDepth: a new depth based algorithm for identification of ligand binding sites in proteins. Journal of structural biology, 161(1):31–42, 2008a. ISSN 1047-8477.
72. Y. Kalidas and N. Chandra. Pocketmatch: A new algorithm to compare binding sites in protein structures. Nature Precedings http://hdl.handle.net/10101/npre.2008.2142.1, 2008b.

73. S. R. Kanade, V. L. Suhas, N. Chandra, and L. R. Gowda. Functional interaction of diphenols with polyphenol oxidase. FEBS J., 274:4177–87, 2007.
74. L. A. Kelley, R. M. MacCallum, and M. J. Sternberg. Enhanced genome annotation using structural profiles in the program 3d-pssm. J Mol Biol, 299(2):499–520, 2000.
75. K. Gopalakrishnan, S.S.Sheik, C. Ranjani, A.Udayakumar, and K.Sekar. Conformation angles database (cadb-3.0). Protein and peptide Letters, 14:665–668, 2007.
76. P. Khodade, R. Prabhu, N. Chandra, S. Raha, and R. Govindarajan. Parallel implementation of Autodock. Journal of Applied Crystallography, 40:598–99, 2007.
77. G. J. Kleywegt. Recognition of spatial mofits in protein structures. Journal of Molecular Biology, 285:1887–1897, 1999.
78. G. J. Kleywegt and T. A. Jones. Detection, delineation, measurement and display of cavities in macromolecular structures. Acta Crystallogr D Biol Crystallogr, 50(Pt 2):178–185, 1994.
79. V. Konkimalla and N. Chandra. Determinants of histamine recognition: implications for the design of antihistamines. Biochem Biophys Res Commun, 309:425–431, 2003.
80. V. B. Konkimalla, V. L. Suhas, N. R. Chandra, E. Gebhart, and T. Efferth. Diagnosis and therapy of oral squamous cell carcinoma. Expert. Rev. Anticancer Ther., 7:317–29, 2007.
81. B. Korber, M. LaBute, and K. Yusim. Immunoinformatics comes of age. PLoS Comput Biol, 2:e71, 2006.
82. E. Krissinel and K. Henrick. Secondary-structure matching (ssm), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr, 60(Pt 12 Pt 1):2256–2268, 2004.
83. D. Kuhn, N. Weskamp, E. H¨ullermeier, and G. Klebe. Functional classification of protein kinase binding sites using cavbase. ChemMedChem, 2(10):1432–1447, 2007.
84. I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin. A geometric approach to macromolecule-ligand interactions. J Mol Biol, 161(2):269–288, 1982.
85. N. C. Kyrpides. Genomes online database (gold 1.0): a monitor of complete and ongoing genome projects world-wide. Bioinformatics, 15(9):773–774, 1999.
86. M. Landon, D. Lancia Jr., J. Yu, S. Thiel, and S. Vajda. Identification of hot spots within druggable binding regions by compuatational solvent mapping of proteins. J. Med. Chem., 50:1231–1240, 2007.
87. R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton. Procheck: a program to check the stereochemical quality of protein structures. Journal of Applied Crystallography, 26:283–291, 1993.
88. R. A. Laskowski, J. D. Watson, and J. M. Thornton. Protein function prediction using local 3d templates. J Mol Biol, 351(3):614–626, 2005a.
89. R. A. Laskowski, J. D. Watson, and J. M. Thornton. Profunc: a server for predicting protein function from 3d structure. Nucleic Acids Res, 33(Web Server issue):89–93, 2005b.
90. A. T. Laurie and R. M. Jackson. Q-sitefinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics, 21(9):1908–1916, 2005.
91. M. P. Lefranc, V. Giudicelli, L. Regnier, and P. Duroux. IMGT, a system and an ontology that bridge biological and computational spheres in bioinformatics. Brief Bioinform, 9(4):263–275, 2008.
92. D. G. Levitt and L. J. Banaszak. Pocket: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. J Mol Graph, 10 (4):229–234, 1992.
93. J. Liang, H. Edelsbrunner, and C.Woodward. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci, 7(9):1884–1897, 1998.
94. S. Linnainmaa, D. Harwood, and L. Davis. Pose determination of a three-dimensional object using triangle pairs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 10:634–647, 1988.
95. K. Liolios, K. Mavromatis, N. Tavernarakis, and N. C. Kyrpides. The genomes on line database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res, 36:D475–D479, 2008.
96. W. C. Lo, P. J. Huang, C. H. Chang, and P. C. Lyu. Protein structural similarity search by Ramachandran codes. BMC Bioinformatics, 8:307–307, 2007.
97. T. Madej, J. Gibrat, and S. Bryant. Threading a database of protein cores. Proteins Structure Function and Bioinformatics, 23:356–69, 1995.
98. L. J. McGuffin, K. Bryson, and D. T. Jones. The psipred protein structure prediction server. Bioinformatics, 16(4):404–405, 2000.
99. M. Meissner, O. Koch, G. Klebe, and G. Schneider. Prediction of turn types in protein structure by machine-learning classifiers. Proteins, 2008.
100. R. Minai, Y. Matsuo, H. Onuki, and H. Hirota. Method for comparing the structures of protein ligand-binding sites and application for predicting protein-drug interactions. Proteins: Structure, Function, and Bioinformatics, 72:367–81, 2008.
101. S. Mitra-Kaushik, M. Shaila, A. Karande, and Nayak. R. Idiotype and antigen-specific T-cell responses in mice on immunization with antigen, antibody, and anti-idiotypic antibody. Cell Immunol., 209:109–19, 2001.
102. G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. Journal of Computational Chemistry, 19 (14):1639–1662, 1999. ISSN 1096-987X.
103. R. J. Morris, R. J. Najmanovich, A. Kahraman, and J. M. Thornton. Real spherical harmonic expansion coefficients as 3d shape descriptors for protein binding pocket and ligand comparisons. Bioinformatics, 21:2347–2355, 2005.
104. J. Moult, J. T. Pedersen, R. Judson, and K. Fidelis. A large-scale experiment to assess protein structure prediction methods. Proteins, 23(3), 1995.
105. J. Moult, K. Fidelis, A. Kryshtafovych, B. Rost, T. Hubbard, and A. Tramontano. Critical assessment of methods of protein structure prediction-round vii. Proteins, 69 Suppl 8:3–9, 2007.
106. I. Muegge and Y. C. Martin. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. Journal of Medicinal Chemistry, 42: 791–804, 1999.
107. A. Murzin, S. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology, 247:536–540, 1995.
108. A. Nair, I. Bonin, A. Tossi, W. Wels, and S. Miertus. Computational studies of the resistance patterns of mutant hiv-1 aspartic proteases towards abt-538 (ritonavir) and design of new derivatives. J Mol Graph Model, 21:171–9, 2002.
109. S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol, 48(3):443–453, 1970.
110. M. Ondetti, B. Rubin, and D. Cushman. Design of specific inhibitors of angiotensinconverting enzyme: new class of orally active antihypertensive agents. Science, 196: 441–4, 1977.
111. C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. Cath–a hierarchic classification of protein domain structures. Structure, 5(8):1093–1108, 1997.
112. R. Parida, M. Shaila, S. Mukherjee, N. Chandra, and R.

Nayak. Computational analysis of proteome of h5n1 avian influenza virus to define t cell epitopes with vaccine potential. Vaccine, 25:7530–9, 2007.
113. K. Park and D. Kim. A method to detect important residues using protein binding site comparison. Genome Informatics, 17(2):216–225, 2006.
114. M. C. Peitsch. Large scale protein modelling and model repository. Proc Int Conf Intell Syst Mol Biol, 5:234–236, 1997.
115. H. Peng, N. Huang, J. Qi, P. Xie, C. Xu, J. Wang, and C. Yang. Identification of novel inhibitors of bcr-abl tyrosine kinase via virtual screening. Bioorganic and Medicinal Chemistry Letters, 13:3693–3699, 2003.
116. K. P. Peters, J. Fauck, and C. Frmmel. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. Journal of Molecular Biology, 256:210–213, 1996.
117. N. Petrovsky and V. Brusic. Computational immunology: The coming of age. Immunol Cell Biol, 80:248–54, 2002.
118. U. Pieper, N. Eswar, H. Braberg, M. S. Madhusudhan, F. P. Davis, A. C. Stuart, N. Mirkovic, A. Rossi, M. A. Marti-Renom, A. Fiser, B. Webb, D. Greenblatt, C. C. Huang, T. E. Ferrin, and A. Sali. Modbase, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res, 32(Database issue):217–222, 2004.
119. J. Pillardy, C. Czaplewski, A. Liwo, J. Lee, D. R. Ripoll, R. Kamierkiewicz, S. Oldziej, W. J. Wedemeyer, K. D. Gibson, Y. A. Arnautova, J. Saunders, Y. J. Ye, and H. A. Scheraga. Recent improvements in prediction of protein structure by global optimization of a potential energy function. Proc Natl Acad Sci U S A, 98(5):2329–2333, 2001.
120. A. Poddar, N. Chandra, M. Ganapathiraju, K. Sekar, J. Klein, Seetharaman, R. Reddy, and N. Balakrishnan. Evolutionary insights from suffix array-based genome sequence analysis. Journal of Bioscience, August Special Issue: 871–881, 2007.
121. R. Powers, J. C. Copeland, K. Germer, K. A. Mercier, V. Ramanathan, and P. Revesz. Comparison of protein active site structures for functional annotation of proteins and drug design. Proteins, 65(1):124–135, 2006.
122. R. Prabu, V. Rangannan and N. R. Chandra. Carbohydrate based drug design: Recognition fingerprints and their use in lead identification. Indian Journal of Chemistry: Special issue on Structure based drug design, 2006.
123. T. Prasad, M. Prathima, and N. Chandra. Detection of hydrogen-bond signature patterns in protein families. Bioinformatics, 19:167–168, 2003.
124. T. Prasad, T. Subramanian, S. Hariharaputran, H. Chaitra, and N, Chandra. Extracting hydrogen-bond signature patterns from protein structure data. Applied Bioinformatics, 3:125–135, 2004.
125. G. Pugalenthi, P. N. Suganthan, R. Sowdhamini, and S. Chakrabarti. SMotif: a server for structural motifs in proteins. Bioinformatics, 23(5):637–638, 2007.
126. G. Pugalenthi, P. N. Suganthan, R. Sowdhamini, and S. Chakrabarti. Megamotifbase: a database of structural motifs in protein families and superfamilies. Nucleic Acids Res, 36(Database issue):218–221, 2008.
127. Rajan Prabu, Vetriselvi Rangannan and Chandra, N.R. Carbohydrate based drug design: Recognition fingerprints and their use in lead identification. Indian Journal of Chemistry: Special issue on Structure based drug design, 2006.
128. G. Ramachandraiah and N. R. Chandra. Sequence and structural determinants of mannose recognition. Proteins, 39(4):358–364, 2000.
129. G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. J Mol Biol, 7:95–99, 1963.
130. K. Raman, Y. Kalidas, and N. Chandra, Editors : J. Chen and A.S. Sidhu. Biological database modeling Model driven drug discovery - principles and practices Artech House, 2007
131. M. Rarey, B. Kramer, T. Lengauer, and G. Klebe. A fast flexible docking method using an incremental construction algorithm. J Mol Biol, 261(3):470–489, 1996.
132. S. Raval, S. B. Gowda, D. D. Singh, and N. R. Chandra. A database analysis of jacalin-like lectins: sequence structure function relationships. Glycobiology, 14:1247–1263, 2004.
133. J. Ren, P. P. Chamberlain, A. Stamp, S. A. Short, K. L. Weaver, K. R. Romines, R. Hazen, A. Freeman, R. G. Ferris, C. W. Andrews, L. Boone, J. H. Chan, and D. K. Stammers. Structural basis for the improved drug resistance profile of new generation benzophenone non-nucleoside hiv-1 reverse transcriptase inhibitors. J Med Chem, 2008.
134. B. Rost, R. Schneider, and C. Sander. Protein fold recognition by prediction-based threading. J Mol Biol, 270(3):471–480, 1997.
135. R. S'anchez and A. Sali. Advances in comparative protein-structure modelling. Curr Opin Struct Biol, 7(2):206–214, 1997.
136. M. T. Shamim, M. Anwaruddin, and H. A. Nagarajaram. Support vector machine based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs. Bioinformatics, 23(24):3320–3327, 2007.
137. D. Shanmugham, P. N. Rangarajan, Nagasuma R. Chandra, B. K. Chandrasekhar Sagar, and G. Padmanaban. Delta-aminolevulinic acid dehydratase from *Plasmodium falciparum*. Journal of Biochemistry, 279:6934–6942, 2004.
138. S. Shi, Y. Zhong, I. Majumdar, S. Sri Krishna, and N. V. Grishin. Searching for three-dimensional secondary structural patterns in proteins with PRoSMoS. Bioinformatics, 23(11):1331–1338, 2007.
139. J.-M. Shin and D.-H. Cho. Pdb-ligand: a ligand database based on PDB for the automated and customized classification of ligand-binding structures. Nucleic Acids Research, 33:D238–D241, 2005.
140. I. N. Shindyalov and P. E. Bourne. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng, 11(9):739–747, 1998.
141. D. Singh, K. Saikrishnan, P. Kumar, Z. Dauter, K. Sekar, A. Surolia, and M. Vijayan. Purification, crystallization and preliminary x-ray structure analysis of the banana lectin from *Musa paradisiaca*. Acta Crystallogr D Biol Crystallogr, 60:2104–6, 2004.
142. K. Sivapriya, S. Hariharaputran, V. L. Suhas, N. Chandra, and S. Chandrasekaran. Conformationally locked thiosugars as potent alpha-mannosidase inhibitors: Synthesis, biochemical and docking studies. Bioorg. Med. Chem., 15:5659–65, 2007.
143. B. Snel, G. Lehmann, P. Bork, and M. A. Huynen. String: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res, 28 (18):3442–3444, 2000.
144. V. Sobolev, A. Sorokine, J. Prilusky, E. E. Abola, and M. Edelman. Automated analysis of interatomic contacts in proteins. Bioinformatics, 15:327–332, 1999.
145. S. Soga, H. Shirai, M. Kobori, and N. Hirayama. Use of amino acid composition to predict ligand-binding sites. J Chem Inf Model, 47(2):400–406, 2007.
146. R. Sowdhamini, D. F. Burke, J. F. Huang, K. Mizuguchi, H. A. Nagarajaram, N. Srinivasan, R. E. Steward, and T. L. Blundell. CAMPASS: a database of structurally aligned protein superfamilies. Structure, 6(9):1087–1094, 1998.
147. M. Stahl, C. Taroni, and G. Schneider. Mapping of protein surface cavities and prediction of enzyme class by a self-organizing neural network. Protein Engineering, 13: 83–88, 2000.
148. A. Stark and R. B. Russell. Annotation in three dimensions. PINTS: Patterns in nonhomologous tertiary structures. Nucleic Acids Research, 31:3341–3344, 2003.

149. G. R. Stockwell and J. M. Thornton. Conformational diversity of ligands bound to proteins. Journal of Molecular Biology, 356:928–44, 2006.
150. S. Sun. Reduced representation model of protein structure prediction: statistical potential and genetic algorithms. Protein Sci, 2(5):762–785, 1993.
151. J. D. Szustakowski and Z. Weng. Protein structure alignment using a genetic algorithm. Proteins, 38(4):428–440, 2000.
152. Y. Tanrikulu and G. Schneider. Pseudoreceptor models in drug design: bridging ligand and receptor-based virtual screening. Nature Reviews Drug Discovery, 7:667–677, 2008.
153. W. Taylor and C. Orengo. Protein structure alignment. Journal of Molecular Biology, 208:1–22, 1989.
154. W. Tong, R. J. Williams, Y. Wei, L. F. Murga, J. Ko, and M. J. Ondrechen. Enhanced performance in prediction of protein active sites with thematics and support vector machines. Protein Sci, 17(2):333–341, 2008.
155. R. Unger. The genetic algorithm approach to protein structure prediction. Structure and Bonding, 110:153–175, 2004.
156. J. Vani, M. Shaila, N. Chandra, and N. R. A combined immuno-informatics and structure-based modeling approach for prediction of t cell epitopes of secretory proteins of mycobacterium tuberculosis. Microbes Infect, 8:738–46, 2006.
157. J. Vani, J. Jeyakani, N. R. Chandra, R. Nayak, and M. S. Shaila. Peptidomimetics of antigen are present in variable region of heavy and light chains of anti-idiotypic antibody and function as surrogate antigens for perpetuation of immunological memory. Mol. Immunol., 44:3345–3354, 2007.
158. C. M. Venkatachalam, X. Jiang, T. Oldfield, and M. Waldman. Ligandfit: a novel method for the shape-directed rapid docking of ligands to protein active sites. J Mol Graph Model, 21(4):289–307, 2003.
159. P. Vinod, B. Konkimalla, and N. Chandra. *In-silico* pharmacodynamics: correlation of adverse effects of h2-antihistamines with histamine n-methyl transferase binding potential. Appl Bioinformatics, 5:141–50, 2006.
160. H. A. Watkins and E. N. Baker. Structural and functional analysis of Rv3214 from *Mycobacterium tuberculosis*, a protein with conflicting functional annotations, leads to its characterization as a phosphatase. Journal of Bacteriology, 188:3589–3599, 2006.
161. Y. Xiang, D. W. Zhang, and J. Z. H. Zhang. Fully quantum mechanical energy optimization for protein-ligand structure. Journal of Computational Chemistry, 25:1431–1437, 2004.
162. Y. Ye and A. Godzik. Flexible structure alignment by chaining aligned fragment pairs allowing twists. Bioinformatics, 19 Suppl 2:246–255, 2003.
163. J. Zhu and Z. Weng. Fast: a novel protein structure alignment algorithm. Proteins, 58 (3):618–627, 2005.
164. V. Zoete, O. Michelin, and M. Karplus. Protein-ligand binding free energy estimation using molecular mechanics and continuum electrostatics. Application to HIV-1 protease inhibitors. Journal of Computer-Aided Molecular Design, 17:861–880, 2003.

Yeturu Kalidas is a PhD student at the Indian Institute of Science. He has B.Tech in Computer Science and is currently working in the area of systems level integration of structural information about protein ligand interactions.

Nagasuma Chandra holds bachelors and masters degrees in pharmaceutical sciences and pharmacognosy, both from Bangalore and a PhD in structural biology from the University of Bristol, UK. Following postdoctoral work at Molecular Biophysics Unit, IISc, and a research scientistship awarded by the Poornaprajna research institute, she now serves on the faculty of Bioinformatics at the Indian Institute of Science, Bangalore. Her current research interests are in computational systems biology, cell modelling and structural bioinformatics and in applying these to address fundamental issues in drug discovery and in rational vaccine design.