Interdiscip Sci Comput Life Sci (2010) 2: 347–366
DOI: 10.1007/s12539-010-0045-6

Structural Bioinformatics: Deriving Biological Insights from Protein Structures

Nagasuma CHANDRA*, Praveen ANAND, Kalidas YETURU
(Bioinformatics Centre, Indian Institute of Science, Bangalore 560012, India)

Received 23 April 2010 / Revised 18 June 2010 / Accepted 21 June 2010

Abstract: Structural bioinformatics can be described as an approach that will help decipher biological insights from protein structures. As an important component of structural biology, this area promises to provide a high resolution understanding of biology by assisting comprehension and interpretation of a large amount of structural data. Biological function of protein molecules can be inferred from their three-dimensional structures by comparing structures, classifying them and transferring function from a related protein or family. It is well known now that the structure space of protein molecules is more conserved than the sequence space, making it important to seek functional associations at the structural level. An added advantage of structural bioinformatics over simpler sequence-based methods is that the former also provides ultimate insights into the mechanisms by which various biological events take place. A bird's eye-view of the different aspects of structural bioinformatics is given here along with various recent advances in the area including how knowledge obtained from structural bioinformatics can be applied in drug discovery.
Key words: protein structures, structural genomics, structure-function relationship, structural analysis, biological data mining.

1 Introduction

Deciphering complete genome sequences of several organisms including that of the human genome, has been marking defining moment in the history of biology (Fleischmann *et al.*, 1995; Forster and Church, 2006; Venter *et al.*, 2001). With the architectural blue-print of life of several different organisms in hand, the next step is to comprehend the huge pool of data (Kyrpides, 1999; Liolios *et al.*, 2008), identify and understand the function of the individual gene products.

In biology, knowledge available for one system heavily influences understanding of a related system. It is quite understandable therefore, why recognizing similarities and deriving relationships are crucial for all further knowledge, making bioinformatics an integral and important component of modern biology. This need is not only heightened, but is also rendered with the large number of genomes sequenced in the last few years. Where available, protein structures provide much better functional insight than their sequences alone. The reasons are that: as compared to the sequences, two-fold structures provide (a) a much higher resolution of information about the protein molecules and (b) a much more sensitive approach for detecting similarities among proteins. This is because protein structures are seen to cluster only into certain regions of the entire fold space suggesting that the same fold is repeatedly sampled in nature (Holm and Sander, 1996; Russell *et al.*, 1997).

The need to navigate and comprehend this large resource of experimental and theoretical structural data, has automatically led to genesis of a new discipline called structural bioinformatics (Burley, 2000; Bourne and Weissig, 2008), which has become well established in the last decade. Structural Bioinformatics is probably the 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. This ultimately helps us to understand 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 they help associate a molecule with a function, and also provide ultimate insights into the mechanisms by which various biological events take place.

In principle, the term 'structural bioinformatics' could encompass all biological macromolecules, but is used here predominantly in the context of protein

*Corresponding author.
E-mail: nchandra@serc.iisc.ernet.in

molecules, given the focus of this review. Comparing proteins, deriving structural patterns, correlating with function and ultimately utilizing such patterns of prediction are all integral components of structural bioinformatics. Given the complexities involved in solving new X-Ray or NMR structures of protein molecules, structure determination might often feel like a successful end to a long effort, but in reality a structure of a protein molecule is just the beginning of a journey to understand the function of protein molecule. Structural bioinformatics is an important area that serves as a bridge in transforming protein structures into biological insights.

2 Generation of structural data

Protein structural data is growing rapidly, with the current holdings going beyond 64,000 entries in the Protein Data Bank (Berman *et al.*, 2000), as illustrated in Fig. 1. Various structural genomics projects are also underway to obtain structural data on a genome-wide scale (Lesley *et al.*, 2002; Goulding *et al.*, 2003; Marsden *et al.*, 2007). Although the focus of this work is structural bioinformatics, a brief overview of structure determination, highlighting the importance of experimental structure determination is included here. Structural bioinformatics takes these structures as input and provides that crucial link between structure and function.

2.1 X-ray crystallography

Protein crystallography, which is essentially a form of very high-resolution microscopy, facilitates visualization of protein structures at the atomic level (Fig. 2(a)). This technique is now used routinely to determine the structures of protein molecules. It is also used commonly to understand how natural ligands, inhibitors and drugs bind to different proteins, as well as to derive guidelines for designing novel drugs or rationally engineering enzymes with enhanced capabilities (Chen, 2001). The main requirement for employing this method is to obtain diffraction quality single crystals from the pure form of the protein sample. Following data collection and processing, structures are solved using an appropriate method such as multiple isomorphous replacement or molecular replacement (Ilari and Savino, 2008). The structures are then refined, checked and analyzed.

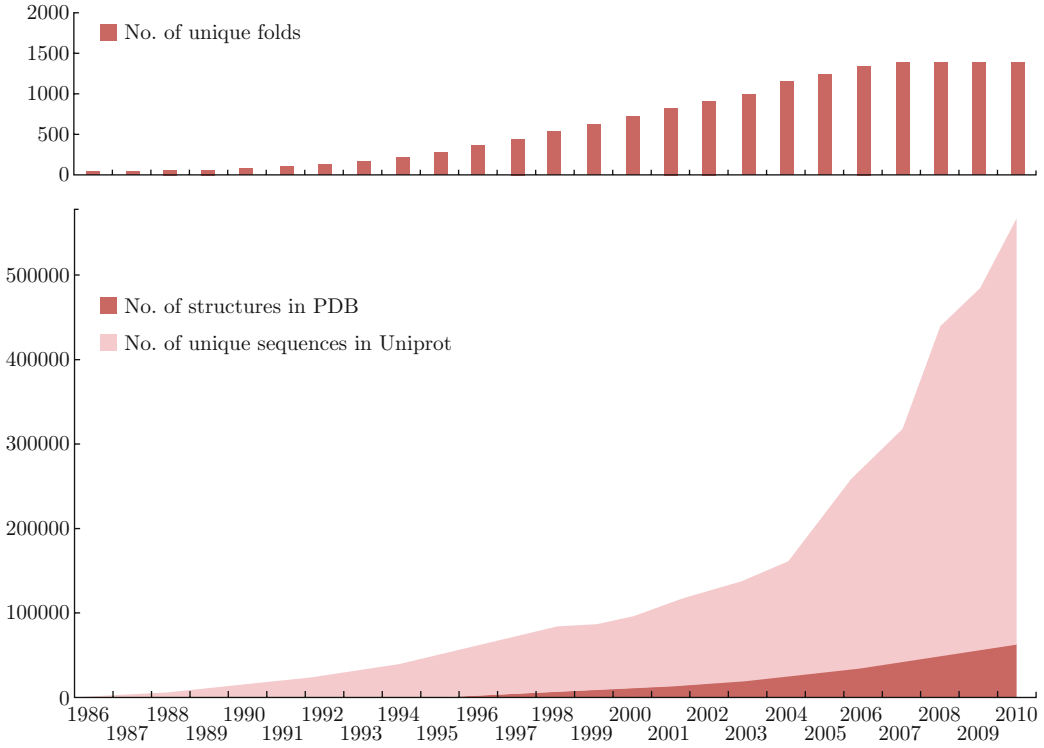


Fig. 1 Growth of sequence and structural data in the past two decades. About 580,000 unique protein sequences are available in Uniprot (shown in the lighter shade), but the numbers of structures in PDB are only about 64000 (shown in dark shade). The huge gap between sequence and structure is evident from this graph. The bottom panel shows the number of sequences and structures in the databases, whereas the top panel indicates the number of unique folds (filled bars) that the structures in PDB belong to. The numbers of folds in the recent years have remained about the same, which is illustrative of the fact that very few novel folds are now being found

2.2 Nuclear magnetic resonance

The nearest competing method for solving protein structures is nuclear magnetic resonance (NMR) spectroscopy (Fig. 2(b)), which has produced more than 7800 structures, so far as seen in PDB (Berman *et al.*, 2000). Purified protein is taken in a solution form, resonances are assigned, restraints are generated and a structure is calculated and validated, an example of a structure is shown in Fig. 2 (Wuthrich, 2003; McDermott, 2004; Baldus, 2006; Hong, 2006). Difficulties arise, typically at the resonance and Nuclear Overhauser Effect (NOE) assignment steps, hence leading to the development of a number of methods to simplify the task (Tzakos *et al.*, 2006). NMR spectroscopy has been useful in solving the structures of proteins (e.g., membrane proteins) that may not be readily amenable for investigation through crystallography. A limitation of this technique is that it can only be used easily for small proteins.

2.3 Electron microscopy

To understand biological phenomenon at high resolution, it is important to progress from studying individual domains or proteins towards studying multidomain proteins and larger assemblies. Advances in electron microscopic technologies have enabled the visualization of the structure and dynamics of a range of biological assemblies at resolution varying from 2–3nm to 0.3nm (Chiu *et al.*, 2005; Jiang and Ludtke, 2005; Lucic *et al.*, 2005; Frey *et al.*, 2006; Renault *et al.*, 2006). Biological samples are usually studied at cryo-temperatures to reduce the thermal fluctuations. This technique can help in studying large macromolecular complexes, larger assemblies (Fig. 2(c)), providing a comprehensive picture including cellular localization and interaction, thus tending towards merging cell biology and molecular biophysical processes.

2.4 Homology modeling

Millions of gene sequences translated quite confidently into their corresponding protein sequences are now available. Determining three-dimensional structural data on the other hand is much harder, requiring large quantities of purified protein in hand, besides being amenable to individual structure determination methods. To bridge the wide gap between sequence and structure, various computational methods that can predict the structure of a protein molecule with high confidence in many cases have emerged (Sun, 1993; Sánchez and Sali, 1997; Pillardy *et al.*, 2001; Unger, 2004; Jones, 2005). Of these, homology modeling 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 (Sánchez and Sali, 1997). This is based on the premise that two sequences that are homologous also share the same structural fold. Energy minimization 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 neighborhoods, force-field parameters, and model refinement techniques (Sánchez 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 built models built (Moult *et al.*, 1995; Moult *et al.*, 2007).

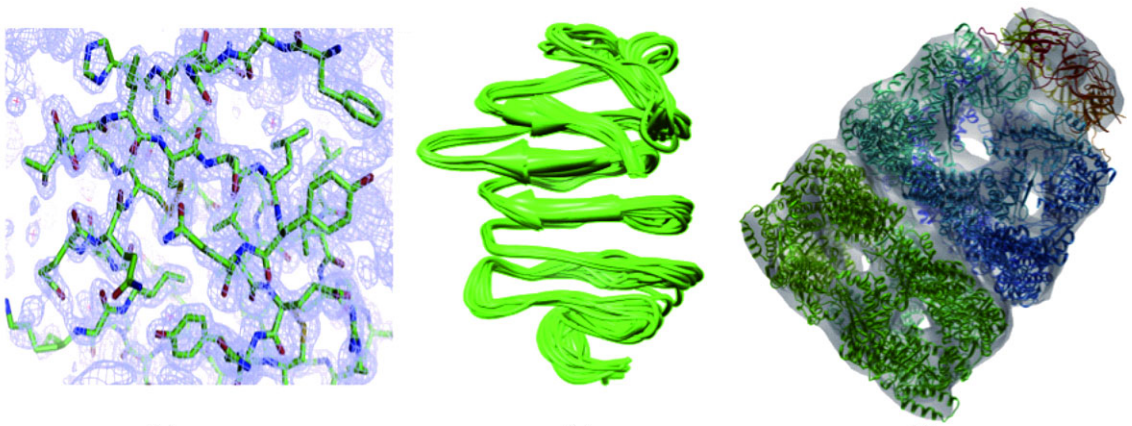


Fig. 2 An illustration to indicate three methods of structure determination. The crystal structure of a part of human insulin (3E7Y) along with its electron density is shown in (a). Spruce bud form antifreeze protein structure ensemble consisting of 20 models (1EWW) determined by solution NMR is shown in (b), where ascryo electron microscopy structure of GROES-ADP7-GROEL-ATP7 complex along with its electron density map is shown in (c)

2.5 Structure verification

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 allowed 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 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, are used quite routinely in crystal structure refinement and for quality estimation (Laskowski *et al.*, 1993). Table 1 lists some of the commonly used databases as well as web-servers hosting software tools for structural bioinformatics.

Numerous examples of molecular models of proteins can be seen in the literature, where they have been used for obtaining a variety of biological insights (Jackson,

Table 1 A list of important resources for structural bioinformatics. The URLs of the various web-servers hosting the databases and the software tools, along with their associated publications are also shown

Description	*URL*	*Reference*
Protein Data Bank (PDB) Repository containing all the 3D structures of the biological molecules	http://www.pdb.org	(Berman *et al.*, 2000)
The Macromolecular Structure Database (MSD) Also known as PDBe, is a European project for collection, management and distribution of data regarding biological macromolecules	http://www.ebi.ac.uk/msd/	(Tagari *et al.*, 2006)
Fold Classification based on structure-structure assignments (FSSP) Families of Structurally Similar Proteins superimposed and generated using DALI algorithm.	http://ekhidna.biocenter.helsinki.fi/dali_server/	(Holm and Sander, 1998)
MSDChem: Ligand Chemistry Provides access to ligands and small molecule dictionary, this repository defines the link between proteins and chemistry.	http://www.ebi.ac.uk/msd-srv/chempdb/	(Dimitropoulos *et al.*, 2006)
Structural classification of proteins (SCOP) Database containing the protein structural domains largely classified manually depending upon similarities of sequences and 3D structures.	http://scop.mrc-lmb.cam.ac.uk/scop/	(Murzin *et al.*, 1995)
Class architecture topology and hierarchical classification of proteins (CATH) Semiautomatic hierarchical classification of protein domains	http://www.cathdb.info	(Orengo *et al.*, 1997)
Protein Function Prediction ProFunc Web server for predicting the likely function of proteins whose 3D structures are known.	http://www.ebi.ac.uk/thornton-srv/databases/ProFunc/	(Laskowski *et al.*, 2005)
PDB-Ligand Database of small molecular ligands that are bound to macromolecular structures in PDB.	http://www.idrtech.com/PDB-Ligand/	(Shin and Cho, 2005)
PubChem Database of chemical molecules maintained by NCBI containing information about bioassay, bioactivity and results from high-throughput screening as well.	http://pubchem.ncbi.nlm.nih.gov/	(Wang *et al.*, 2009)
ChemBank Public web-based informatics environment created by Broads Institute storing informations of small molecues and biomedically relevant assays.	http://chembank.broad.harvard.edu/	(Seiler *et al.*, 2008)
LigASite Database of biologically relevant binding sites	http://www.bigre.ulb.ac.be/Users/benoit/LigASite/	(Dessailly *et al.*, 2008)
Structural motif databases (MALISAM) A database of structurally analogous motifs in proteins.	http://prodata.swmed.edu/malisam/	(Cheng *et al.*, 2008)
PINTS Database for detection of local structural patterns in proteins.	http://www.russell.embl.de/pints/	(Stark and Russell, 2003)
MegaMotifBase Database of structural motifs in protein families and superfamilies.	http://caps.ncbs.res.in/MegaMotifbase/index.html	(Pugalenthi *et al.*, 2008)
SURFACE Database of annotated and compared protein surface regions.	http://cbm.bio.uniroma2.it/surface/	(Ferrè *et al.*, 2004)

1991; Wallace, 1993; Lee *et al.*, 2007). An example of the use of this technique, is the molecular modelling of the closed conformation of a ternary complex of phosphoglycerate kinase. (Chandra *et al.*, 1998), 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 led to an 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 (Bharath *et al.*, 2003). 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 optimization stage (Tanrikulu and Schneider, 2008). An early notable example 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.*, 2008). Models of a number of G-protein coupled receptors, ion channel and voltage gated channel proteins have been built and utilized for guiding lead identification and rational design of new lead compounds (Hillisch *et al.*, 2004).

2.6 Fold recognition

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 are very low and thus such relationships are not obvious. It is now well accepted that conservation at the structure level is higher and thus more detectable than at the sequence level (e.g., 1B3A and 1TVX have low sequence identity but high structural similarity (Lo *et al.*, 2007). In a different context, this issue can be 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 recognized. Nevertheless, many more structures can be predicted by recognizing with which of the known folds a given sequence is most compatible. One of the first methods reported for this purpose, is popularly known as 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 measurement of all pair-wise interaction potentials of the individual residues (Jones *et al.*, 1992; Rost *et al.*, 1997). Profile-based methods have also been commonly used, which 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 (Bowie *et al.*, 1991). 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. Although structure prediction by threading is conceptually very appealing and works well in a number of cases, it has also witnessed failures in some other cases (Moult *et al.*, 2007). With development of newer methods to overcome the existing limitations, structure prediction by recognizing the appropriate template can be envisaged to be utilized more extensively. Another category of modeling 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, three dimensions as well as the difficulty in discriminating real structures from the decoys, and thus it cannot as yet be used as a routine technique (Moult, 2005; Cozzetto *et al.*, 2009).

3 Molecular visualization and structural analysis

Visualization of protein structures has undergone tremendous transformation, starting from brass models of the first crystal structures (Kendrew *et al.*, 1958) followed by physical wireframe models, the first cartoon representations on the computer, to 3D interactive graphics and the more emerging virtual reality representations (Richardson and Richardson, 1992). A variety of programs are available for visualizing of proteins, which are used routinely, some examples being Rasmol, Pymol, DeepView, Mage, VMD, UCSF-Chimera and the web-based tools such as Jmol, Chime and WebMol.

Computer graphics tools have enabled routine use of wireframe, space filling, ball and stick, cartoon, ribbon, and surface representations (Fig. 3). Visualization of molecules in the chosen representations is often critical for obtaining insights into various aspects on the motifs present, how the molecule compares in the fold or in the selected regions with other proteins an investigator may be interested in, often also in appreciating conformational changes in the protein in its different states, thus forming an important step in understanding function. Various properties such as the electrostatic potential, solvent accessible surface area or volume can

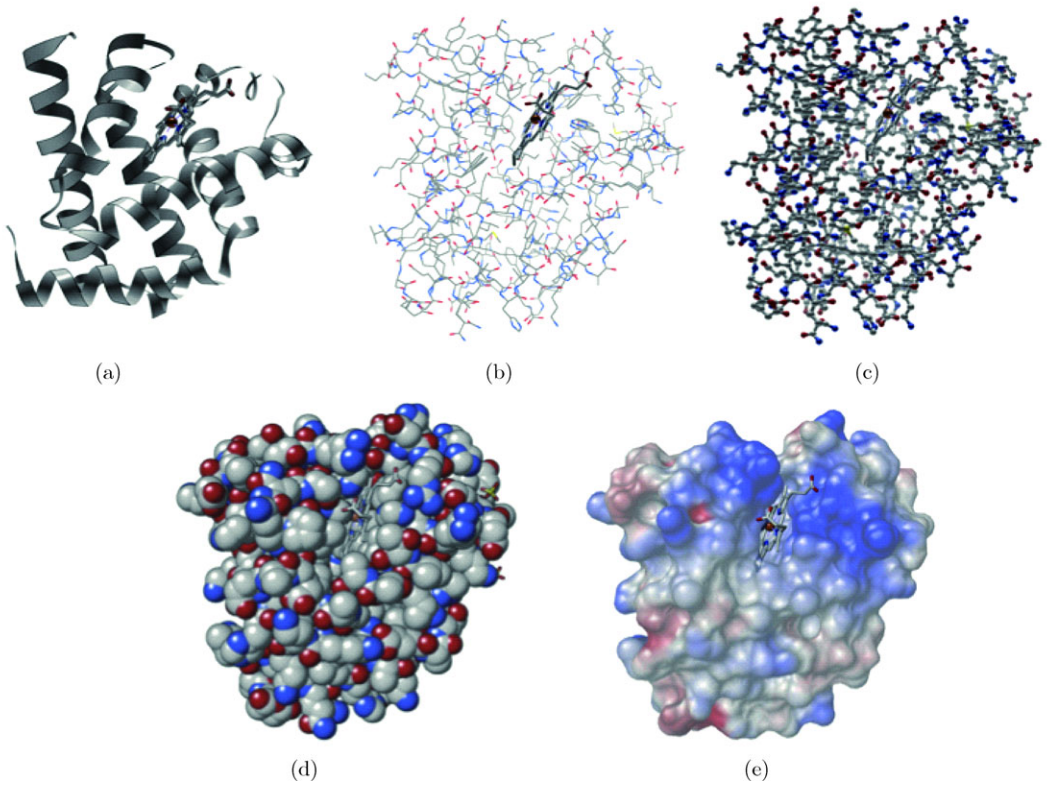


Fig. 3 Different modes of visualization of protein structures. The protein shown here is sperm whale myoglobin (1A6N). The structures can be rendered in a variety of ways, ranging from cartoon representation as shown in (a), wireframes in (b), ball and stick representation in (c), CPK in (d) and columbic surface representation showing the charge distribution on the surface of the protein in (e). The protophyrin IX-containing Fe atom bound to protein is shown as sticks in all types of rendering

be computed for the whole protein or the individual amino acid residues that can also be visualized. Development of new methods and new algorithms for molecular visualization is an active area of research, which can be expected to result in new ways of representing different molecular properties, real time visualization of results of complex computations and highly interactive graphical systems (Richardson and Richardson, 1992; O'Donoghue *et al.*, 2010).

4 Structural comparison of protein structures and 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 sub-graph detection or local geometry matching. For example, DALI (Holm and Sander, 1993) uses distance matrices, CE (Shindyalov and Bourne, 1998) uses combinatorial extension of alignment path, the method by Taylor and Orengo (1989) uses Taylor and Orengouses dynamic programming, that by Szustakowski and Weng (2000) uses genetic algorithms, that by Zhu and Weng (2005) uses maximal common sub-graphs 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 and overcomes some of the drawbacks of the RMSD metric. The dynamic programming method by Taylor and Orengo (1989) is similar to that of Needleman and Wunsch (1970) for sequence alignment, but has the drawback of requiring huge computational resources - time and memory. The maximal common sub-graph detection by 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 to characterize 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.

5 Structural classification of proteins

Murzin and co-workers (Murzin *et al.*, 1995; Andreeva *et al.*, 2008) developed a database called SCOP (Structural classification of proteins), through visual comparison, guided by experience and intuition (Table 1). A hierarchical organization consisting of four levels was used: 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). Here also, 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 1. 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 super-families (Sowdhamini *et al.*, 1998), are examples of databases resulting from structural bioinformatics analysis. 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).

6 Deciphering protein function through structure

6.1 Function annotation at the fold level

The ultimate purpose of studying a protein structure is to gain functional insights on how the given structure achieves the associated function. It is not surprising then, that once a structure is available, any known function from biochemical and biophysical experimental studies will be mapped onto it by associating the fold with the function as well as marking the binding site residues of the associated ligands in the protein. This type of information however, is not available in all cases, making it necessary to explore bioinformatics methods for functional annotation. Functional information can be obtained in some cases by comparison of the protein fold to that of another related 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. Sequence-structure-function relationships however are a bit more complex than the simple linear relationships among the three aspects. There are a number of instances in literatures, where dissimilar structures exhibit similar functions while also other examples that show different functions for proteins adopt

the same structural fold. Fig. 4 illustrates both these cases with known examples. Nevertheless, it is clear that high sequence similarity leads to high structural similarity and a strong likelihood of similarity in function, both in the molecular function as well as in the cellular functions. In cases where structural fold is the same, but functions are different, the two proteins are likely to have significant differences in their finer arrangement of residues, particularly at the functional sites. In the same spirit, it is possible to have proteins adopting different folds, but with similar sub-structures (Fig. 5), hence similar functions. It is also possible to have cases where a given function has arisen in two proteins independently using different structural folds and different binding site architectures as well. A case-by-case analysis of the structures at hand, generally reveals the patterns that one would expect and hence strategies that will be useful in their analyses and annotations.

It must be remembered that 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.

Structure to function models work best when there is high conservation in the entire protein, applications of which have been 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

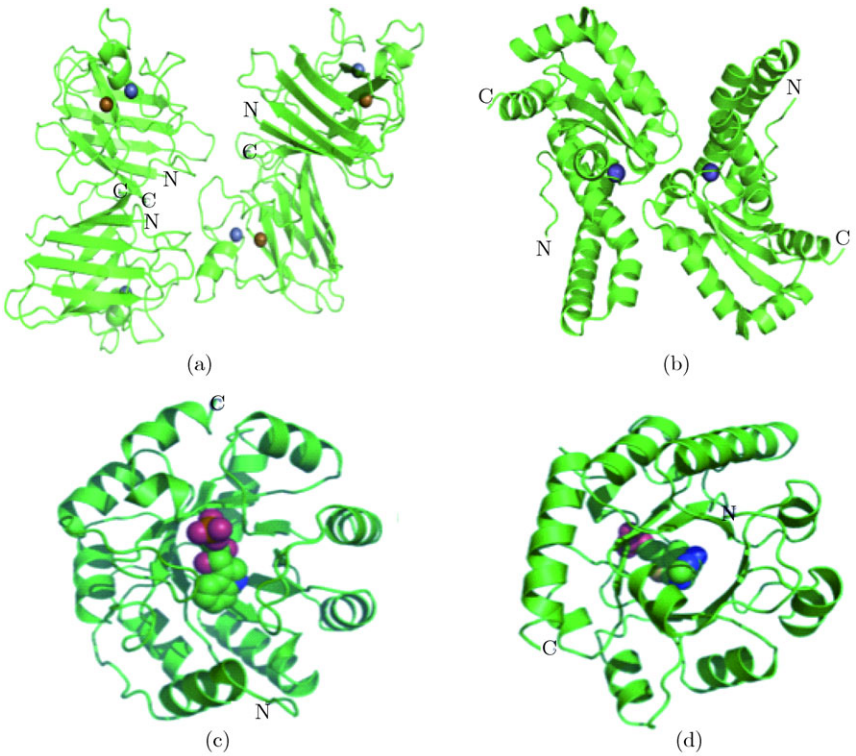


Fig. 4 Protein structure and function. Examples (a) and (b) show proteins that have different structures but the same function; while (c) and (c) show proteins that have similar structures but different functions. The specific examples shown are Cu/Zn superoxide dismutase (1SDY, SCOP-b.1.8.1) shown in (a), and Fe/Mn superoxide dismutase (1N0J, SCOP- c.87.1) shown in (b). The SCOP classes reveal the difference at the fold and class levels. On the other hand, a Thiamine Phosphate Synthase (2TPS) shown in (c) and Indole-3-Glycerolphosphate (1A53) shown in (d), both belong to the TIM barrel fold but have completely different functions. The corresponding substrates are shown in CPK form

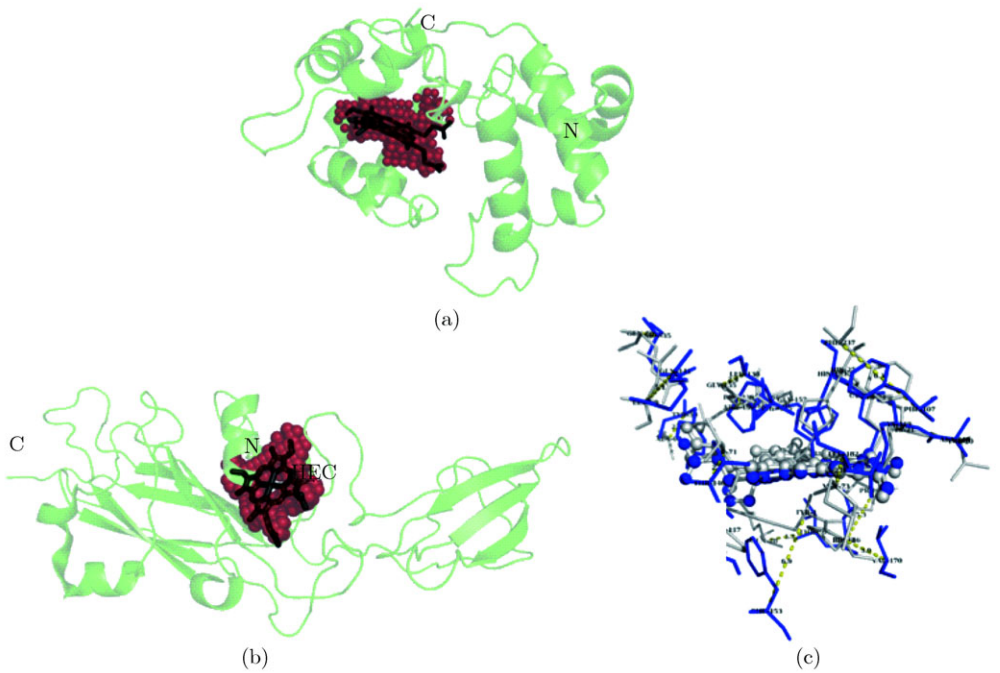


Fig. 5 Substructure comparison. An example to show two proteins belonging to completely different classes, yet exhibiting very high similarity at binding site level; Lignin Peroxidase (1LGA, SCOP-a.93.1.1) shown in (a) where as Cytochrome F (1E2Z, SCOP- b.2.6.1) shown in (b). A superposition of their binding sites along with the heme ligand is shown in (c). The ligand is shown in ball and stick and the residues are in wireframe representations

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, suggesting an alternative function in siderophore biosynthesis, rather than its annotation as a putative aminoglycoside N-acetyltransferase (Card *et al.*, 2005).

The success in deriving various relationships is of course, dependent on the method 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 have 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 (Gen-threader (McGuffin *et al.*, 2000), 3D-PSSM (Kelley *et al.*, 2000)) and structure comparison methods (DALI, VAST). Most of the sequence alignment methods are based on recognizing common sequence patterns whereas the structural alignment methods are based on recognition of common topological arrangement of sub-structures (such as the secondary structural elements).

6.2 Function annotation at binding site level

It has long been recognized that understanding ligand binding to a protein molecule holds the key to understanding function of the molecule. Therefore a different level of understanding protein function is to extract functionally important regions in them and associate them with particular function(s). A complete description of the binding sites is not always obtained, even for crystallographically determined structures, because the protein may not be complexed with all the ligands required for the function of the molecule or because the complexed ligands are often substitutes for the natural ligands. Identification of all relevant binding sites in protein molecules, therefore becomes a key step in the process of gaining functional insights from protein structures. A number of methods have emerged in the last decade for the task of locating binding sites in proteins (Goodford, 1985; Levitt and Banaszak, 1992; Kleywegt and Jones, 1994; Peters *et al.*, 1996; Hendlich *et al.*, 1997; Liang *et al.*, 1998; Brady and Stouten, 2000; Venkatachalam *et al.*, 2003; Bhinge *et al.*, 2004; An *et al.*, 2005; Coleman *et al.*, 2006; Glaser *et al.*,

2006; Huang and Schroeder, 2006; Brylinski *et al.*, 2007; Chakrabarti and Lanczycki, 2007; Landon *et al.*, 2007; Soga *et al.*, 2007; Kalidas and Chandra, 2008; Tong *et al.*, 2008; Yeturu and Chandra, 2008). 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. 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 the use of different features to represent protein molecules and subsequently predict binding sites.

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, 1999), 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).

Once the binding sites are identified through one or more of the above methods, the next task is to compare or align them with binding pockets from known structures, or in other words, known recognition sites of different ligands. Similar to the transfer of function from homologous sequences or highly similar structures, ligand binding function and hence the broader function of the protein can be inferred when there is a significant similarity of the binding sites (Fig. 5). Comparison of binding sites at the structural level however is not a trivial task and requires specialized algorithms. A number of methods are available for this purpose, including Sitesbase (Gold and Jackson, 2006), Cavbase (Kuhn *et al.*, 2007), and Pocketmatch (Yeturu and Chandra, 2008).

7 Macromolecular recognition

7.1 Protein-ligand interactions

Understanding the molecular basis of recognition of the ligands by the proteins is an important aspect of structural and molecular biology, so as to understand how proteins are capable of specific and reversible interactions with ligands. This can be achieved by studying the interactions among proteins, their internal molecular dynamics guided by its intra-molecular forces, influence of other substances such as allosteric factors and function in terms of ligand binding. The basis for protein-ligand interactions can be understood by studying the thermodynamic components, which are the driving forces for ligand binding. 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 analyzing 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 analyzing the number and nature of the hydrogen bonds involved in interaction (Fig. 6). Since a large number of high resolution protein-ligand complexes have been available for a couple of decades now, they have been utilized to derive scoring functions to compute relative binding affinities of the same protein with different ligands. The most popular among these are the empirical scoring function of Bohm and co-workers (Bohm, 1994), which take into consideration 82 protein-ligand complexes to derive relative contributions of different types of interactions, deriving an expression, so as to fit the experimentally observed values, leading to deriving appropriate weights for electrostatic interactions, surface area of interaction and other such parameters.

7.2 Protein-protein interactions

Most biological processes are carried out by macromolecular assemblies and regulated through a complex network of protein-protein interactions. These interactions with other proteins and sometimes nucleic acids are known to be important for maintaining normal physiology. 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-protein interactions would pertain mainly to the first category of interactions. A number of protein structure complexes are being determined ex-

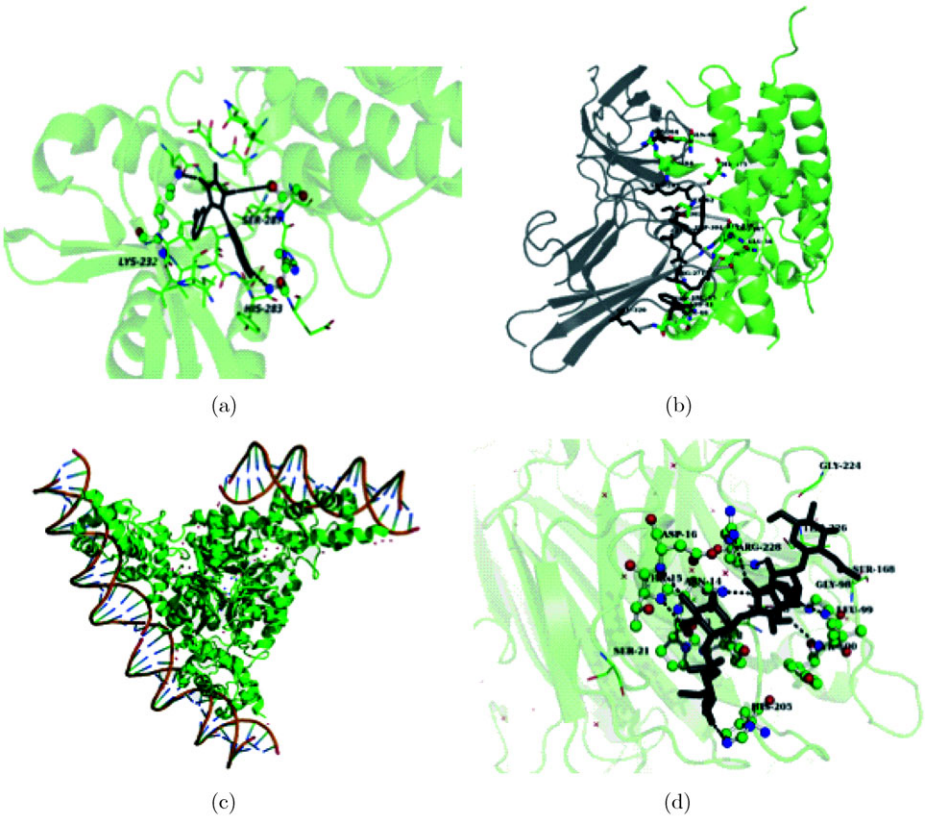


Fig. 6 Different types of molecular complexes: Protein-ligand interactions are seen in (a), protein-protein interactions are seen in (b), while protein-DNA interactions is in (c) and protein-carbohydrate interactions are shown in (d). The examples chosen for the illustration are the X-ray crystal structure of TFGBRI complexed with its inhibitor pyrazolone (3KCF), the crystal structure of human growth hormone ((1A22) shown in dark grey bound to its single receptor (light), crystal structure of arginine repressor (3FHZ), bound with its co-repressor along with DNA operator sequence and the structure of concanavalinA (1TEI) complexed with (pentose sugar shown in black), respectively

perimentally and the current release of PDB contains several protein-protein complexes (Fig. 6), 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 considering gene neighborhoods (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. There are methods that depend upon identification and comparison of surface patches (Jones and Thornton, 1997) on protein surfaces, differentiating between core and the rim residues present at the protein-protein interface to map the conservation (Guharoy and Chakrabarti, 2005), but methods in this category are in general still in their infancy with a lot of scope for improvement.

7.3 Protein-DNA interactions

DNA binding proteins have a fundamental role to play in any living organisms because they are involved in various processes such as DNA recombination (replication and maintaining genome integrity); expression (transcription and translation), genome packaging (histones and protamines), and gene regulation (promot-

ers and repressors). It therefore becomes very important for us to understand the kind of interactions made by the proteins with DNA. The interactions made by the proteins can be non-specific, examples of which are Taq-polymerase and DnaseI. Interactions in this category generally involve backbone of the DNA and are assisted by water molecules. On the other hand, proteins such as Trp repressor, Rel homology region, TATA box-binding protein (TBP) make specific interactions with specific DNA bases, making them dependent on DNA sequence at that region. Fig. 6 illustrates a similar example of a protein-DNA complex. The complexity increases further in the case of multi-specific proteins such as homeodomain, LacI and CAP, which recognize a number of different DNA segments with high specificity. Majority of protein-DNA interactions comprise of DNA backbone interactions that provide stability rather than specificity, followed by van der Waals contacts and then by hydrogen and water mediated bonds. Various computational tools are available that can predict the DNA-protein interactions. Most of these can be classified into two different categories: the first one utilizes structure-based information and requires 3D protein structures, while the second class utilizes only sequence patterns. Some examples of such tools are DISIS, DNABindR, DISPLAR; and BindN, DP-Bind, DBS-PSSM, DBS-Pred respectively (Sarai *et al.*, 2005).

7.4 Protein-carbohydrate interactions

Carbohydrates observed naturally in biological systems, are among the most diverse of molecular components, although our current knowledge on how they code for biological information is limited. The potential information in these kinds of interactions is immense because the conformational space which carbohydrates can explore is vast. X-ray crystallographic studies of these protein-sugar interactions with the current methods can give us snapshots of certain conformations, leaving a large number of other possible conformations and the interacting states of the system, to be explored. Protein-carbohydrate interactions are seen to drive many biological processes like cell adhesion, signal transduction, host-pathogen recognition, inflammation, often also serve as molecular switches (O'Conner and Imperiali, 1998) in addition to providing specific substrates for cellular interaction. Among the proteins that are known to bind to carbohydrates, the family of lectins and antibodies are well studied. Lectins have carbohydrate recognition domains and exhibit a wide range of forms with specificities for diverse carbohydrates, a remarkable range of strategies for achieving selective binding (Fig. 6) (Chandra *et al.*, 2006). Specific antibodies exist for specific recognition of carbohydrates as antigens, the binding in some cases is capable of triggering a variety of immunological reactions (Sacchettini *et al.*, 2001).

8 Dynamics in proteins

Several studies have now shown that protein molecules are not rigid bodies, rather specific movements within them are crucial for their function (Schultz-Heienbrok *et al.*, 2005). Some proteins are known to undergo substantial rearrangements in their domains or in smaller segments upon binding to other molecules, to switch between active and inactive states (Kern *et al.*, 1999). Flexibility of proteins has been associated with various functional aspects such as enabling catalytic activity, signal transduction, and various allosteric mechanisms. Currently our understanding about conformational changes in proteins is limited to a few well-studied examples (Dodson and Verma, 2006). Modeling these conformational changes is therefore of interest, to simulate and predict the nature of conformational changes and hence the flexibility in different proteins. This understanding in turn has broad implications in the field of protein design, assigning function to uncharacterized protein and in mechanism of molecular recognition. Various computational approaches have been adapted to study protein dynamics and allostery, which has an advantage of being fast in comparison and giving deeper mechanistic insights that is not possible to trace experimentally.

The database of macromolecular movements (http://molmovdb.org/) that describe and classify the motions that occur in proteins and other macromolecules, hosts a large collection of protein structures in different states. Traditionally theoretical studies on protein motion have focused on structures of single molecules to study phenomenon such as domain movement in response to ligand binding (Qi and Hayward, 2009). A few cases of different conformations of the same protein have been indeed studied by crystallography. Fig. 7 illustrates one such example. In the recent years, new computational methodologies are being explored to predict the flexibility of proteins. Techniques from graph theory are applied to analyze the bond networks in proteins with covalent, hydrogen bonds and salt bridges considered to be distance constraints to distinguish between flexible and rigid residues (Ghosh and Vishveshwara, 2007; Kuhn *et al.*, 2007). A statistical thermodynamics algorithm (Vertrees *et al.*, 2005) predicts network of cooperative residues within the proteins (Hilser, 1996), providing insights about routes mediating conformational changes. Predicting flexibility from sequence alone (Schlessinger and Rost, 2005) has also been explored.

Molecular dynamics and its other variations are one of the most commonly used methods for studying protein flexibility, by mapping the atomic positions in the trajectories and deriving insights into regions that can undergo large movements. Although these have the advantage of being highly accurate and the conforma-

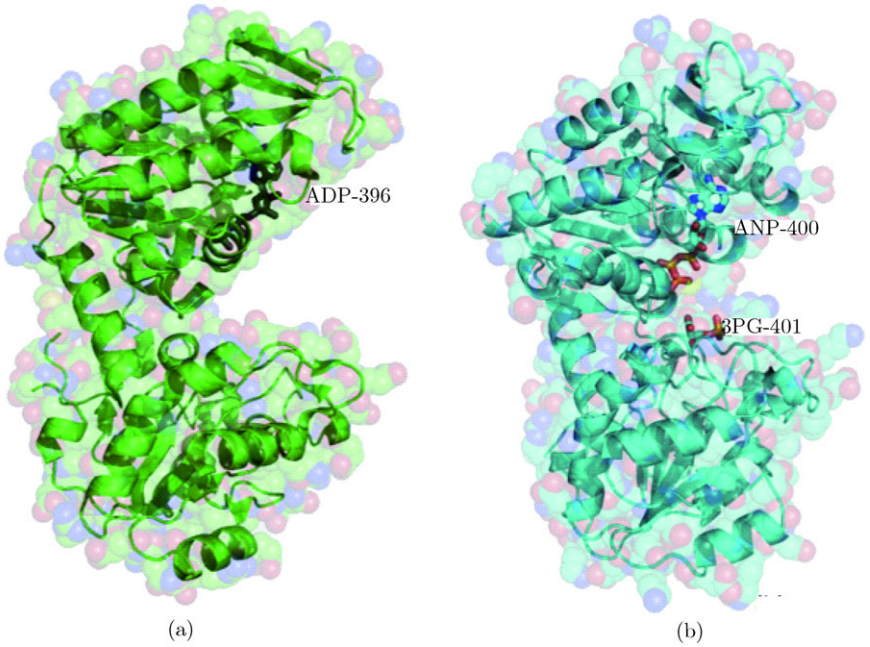


Fig. 7 An example to illustrate movements in proteins, linked to their function. The conformation change observed upon ligand binding in phosphoglycerate kinase is shown, the inactive open, but more stable conformational (1PHP) is in (a), while the active, closed conformation is in (b) (1VPE). The two ligands, an ATP analogue and the triose phosphate are also shown in the closed form, while the position of the nucleotide is indicated in the open form also

tional trajectories can be studied in details, a main drawback is the computational cost. An alternate approach to studying the conformations produced in the protein structure is through morphing. In this technique, one image is gradually changed into another, through the generation of intermediate images that help transform the starting conformation to the desired or expected end-point. It is expected that in nature that the transition would be smooth and be energetically favorable. The morph server (Flores *et al.*, 2006) uses energy minimization to calculate the intermediate frames, which produces results that are usually much better than morphs made by simpler linear interpolation, but still relatively quick. The area of predicting flexibility and different conformational states needs much more attention, and more advances in this direction can be expected in the future.

9 Applications

9.1 Structure based drug discovery

9.1.1 Target identification

Targetability refers to the assessment of the feasibility of a protein as a drug target molecule (Raman *et al.*, 2008). A further measure of feasibility is to understand the ability of the protein molecule to be accessible and specifically bind a drug-like small molecule; all of these can be studied by sequence- and structure-based methods. Knowledge of the structure of the target macromolecule helps us to estimate the feasibility of the protein as a target and also facilitates computational docking of the ligand molecule into its binding site. Function can also be better appreciated by analyzing protein structures than sequence alone. Several methods have emerged in the last few years to analyze protein structures, which can be used for evaluating their feasibility as drug targets (Scapin, 2006). Besides providing functional clues, the structures also provide a framework to understand the molecular basis of recognition, which is required both for lead design as well as for analyzing the feasibility of the target molecule. This type of analysis, however, is restricted to those proteins whose structures are either experimentally determined or predicted with high confidence by computational methods. 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.

A recent study of *M. tuberculosis* proteome demonstrates how this can be achieved through computational methods (Raman *et al.*, 2008). Possible pockets in the set of bacterial and human structures were first identified by detecting binding sites in all the proteins and then by identifying unique pockets that could serve as feasible drug targets. A similar concept has been termed as a chemical systematic biology approach,

which identifies off-target binding networks through their ligand binding sites, again with the help of binding site detection in protein structures and comparison approaches. Using this, the authors demonstrate their use in identifying drug candidates or multi-drug resistant TB and in explaining adverse effects of CETP inhibitors (Xie *et al.*, 2009).

9.1.2 Lead identification and optimization

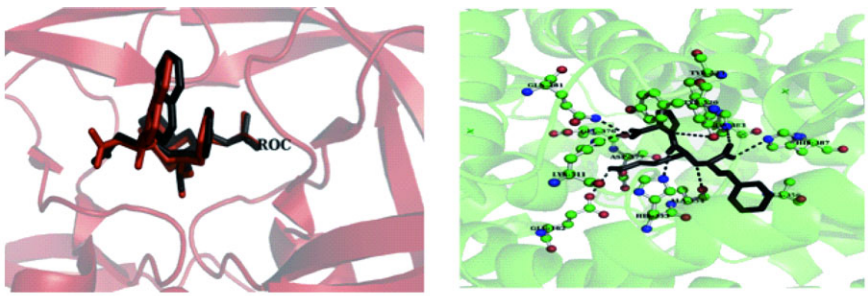
Structure-based drug design (SBDD) is a well established field for designing appropriate small molecules to enhance or inhibit the activity of the protein in question, when the structure of the target protein and the binding site details are known. Some examples of drugs designed by structure-based methods are Zanamivir and Oseltamivir against influenza neuraminidase, Nelfinavir, Amprenavir, and Lopinavir targeting HIV protease (Nair *et al.*, 2002).

9.1.3 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. Rarey and *et al.* (1996) developed FlexX where the base fragment of the ligand is placed into the binding site considering complementary 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 favorable conformations of the ligand are generated and later grouped 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)) and heuristic clique-based searches (DOCK (Ewing and Kuntz, 1998)). An example of docking of saquinavir to HIV protease is shown in Fig. 8. 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 (Zoete *et al.*, 2003; Xiang, 2006). 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.

9.1.4 Virtual high-throughput screening

As the promise of structure-based drug design begins to be realized (Congreve *et al.*, 2005), the need for expanding to a larger scale is becoming more acute. A common need in present 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.



(a) (b)

Fig. 8 Examples of structure-based drug design. Docking of saquinavir (ROC) to HIV protease (3EL4) by using AutoDock. A comparison of the docked pose (darker shade) with the crystallographically observed pose (lighter shade) is shown in (a). Details of the interactions (b) show another example of structure-based design. Lisinopril, an inhibitor of the angiotensin-converting enzyme, as bound to its target (1O86). This drug, a clinically used antihypertensive has been designed with clues derived from the structure of the target protein

10 Future perspective and challenges

Various structural biology and structural bioinformatics studies have already shown us the power of these approaches in understanding biology at high resolution. As the data generated keeps increasing in variety as well as in quality, these approaches only become even more important towards precisely reasoning out the function of the biological system and predicting the effects of modifications or perturbations. There are still some challenges ahead that have to be overcome and help us comprehend structural bases for the biological phenomena.

One such important challenge is experimental structure determination at a genomic scale. With advances in experimental methods and increased attempts in a number of structures determined, a high amount of unprocessed structural data is being generated. A requirement therefore is to understand the gaps that exist in processing such data (for example, data from twinned crystals), and determine the structures much more effectively. The other rate-limiting step is arriving at crystallization condition that works for the protein of interest. The data on crystallization conditions can also be explored; computational approaches including machine-learning methods can be applied to increase the yield of crystals (Hennessy *et al.*, 2000; Gopalakrishnan *et al.*, 2004).

Other challenges include better visualization tools as complexity of molecules demands novel display methods. It should help us synchronize the structural data with important clues such as location of functional sites, areas of structural and genetic variability. The database of known structures available today is large enough, and better classification systems are required. An effort towards this end is a database termed SIFTS (Structure Integration with function, taxonomy and sequence initiative), which maps the protein structures in PDB with the corresponding gene ontology terms. Much more work needs to be done for higher order integration with knowledge available in literature as well as those which can be computed with various high confidence bioinformatics approaches. Prediction of three-dimensional structure from just the sequence still remains an area of interest. CASP (Critical Assessment of Protein Structures) meetings still continue to be beneficial in terms of understanding the performance of different methods and the quality of homology modeling, threading, binding site prediction and *ab initio* structure prediction. A new area that can be envisaged in the near future is that of structural systems biology, integrating systems level analysis of large complex systems with the details that are obtained through high resolution studies of protein structures, which aims to predict the behavior of biological systems on the basis of a set of molecules involved. Inclusion of structural details can ultimately turns abstract system representations into models that reflect biological reality (Aloy and Russell, 2006). Recently with the use of structural genomics and systems biology, a three-dimensional reconstruction of the central metabolic network of bacterium *Thermotoga maritima* was obtained (Zhang *et al.*, 2009). Integration of structural data with network analysis can also give us insights into function, mechanism and evolution of biological systems. Eventually structural bioinformatics when cross-linked with the experimental data should provide us with valuable information about the macromolecular interactions within the cell, and their localization into compartments. It can help us give a more comprehensive view of the working of the cell, enable development of more complete computational models and enable answering questions about how various diverse molecules work together inside a cell, provide mechanisms by which they are orchestrated and, unravel the physical basis of life itself.

References

[1] Aloy, P., Russell, R.B. 2006. Structural systems biology: Modelling protein interactions. Nat Rev Mol Cell Biol 7, 188–197.

[2] An, J., Totrov, M., Abagyan, R. 2005. Pocketome via comprehensive identification and classification of ligand binding envelopes. Mol Cell Proteomics 4, 752–761.

[3] Ananthalakshmi, P., Samayamohan, K., Chokalingam, C., Mayilarasi, C., Sekar, K. 2005. Psst-2.0: Protein data bank sequence search tool. Applied Bioinformatics 4, 141–145.

[4] Andreeva, A., Howorth, D., Chandonia, J.M., Brenner, S.E., Hubbard, T.J.P., Chothia, C., Murzin, A.G. 2008. Data growth and its impact on the scop database: New developments. Nucleic Acids Research 36, D419–D425.

[5] Baldus, M. 2006. Molecular interactions investigated by multi-dimensional solid-state nmr. Curr Opin Struct Biol 16, 618–623.

[6] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E. 2000. The protein data bank. Nucleic Acids Research 28, 235–242.

[7] Bernstein, B.E., Williams, D.M., Bressi, J.C., Kuhn, P., Gelb, M.H., Blackburn, G.M., Hol, W.G. 1998. A bisubstrate analog induces unexpected conformational changes in phosphoglycerate kinase from trypanosoma brucei. J Mol Biol 279, 1137–1148.

[8] Bharath, M.M.S., Chandra, N.R., Rao, M.R.S. 2003. Molecular modeling of the chromatosome particle. Nucleic Acids Research 31, 4264–4274.

[9] Bhat, T.N., Saikrishnan, V., Vijayan, M. 1979. An analysis of side chain conformation in proteins. International Journal of Peptide and Protein Research 13, 170–184.

[10] Bhattacharya, S., Bhattacharyya, C., Chandra, N.R. 2006. Projections for fast protein structure retrieval. BMC Bioinformatics 7, Suppl 5, S5.

[11] Bhinge, A., Chakrabarti, P., Uthanumallian, K., Bajaj, K., Chakraborty, K.R. 2004. Accurate detection of protein: Ligand binding sites using molecular dynamics simulations. Structure 12, 1989–1999.

[12] Bohm, H.-J. 1994. 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.

[13] Bourne, P.E., Weissig, H. 2008. Structural Bioinformatics. Wiley InterScience, New York.

[14] Bowie, J.U., Luthy, R., Eisenberg, D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. Science 253, 164–170.

[15] Brady, G.P., Stouten, P.F. 2000. Fast prediction and visualization of protein binding pockets with pass. J Comput Aided Mol Des 14, 383–401.

[16] Brylinski, M., Kochanczyk, M., Broniatowska, E., Roterman, I. 2007. Localization of ligand binding site in proteins identified *in silico*. J Mol Model 13, 665–675.

[17] Burley, S.K. 2000. An overview of structural genomics. Nature Structure and Molecular Biology 7, 932–934.

[18] Cai, W., Shao, X., Maigret, B. 2002. Protein-ligand recognition using spherical harmonic molecular surfaces: Towards a fast and efficient filter for large virtual throughput screening. J Mol Graph Model 20, 313–328.

[19] Card, G.L., Peterson, N.A., Smith, C.A., Rupp, B., Schick, B.M., Baker, E.N. 2005. 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, 13978–13986.

[20] Chakrabarti, S., Lanczycki, C.J. 2007. Analysis and prediction of functionally important sites in proteins. Protein Sci 16, 4–13.

[21] Chandra, N., Kumar, N., Jeyakani, J., Singh, D., Gowda, S., Prathima, M. 2006. Lectindb: A plant lectin database. Glycobiology 16, 938–946.

[22] Chandra, N., Muirhead, H., Holbrook, J., Bernstein, B., Hol, W., Sessions, R. 1998. 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, 372–380.

[23] Chen, R. 2001. Enzyme engineering: Rational redesign versus directed evolution. Trends Biotechnol 19, 13–14.

[24] Cheng, H., Kim, B.H., Grishin, N.V. 2008. Malisam: A database of structurally analogous motifs in proteins. Nucleic Acids Res 36 (Database issue), 211–217.

[25] Chiu, W., Baker, M.L., Jiang, W., Dougherty, M., Schmid, M.F. 2005. Electron cryomicroscopy of biological machines at subnanometer resolution. Structure 13, 363–372.

[26] Coleman, R.G., Salzberg, A.C., Cheng, A.C. 2006. Structure based identification of small molecule binding sites using free energy model. J Chem Inf Model 46, 2631–2637.

[27] Congreve, M., Murray, C.W., Blundell, T.L. 2005. Structural biology and drug discovery. Drug Discovery Today 10, 895–907.

[28] Cozzetto, D., Kryshtafovych, A., Fidelis, K., Moult, J., Rost, B., Tramontano, A. 2009. Evaluation of template-based models in casp8 with standard measures. Proteins 77 Suppl 9, 18–28.

[29] Dandekar, T., Snel, B., Huynen, M., Bork, P. 1998. Conservation of gene order: A fingerprint of proteins that physically interact. Trends Biochem Sci 23, 324–328.

[30] Dessailly, B.H., Lensink, M.F., Orengo, C.A., Wodak, S.J. 2008. Ligasite - a database of biologically relevant binding sites in proteins with known apo-structures. Nucleic Acids Res 36 (Database issue), 667–673.

[31] Dimitropoulos, D., Ionides, J., Henrick, K. 2006. Using msdchem to search the pdb ligand dictionary. Curr Protoc Bioinformatics, Chapter 14.

[32] Dodson, G., Verma, C.S. 2006. Protein flexibility: Its role in structure and mechanism revealed by molecular simulations. Cell Mol Life Sci 63, 207–219.

[33] Dunbrack, R.L., Karplus, M. 1994. Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. Nat Struct Biol 1, 334–340.

[34] Enright, A.J., Ouzounis, C.A. 2001. Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. Genome Biol 2.

[35] Ewing, T.J.A., Kuntz, I.D. 1998. Critical evaluation of search algorithms for automated molecular docking and database screening. Journal of Computational Chemistry 18, 1175–1189.

[36] Ferrè F., Ausiello, G., Zanzoni, A., Helmer-Citterich, M. 2004. Surface: A database of protein surface regions for functional annotation. Nucleic Acids Res 32 (Database issue), 240–244.

[37] Flieschmann, R.D., Adams,M.D., White, O. Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M. 1995. Whole-genome random sequencing and assembly of haemophilus influenzaerd. Science 269, 496–512.

[38] Flores, S., Echols, N., Milburn, D., Hespenheide, B., Keating, K., Lu, J., Wells, S., Yu, E.Z., Thorpe, M., Gerstein, M., 2006. The database of macromolecular motions: New features added at the decade mark. Nucleic Acids Res 34 (Database issue), D296–D301.

[39] Forster, A.C., Church, G.M. 2006. Towards synthesis of a minimal cell. Mol Syst Biol 2, 45.

[40] Frey, T.G., Perkins, G.A., Ellisman, M.H. 2006. Electron tomography of membrane-bound cellular organelles. Annu Rev Biophys Biomol Struct 35, 199–224.

[41] Gabb, H.A., Jackson, R.M., Sternberg, M.J. 1997. Modelling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol 272, 106–120.

[42] Ghosh, A., Vishveshwara, S. 2007. A study of communication pathways in methionyl-trna synthetase by molecular dynamics simulations and structure network analysis. Proc Natl Acad Sci USA 104, 15711–15716.

[43] Glaser, F., Morris, R.J., Najmanovich, R.J., Laskowski, R.A., Thornton, J.M. 2006. A method for localizing ligand binding pockets in protein structures. Proteins 62, 479–488.

[44] Gold, N.D., Jackson, R.M. 2006. Sitesbase: A database for structur-based protein-ligand binding site comparisons. Nucleic Acids Research 34, D231–D234.

[45] Goodford, P.J., 1985. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J Med Chem 28, 849–857.

[46] Gopalakrishnan, V., Livingston, G., Hennessy, D., Buchanan, B., Rosenberg, J.M. 2004. Machine-learning techniques for macromolecular crystallization data. Acta Crystallogr D Biol Crystallogr 60 (Pt 10), 1705–1716.

[47] Goulding, C.W., Perry, L.J., Anderson, D., Sawaya, M.R., Cascio, D., Apostol, M.I., Chan, S., Parseghian, A., Wang, S.S., Wu, Y., Cassano, V., Gill, H.S., Eisenberg, D. 2003. Structural genomics of mycobacterium tuberculosis: A preliminary report of progress at ucla. Biophys Chem 105, 361–370.

[48] Gowri, V.S., Pandit, S.B., Karthik, P.S., Srinivasan, N., Balaji, S. 2003. Integration of related sequences with protein three-dimensional structural families in an updated version of pali database. Nucleic Acids Res 31, 486–488.

[49] Guharoy, M., Chakrabarti, P. 2005. Conservation and relative importance of residues across protein-protein interfaces. Proc Natl Acad Sci USA 102, 15447–15452.

[50] Hendlich, M., Rippmann, F., Barnickel, G. 1997. Ligsite: Automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph Model 15, 359–363.

[51] Hennessy, D., Buchanan, B., Subramanian, D., Wilkosz, P.A., Rosenberg, J.M. 2000. Statistical methods for the objective design of screening procedures for macromolecular crystallization. Acta Crystallogr D Biol Crystallogr 56 (Pt 7), 817–827.

[52] Hillisch, A., Pineda, L.F., Hilgenfeld, R. 2004. Utility of homology models in the drug discovery process. Drug Discov Today 9, 659–669.

[53] Holm, L., Sander, C. 1993. Protein structure comparison by alignment of distance matrices. J Mol Biol 233, 123–138.

[54] Holm, L., Sander, C., 1996. Mapping the protein universe. Science 273, 595–603.

[55] Holm, L., Sander, C. 1998. Touring protein fold space with dali/fssp. Nucleic Acids Res 26, 316–319.

[56] Hong, M. 2006. Oligomeric structure, dynamics, and orientation of membrane proteins from solid-state nmr. Structure 14, 1731–1740.

[57] Huang, B., Schroeder, M., 2006. Ligsitecsc: Predicting ligand binding sites using the connolly surface and degree of conservation. BMC Struct Biol 6, 19.

[58] Huang, T.W., Tien, A.C., Huang, W.S., Lee, Y.C., Peng, C.L., Tseng, H.H., Kao, C.Y., Huang, C.Y. 2004. Point: A database for the prediction of protein-protein interactions based on the orthologous interactome. Bioinformatics 20, 3273–3276.

[59] Ilari, A., Savino, C. 2008. Protein structure determination by x-ray crystallography. Methods Mol Biol 452, 63–87.

[60] Jackson, T. 1991. Structure and function of g protein coupled receptors. Pharmacol Ther 50, 425–442.

[61] Jiang, W., Ludtke, S.J. 2005. Electron cryomicroscopy of single particles at subnanometer resolution. Curr Opin Struct Biol 15, 571–577.

[62] Jones, C. 2005. Vaccines based on the cell surface carbohydrates of pathogenic bacteria. An Acad Bras Cienc 77, 293–324.

[63] Jones, D.T., Taylor, W.R., Thornton, J.M. 1992. A new approach to protein fold recognition. Nature 358, 86–89.

[64] Jones, S., Thornton, J.M. 1997a. Prediction of protein-protein interaction sites using patch analysis. J Mol Biol 272, 133–143.

[65] Kalidas, Y., Chandra, N. 2008a. Pocketdepth: A new depth based algorithm for identification of ligand binding sites in proteins. Journal of structural biology 161, 31–42.

[66] Kelley, L.A., Maccallum, R.M., Sternberg, M.J. 2000. Enhanced genome annotation using structural profiles in the program 3d-pssm. J Mol Biol 299, 499–520.

[67] Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H., Phillips, D.C. 1958. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature 181, 662–666.

[68] Kern, D., Volkman, B.F., Luginbühl, P., Nohaile, M.J., Kustu, S., Wemmer, D.E. 1999. Structure of a transiently phosphorylated switch in bacterial signal transduction. Nature 402, 894–898.

[69] Kleywegt, G.J. 1999. Recognition of spatial mofits in protein structures. Journal of Molecular Biology 285, 1887–1897.

[70] Kleywegt, G.J., Jones, T.A. 1994. Detection, delineation, measurement and display of cavities in macromolecular structures. Acta Crystallogr D Biol Crystallogr 50, 178–185.

[71] Krissinel, E., Henrick, K. 2004. Secondary-structure matching (ssm), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr 60, 2256–2268.

[72] Kuhn, D., Weskamp, N., Hüllermeier, E., Klebe, G. 2007. Functional classification of protein kinase binding sites using cavbase. ChemMedChem 2, 1432–1447.

[73] Kyrpides, N.C. 1999. Genomes online database (gold 1.0): A monitor of complete and ongoing genome projects world-wide. Bioinformatics 15, 773–774.

[74] Landon, M.R., Lancia Jr., D.R., Yu, J., Thiel, S.C., Vajda, S. 2007. Identification of hot spots within druggable binding regions by compuatational solvent mapping of proteins. J Med Chem 50, 1231–1240.

[75] Laskowski, R.A., Macarthur, M.W., Moss, D.S., Thornton, J.M. 1993. Procheck: A program to check the stereochemical quality of protein structures. Journal of Applied Crystallography 26, 283–291.

[76] Laskowski, R.A., Watson, J.D., Thornton, J.M. 2005. Profunc: A server for predicting protein function from 3d structure. Nucleic Acids Res 33, 89–93.

[77] Laurie, A.T., Jackson, R.M. 2005. Q-sitefinder: An energy-based method for the prediction of protein-ligand binding sites. Bioinformatics 21, 1908–1916.

[78] Lee, D., Redfern, O., Christine Orengo, C.A. 2007. Predicting protein function from sequence and structure. Nature Reviews Molecular Cell Biology 8, 995–1005.

[79] Lesley, S.A., Kuhn, P., Godzik, A., Deacon, A.M., Mathews, I., Kreusch, A., Spraggon, G., Klock, H.E., Mcmullan, D., Shin, T., Vincent, J., Robb, A., Brinen, L.S., Miller, M.D., Mcphillips, T.M., Miller, M.A., Scheibe, D., Canaves, J.M., Guda, C., Jaroszewski, L., Selby, T.L., Elsliger, M.A., Wooley, J., Taylor, S.S., Hodgson, K.O., Wilson, I.A., Schultz, P.G., Stevens, R.C. 2002. Structural genomics of the thermotoga maritima proteome implemented in a high-throughput structure determination pipeline. Proc Natl Acad Sci USA 99, 11664–11669.

[80] Levitt, D.G., Banaszak, L.J. 1992. Pocket: A computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. J Mol Graph 10, 229–234.

[81] Liang, J., Edelsbrunner, H., Woodward, C. 1998. Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. Protein Sci 7, 1884–1897.

[82] Linnainmaa, S., Harwood, D.A., Davis, L.S. 1988. Pose determination of a three-dimensional object using triangle pairs. IEEE Transactions on Pattern Analysis and Machine Intelligence 10, 634–647.

[83] Liolios, K., Mavromatis, K., Tavernarakis, N., Kyrpides, N.C. 2008. The genomes on line database (gold) in 2007: Status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 36, D475–D479.

[84] Lo, W.C., Huang, P.J., Chang, C.H., Lyu, P.C. 2007. Protein structural similarity search by ramachandran codes. BMC Bioinformatics 8, 307.

[85] Lucic, V., Förster, F., Baumeister, W. 2005. Structural studies by electron tomography: From cells to molecules. Annu Rev Biochem 74, 833–865.

[86] Marsden, R.L., Lewis, T.A., Orengo, C.A. 2007. Towards a comprehensive structural coverage of completed genomes: A structural genomics viewpoint. BMC Bioinformatics 8, 86.

[87] Mcdermott, A.E. 2004. Structural and dynamic studies of proteins by solid-state nmr spectroscopy: Rapid movement forward. Curr Opin Struct Biol 14, 554–561.

[88] Mcguffin, L.J., Bryson, K., Jones, D.T. 2000. The psipred protein structure prediction server. Bioinformatics 16, 404–405.

[89] Morris, G.M., Goodsell, D.S., Halliday, R.S., Huey, R., Hart, W.E., Belew, R.K., Olson, A.J. 1999. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. Journal of Computational Chemistry 19, 1639–1662.

[90] Moult, J. 2005. A decade of casp: Progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol 15, 285–289.

[91] Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B., Hubbard, T., Tramontano, A. 2007. Critical assessment of methods of protein structure prediction-round vii. Proteins 69 Suppl 8, 3–9.

[92] Moult, J., Pedersen, J.T., Judson, R., Fidelis, K. 1995. A large-scale experiment to assess protein structure prediction methods. Proteins 23.

[93] Muegge, I., Martin, Y.C. 1999. A general and fast scoring function for protein-ligand interactions: A simplified potential approach. Journal of Medicinal Chemistry 42, 791–804.

[94] Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C. 1995. Scop: A structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology 247, 536–540.

[95] Nair, A., Bonin, I., Tossi, A., Wels, W., Miertus, S. 2002. 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, 171–179.

[96] Needleman, S.B., Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48, 443–453.

[97] O'conner, S.E., Imperiali, B. 1998. A molecular basis for glycosylation-induced conformational switching. Chem Biol 5, 427–437.

[98] O'donoghue, S.I., Goodsell, D.S., Frangakis, A.S., Jossinet, F., Laskowski, R.A., Nilges, M., Saibil, H.R., Schafferhans, A., Wade, R.C., Westhof, E., Olson, A.J. 2010. Visualization of macromolecular structures. Nat Methods 7 (3 Suppl), S42–S55.

[99] Ondetti, M., Rubin, B., Cushman, D. 1977. Design of specific inhibitors of angiotensin-converting enzyme: New class of orally active antihypertensive agents. Science 196, 441–444.

[100] Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., Thornton, J.M. 1997. Cath-a hierarchic classification of protein domain structures. Structure 5, 1093–1108.

[101] Peitsch, M.C. 1997. Large scale protein modelling and model repository. Proc Int Conf Intell Syst Mol Biol 5, 234–236.

[102] Peng, H., Huang, N., Qi, J., Xie, P., Xu, C., Wang, J., Yang, C. 2003. Identification of novel inhibitors of bcr-abl tyrosine kinase via virtual screening. Bioorganic and Medicinal Chemistry Letters 13, 3693–3699.

[103] Peters, K.P., Fauck, J., Frömmel, C. 1996. 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.

[104] Pieper, U., Eswar, N., Braberg, H., Madhusudhan, M.S., Davis, F.P., Stuart, A.C., Mirkovic, N., Rossi, A., Marti-Renom, M.A., Fiser, A., Webb, B., Greenblatt, D., Huang, C.C., Ferrin, T.E., Sali, A. 2004. Modbase, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res 32, 217–222.

[105] Pillardy, J., Czaplewski, C., Liwo, A., Lee, J., Ripoll, D.R., Kaźmierkiewicz, R., Oldziej, S., Wedemeyer, W.J., Gibson, K.D., Arnautova, Y.A., Saunders, J., Ye, Y.J., Scheraga, H.A. 2001. Recent improvements in prediction of protein structure by global optimization of a potential energy function. Proc Natl Acad Sci USA 98, 2329–2333.

[106] Pugalenthi, G., Suganthan, P.N., Sowdhamini, R., Chakrabarti, S. 2007. Smotif: A server for structural motifs in proteins. Bioinformatics 23, 637–638.

[107] Pugalenthi, G., Suganthan, P.N., Sowdhamini, R., Chakrabarti, S. 2008. Megamotifbase: A database of structural motifs in protein families and superfamilies. Nucleic Acids Res 36, 218–221.

[108] Qi, G., Hayward, S. 2009. Database of ligand-induced domain movements in enzymes. BMC Struct Biol 9, 13.

[109] Ramachandran, G.N., Ramakrishnan, C., Sasisekharan, V. 1963. Stereochemistry of polypeptide chain configurations. J Mol Biol 7, 95–99.

[110] Raman, K., Yeturu, K., Chandra, N. 2008. Targettb: A target identification pipeline for mycobacterium tuberculosis through an interactome, reactome and genome-scale structural analysis. BMC Syst Biol 2, 109.

[111] Rarey, M., Kramer, B., Lengauer, T., Klebe, G. 1996. A fast flexible docking method using an incremental construction algorithm. J Mol Biol 261, 470–489.

[112] Renault, L., Chou, H.-T., Chiu, P.-L., Hill, R.M., Zeng, X., Gipson, B., Zhang, Z.Y., Cheng, A., Unger, V., Stahlberg, H. 2006. Milestones in electron crystallography. J Comput Aided Mol Des 20, 519–527.

[113] Richardson, D.C., Richardson, J.S. 1992. The kinemage: A tool for scientific communication. Protein Sci 1, 3–9.

[114] Rost, B., Schneider, R., Sander, C. 1997. Protein fold recognition by prediction-based threading. J Mol Biol 270, 471–480.

[115] Russell, R.B., Saqi, M.A., Sayle, R.A., Bates, P.A., Sternberg, M.J. 1997. Recognition of analogous and homologous protein folds: Analysis of sequence and structure conservation. J Mol Biol 269, 423–439.

[116] Sacchettini, J.C., Baum, L.G., Brewer, C.F. 2001. Multivalent protein-carbohydrate interactions. A new paradigm for supermolecular assembly and signal transduction. Biochemistry 40, 3009–3015.

[117] Sánchez, R., Sali, A. 1997. Advances in comparative protein-structure modelling. Curr Opin Struct Biol 7, 206–214.

[118] Sayers, E. 2005. Pubchem: An entrez database of small molecules. NLM Technical Bulletin 342, e2.

[119] Scapin, G. 2006. Structural biology and drug discovery. Curr Pharm Des 12, 2087–2097.

[120] Schlessinger, A., Rost, B. 2005. Protein flexibility and rigidity predicted from sequence. Proteins 61, 115–126.

[121] Schultz-Heienbrok, R., Maier, T., Sträter, N. 2005. A large hinge bending domain rotation is necessary for the catalytic function of escherichia coli 5'-nucleotidase. Biochemistry 44, 2244–2252.

[122] Seiler, K.P., George, G.A., Happ, M.P., Bodycombe, N.E., Carrinski, H.A., Norton, S., Brudz, S., Sullivan, J.P., Muhlich, J., Serrano, M., Ferraiolo, P., Tolliday, N.J., Schreiber, S.L., Clemons, P.A. 2008. Chembank: A small-molecule screening and cheminformatics resource database. Nucleic Acids Res 36, 351–359.

[123] Shin, J.-M., Cho, D.-H. 2005. Pdb-ligand: A ligand database based on pdb for the automated and customized classification of ligand-binding structures. Nucleic Acids Research 33, D238–D241.

[124] Shindyalov, I.N., Bourne, P.E. 1998. Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. Protein Eng 11, 739–747.

[125] Snel, B., Lehmann, G., Bork, P., Huynen, M.A. 2000. String: A web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res 28, 3442–3444.

[126] Sobolev, V., Sorokine, A., Prilusky, J., Abola, E.E., Edelman, M. 1999. Automated analysis of interatomic contacts in proteins. Bioinformatics 15, 327–332.

[127] Soga, S., Shirai, H., Kobori, M., Hirayama, N. 2007. Use of amino acid composition to predict ligand-binding sites. J Chem Inf Model 47, 400–406.

[128] Sowdhamini, R., Burke, D.F., Huang, J.F., Mizuguchi, K., Nagarajaram, H.A., Srinivasan, N., Steward, R.E., Blundell, T.L. 1998. Campass: A database of structurally aligned protein superfamilies. Structure 6, 1087–1094.

[129] Stark, A., Russell, R.B. 2003. Annotation in three dimensions. Pints: Patterns in non-homologous tertiary structures. Nucleic Acids Research 31, 3341–3344.

[130] Sun, S. 1993. Reduced representation model of protein structure prediction: Statistical potential and genetic algorithms. Protein Sci 2, 762–785.

[131] Szustakowski, J.D., Weng, Z. 2000. Protein structure alignment using a genetic algorithm. Proteins 38, 428–440.

[132] Tagari, M., Tate, J., Swaminathan, G.J., Newman, R., Naim, A., Vranken, W., Kapopoulou, A., Hussain, A., Fillon, J., Henrick, K., Velankar, S. 2006. E-msd: Improving data deposition and structure quality. Nucleic Acids Res 34, D287–D290.

[133] Tanrikulu, Y., Schneider, G. 2008. Pseudoreceptor models in drug design: Bridging ligand- and receptor-based virtual screening. Nature Reviews Drug Discovery 7, 667–677.

[134] Taylor, W.R., Orengo, C.A. 1989. Protein structure alignment. Journal of Molecular Biology 208, 1–22.

[135] Tong, W., Williams, R.J., Wei, Y., Murga, L.F., Ko, J., Ondrechen, M.J. 2008. Enhanced performance in prediction of protein active sites with thematics and support vector machines. Protein Sci 17, 333–341.

[136] Tzakos, A.G., Grace, C.R., Lukavsky, P.J., Riek, R. 2006. Nmr techniques for very large proteins and rnas in solution. Annu Rev Biophys Biomol Struct 35, 319–342.

[137] Unger, R., 2004. The genetic algorithm approach to protein structure prediction. Structure and Bonding 110, 153–175.

[138] Venkatachalam, C.M., Jiang, X., Oldfield, T., Waldman, M. 2003. Ligandfit: A novel method for the shape-directed rapid docking of ligands to protein active sites. J Mol Graph Model 21, 289–307.

[139] Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., Mckusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C. 2001. The sequence of the human genome. Science 291, 1304–1351.

[140] Wallace, C.J. 1993. Understanding cytochrome c function: Engineering protein structure by semisynthesis. FASEB J 7, 505–515.

[141] Wang, Y., Xiao, Suzek, T.O., Zhang, J., Wang, J., Bryant, S.H. 2009. PubChem: A public information system for analyzing bioactivities of small molecules. Nucleic Acids Res 37, W623–W633.

[142] Watkins, H.A., Baker, E.N. 2006. 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.

[143] Wuthrich, K. 2003. Nmr studies of structure and function of biological macromolecules. Biosci Rep 23, 119–168.

[144] Xiang, Z. 2006. Advances in homology protein structure modeling. Curr Protein Pept Sci 7, 217–227.

[145] Xie, L., Li, J., Xie, L., Bourne, P.E. 2009. Drug discovery using chemical systems biology: Identification of the protein-ligand binding network to explain the side effects of cetp inhibitors. PLoS Comput Biol 5, e1000387.

[146] Yeturu, K., Chandra, N. 2008. Pocketmatch: A new algorithm to compare binding sites in protein structures. BMC Bioinformatics 9, 543.

[147] Zhang, Y., Thiele, I., Weekes, D., Li, Z., Jaroszewski, L., Ginalski, K., Deacon, A.M., Wooley, J., Lesley, S.A., Wilson, I.A., Palsson, B., Osterman, A., Godzik, A. 2009. Three-dimensional structural view of the central metabolic network of thermotoga maritima. Science 325, 1544–1549.

[148] Zhu, J., Weng, Z. 2005. Fast: A novel protein structure alignment algorithm. Proteins 58, 618–627.

[149] Zoete, V., Michielin, O., Karplus, M. 2003. 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.