Entwicklung von rechnergestützten Ansätzen für strukturelle Klassifikation, Analyse und Vorhersage von molekularen Erkennungsregionen in Proteinen

DISSERTATION
zur Erlangung des akademischen Grades
Doctor rerum naturalium (Dr. rer. Nat.)
vorgelegt

der Fakultät Mathematik und Naturwissenschaften
der Technischen Universität Dresden
von

Master in Bioinformatik Joan Teyra i Canaleta
geboren am 01.12.1979 in Girona, Spanien

Gutachter:
- Prof. Michael Schröder, Bioinformatics at BIOTEC TU-Dresden
- Prof. Bernard Hoflack, Proteomics at BIOTEC TU-Dresden

Eingereicht am: Juli
Verteidigt am: October

Die Dissertation wurde in der Zeit von August 2005 bis
Juli 2010 im BIOTEC TU Dresden angefertigt.

Development of computational approaches for structural classification, analysis and prediction of molecular recognition regions in proteins

Joan Teyra i Canaleta

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

Technical University of Dresden

2010

Program Authorized to Offer Degree: Biology

The research in this thesis has been carried out at the Structural Bioinformatics Group headed by Dr. María Teresa Pisabarro at the Biotechnology Center (BIOTEC), part of the Technical University of Dresden (TU-Dresden).

Biotechnologisches Zentrum

TECHNISCHE UNIVERSITÄT DRESDEN

The research carried out in this thesis has been fully supported by - Klaus Tschira Foundation - grant awarded to Dr. María Teresa Pisabarro.

KTS
Klaus Tschira Stiftung gemeinnützige GmbH

The author of this dissertation was awarded with two different short term fellowships that allowed him to carry on the experimental work in external laboratories.

EMBO

FEBS

Technical University of Dresden
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Joan Teyra i Canaleta

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of the Supervisory Committee:

_____

Michael Brand

Reading Committee:

_____

Prof. Bernard Hoflack

_____

Prof. Michael Schröder

Date: _____

Selbständigkeitserklärung und Erklärung zur Annerkennung der Promotionsordnung

Hiermit versichere ich, Joan Teyra, dass ich die vorliegende Arbeit:

Strukturelle Klassifikation, Analyse und Vorhersage von Protein-Bindungsregionen.
Studium des Lösungsmittels in Interfaces der Proteinen

ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.
Die Dissertation wurde von Dr. M. Teresa Pisabarro, Structural Bioinformatics, BIOTEC TU-Dresden betreut und im Zeitraum vom August 2005 bis Juli 2010 verfasst.

Meine Person betreffend erkläre ich hiermit, dass keine früheren erfolglosen Promotionsverfahren stattgefunden haben.

Ich erkenne die Promotionsordnung der Fakultät für Mathematik und Naturwissenschaften, Technische Universität Dresden an.

Technical University of Dresden

Summary

Development of computational approaches
for structural classification, analysis and prediction
of molecular recognition regions in proteins

Joan Teyra i Canaleta

Chair of the Supervisory Committee:
Professor Michael Brand
Faculty of Biology

The vast and growing volume of 3D protein structural data stored in the PDB contains abundant information about macromolecular complexes, and hence, data about protein interfaces. Non-covalent contacts between amino acids are the basis of protein interactions, and they are responsible for binding affinity and specificity in biological processes. In addition, water networks in protein interfaces can also complement direct interactions contributing significantly to molecular recognition, although their exact role is still not well understood.

It is estimated that protein complexes in the PDB are substantially underrepresented due to their crystallization difficulties. Methods for automatic classification and description of the protein complexes are essential to study protein interfaces, and to propose putative binding regions. Due to this strong need, several protein-protein interaction databases have been developed. However, most of them do not take into account either protein-peptide complexes, solvent information or a proper classification of the binding regions, which are fundamental components to provide an accurate description of protein interfaces.

In the first stage of my thesis, I developed the SCOWLP platform, a database and web application that structurally classifies protein binding regions at family level and defines accurately protein interfaces at atomic detail. The analysis of the results showed that protein-peptide complexes are substantially represented in the PDB, and are the only source

of interacting information for several families. By clustering the family binding regions, I could identify 9,334 binding regions and 79,803 protein interfaces in the PDB. Interestingly, I observed that 65% of protein families interact to other molecules through more than one region and in 22% of the cases the same region recognizes different protein families. The database and web application are open to the research community (www.scowlp.org) and can tremendously facilitate high-throughput comparative analysis of protein binding regions, as well as, individual analysis of protein interfaces.

SCOWLP and the other databases collect and classify the protein binding regions at family level, where sequence and structure homology exist. Interestingly, it has been observed that many protein families also present structural resemblances within each other, mostly across folds. Likewise, structurally similar interacting motifs (binding regions) have been identified among proteins with different folds and functions. For these reasons, I decided to explore the possibility to infer protein binding regions independently of their fold classification. Thus, I performed the first systematic analysis of binding region conservation within all protein families that are structurally similar, calculated using non-sequential structural alignment methods. My results indicate there is a substantial molecular recognition information that could be potentially inferred among proteins beyond family level. I obtained a 6 to 8 fold enrichment of binding regions, and identified putative binding regions for 728 protein families that lack binding information. Within the results, I found out protein complexes from different folds that present similar interfaces, confirming the predictive usage of the methodology. The data obtained with my approach may complement the SCOWLP family binding regions suggesting alternative binding regions, and can be used to assist protein-protein docking experiments and facilitate rational ligand design.

In the last part of my thesis, I used the interacting information contained in the SCOWLP database to help understand the role that water plays in protein interactions in terms of affinity and specificity. I carried out one of the first high-throughput analysis of solvent in protein interfaces for a curated dataset of transient and obligate protein complexes. Surprisingly, the results highlight the abundance of water-bridged residues in protein interfaces

(40.1% of the interfacial residues) that reinforces the importance of including solvent in protein interaction studies (14.5% extra residues interacting only water-mediated). Interestingly, I also observed that obligate and transient interfaces present a comparable amount of solvent, which contrasts the old thoughts saying that obligate protein complexes are expected to exhibit similarities to protein cores having a dry and hydrophobic interfaces. I characterized novel features of water-bridged residues in terms of secondary structure, temperature factors, residue composition, and pairing preferences that differed from direct residue-residue interactions. The results also showed relevant aspects in the mobility and energetics of water-bridged interfacial residues.

Collectively, my doctoral thesis work can be summarized in the following points:

1. I developed SCOWLP, an improved framework that identifies protein interfaces and classifies protein binding regions at family level.

2. I developed a novel methodology to predict alternative binding regions among structurally similar protein families independently of the fold they belong to.

3. I performed a high-throughput analysis of water-bridged interactions contained in SCOWLP to study the role of solvent in protein interfaces.

These three components of my thesis represent novel methods for exploiting existing structural information to gain insights into protein-protein interactions, key mechanisms to understand biological processes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to acknowledge many people for helping me during my doctoral work. I would especially like to thank my advisor, Mayte Pisabarro, for her generous time and commitment. Throughout my doctoral work she encouraged me to develop independent thinking and research skills. She continually stimulated my analytical thinking and greatly assisted me with scientific writing, an specimen difficult to find nowadays.

I extend many thanks to my colleagues from the Structural Bioinformatics Group: Andres, Aurelie, Carsten, Frank, Gerd, Hongbo, Ionut, Jana, Jens, John, Rainer and especially, Maciej and Sergey for their constant advices and fructiferous discussions, which turned out with lot of fun and resulted on constant research collaborations. I am also very grateful to Mandy and Ralf for their exceptional assistance and help in the day-by-day work.

I am greatly indebted to Carles and Xavi with whom I shared all my University and Master studies. They awakened my interest in the field of Bioinformatics.

I owe a special note of gratitude to Michael Schröder, without him I would most probably have never done my PhD in Dresden. I'd like to thank those from Schröder's group who helped me at the beginning of the PhD, many of them still very good friends.

I wish to thank Irene and Jose for allowing me to use their lab and equipment for the Interleukin-16 work. I am also very grateful to Javi for teaching and supporting me in the lab. I thank also Bertrand to help me in the lab, always there when I needed him.

Unfortunately, I can't thank everyone who helped with my dissertation. All scientific colleagues and the friends that nothing have to do with science that in one way or another positively influenced my PhD: 'Thank you everybody!'

Finally, I'd like to thank my parents and especially my girlfriend, Mònica, a constant source of support and children production during all these years. Without her, definitely, nothing would have been the same. Marc and Pol Teyra i Galiano are good proof of it.

Chapter 1

INTRODUCTION

This introductory chapter contains basic concepts and techniques used during my doctoral thesis time. It first connects the concepts of protein sequence, structure and function to protein interactions. It also provides an overview of the main databases and explains the basics of the computational biomolecular simulation techniques used during the thesis. Finally, it describes the chapters contained in this dissertation.

[1]Snapshot of the authors PhD live. The image in the cover represents the 'word clouds' obtained from the author's bookmarks created during his PhD. The clouds give greater prominence to words that appear more frequently in the source data. The image is generated using Wordle (www.wordle.net)

1.1 Protein sequence, structure and function

1.1.1 Protein sequence

Proteins are the most abundant biological macromolecules, occurring in all cells and all parts of cells. Proteins also occur in great variety; thousands of different kinds, ranging in size from relatively small peptides to huge polymers with molecular weights in the millions, may be found in a single cell. Moreover, proteins exhibit enormous diversity of biological functions and are the most important final products of the information pathways. Proteins are the molecular instruments through which genetic information is expressed. Relatively simple monomeric subunits provide the key to the structure of the thousands of different proteins. All proteins, whether from the most ancient lines of bacteria or from the most complex forms of life, are constructed from the same ubiquitous set of 20 amino acids, covalently linked in characteristic linear sequences. Because each of these amino acids has a side chain with distinctive chemical properties, this group of 20 precursor molecules may be regarded as the alphabet in which the language of protein structure is written. What is most remarkable is that cells can produce proteins with strikingly different properties and activities by joining the same 20 amino acids in many different combinations and sequences. From these building blocks different organisms can make such widely diverse products as enzymes, hormones, antibodies, transporters, muscle fibers, and myriad other substances having distinct biological activities. Among these protein products, the enzymes are the most varied and specialized.

Knowledge of the sequence of amino acids in a protein can offer insights into its three-dimensional structure and its function, cellular location, and evolution. Most of these insights are derived by searching for similarities with other known sequences. Thousands of sequences are known and available in databases accessible through the Internet (Table 1.1). A comparison of a newly obtained sequence with this large bank of stored sequences often reveals relationships both surprising and enlightening. Exactly how the amino acid sequence determines three-dimensional structure is not understood in detail, nor can we always predict function from sequence. However, protein families that have some shared structural or functional features can be readily identified on the basis of amino acid sequence similarities.

Table 1.1: Databases to analyze protein sequences and domains.

Type and name of the tool	**References**
Sequence databases	
UniProt	UniProt Consortium [1]
Ncbi	National Center for Biotechnology Information [2]
Domain databases	
SMART	Schultz et al. [3]
Pfam	Bateman et al. [4]
PROSITE	Hulo et al. [5]
CDD	Marchler-Bauer et al. [6]
Domain classification	
CATH	Orengo et al. [7]
SCOP	Murzin et al. [8]
FSSP	Holm et al. [9]

Individual proteins are assigned to families based on the degree of similarity in amino acid sequence. Members of a family are usually identical across 25% or more of their sequences, and proteins in these families generally share at least some structural and functional characteristics. Some families are defined, however, by identities involving only a few amino acid residues that are critical to a certain function (Table 1.1).

1.1.2 Protein structure

The covalent backbone of a typical protein contains hundreds of individual bonds. Because free rotation is possible around many of these bonds, the protein can assume an unlimited number of conformations. However, each protein has a specific chemical or structural function, strongly suggesting that each has a unique three-dimensional structure.

In a globular protein, different segments of a polypeptide chain (or multiple polypeptide chains) fold back on each other. This folding generates a compact form relative to polypeptides in a fully extended conformation. The folding also provides the structural diversity necessary for proteins to carry out a wide array of biological functions. Globular proteins include enzymes, transport proteins, motor proteins, regulatory proteins, immunoglobulins,

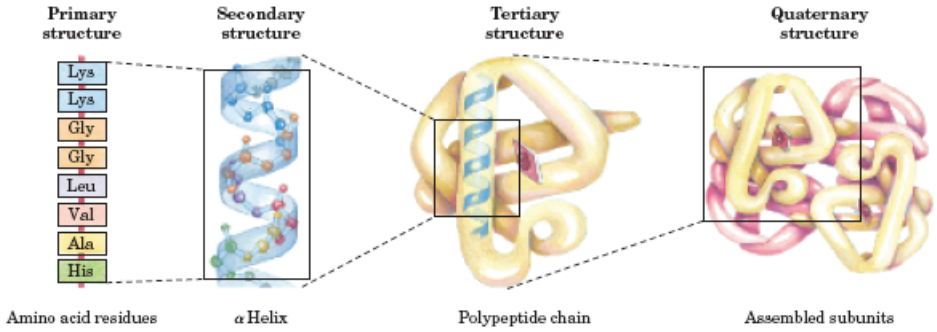


Figure 1.1: Levels of structure in proteins. The primary structure consists of a sequence of amino acids linked together by peptide bonds and includes any disulfide bonds. The resulting polypeptide can be coiled into units of secondary structure, such as an α helix. The helix is a part of the tertiary structure of the folded polypeptide, which is itself one of the subunits that make up the quaternary structure of the multisubunit protein, in this case hemoglobin.

and proteins with many other functions.

For large macromolecules such as proteins, the tasks of describing and understanding structure are approached at several levels of complexity, arranged in a kind of conceptual hierarchy. Four levels of protein structure are commonly defined (Fig 1.1). A description of all covalent bonds (mainly peptide bonds and disulfide bonds) linking amino acid residues in a polypeptide chain is its primary structure. The most important element of primary structure is the sequence of amino acid residues. Secondary structure refers to particularly stable arrangements of amino acid residues giving rise to recurring structural patterns. Tertiary structure describes all aspects of the three-dimensional folding of a polypeptide. When a protein has two or more polypeptide subunits, their arrangement in space is referred to as quaternary structure.

As a new millennium begins, the number of known three-dimensional protein structures is in the thousands and more than doubles every two years. This wealth of structural information is revolutionizing our understanding of protein structure, the relation of structure to function, and even the evolutionary paths by which proteins arrived at their present state, which can be glimpsed in the family resemblances that are revealed as protein databases are sifted and sorted. The sheer variety of structures can seem daunting. Yet as new protein structures become available it is becoming increasingly clear that they are manifestations

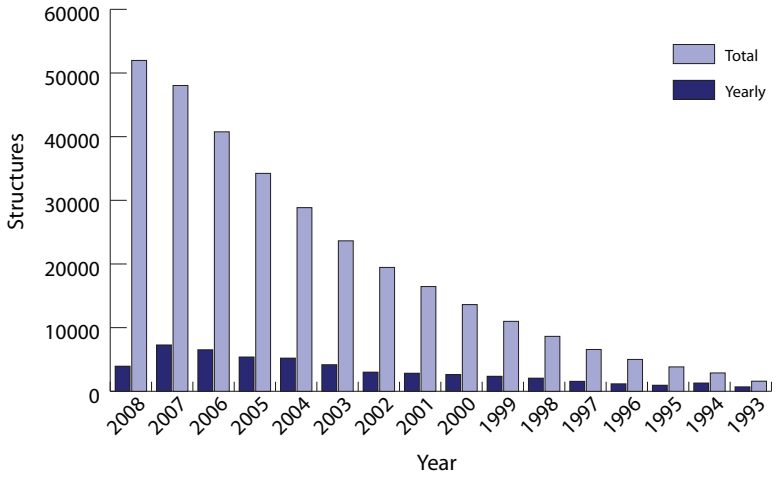


Figure 1.2: PDB statistics. Number of structures deposited in the PDB per year (single and accumulated).

of a finite set of recognizable, stable folding patterns.

Such discussions are possible only because of the vast amount of information available over the Internet from resources such as the Protein Data Bank (PDB; [10]), an archive of experimentally determined three-dimensional structures of biological macromolecules, including proteins and nucleic acids. When the PDB was originally founded in 1973 it contained just 7 protein structures. Since then it has undergone an almost exponential growth in the number of structures, which does not show any sign of falling off (Fig 1.2).

1.1.3 Protein motifs

The three-dimensional structure of a typical globular protein can be considered an assemblage of polypeptide segments in the α-helix and β-sheet conformations, linked by connecting segments. The structure can then be described to a first approximation by defining how these segments stack on one another and how the segments that connect them are arranged. This formalism has led to the development of databases that allow informative comparisons of protein structures, complementing other databases that permit comparisons of protein sequences (1.1). An understanding of a complete three-dimensional structure is built upon an analysis of its parts.

Supersecondary structures, also called motifs or simply folds, are particularly stable arrangements of several elements of secondary structure and the connections between them. There is no universal agreement among biochemists on the application of the three terms, and they are often used interchangeably. The terms are also applied to a wide range of structures. Recognized motifs range from simple to complex, sometimes appearing in repeating units or combinations. A single large motif may comprise the entire protein.

Polypeptides with more than a few hundred amino acid residues often fold into two or more stable, globular units called domains. In many cases, a domain from a large protein will retain its correct three-dimensional structure even when it is separated (for example, by proteolytic cleavage) from the remainder of the polypeptide chain. Different domains often have distinct functions, such as the binding of small molecules or interaction with other proteins. Small proteins usually have only one domain (the domain is the protein).

Following these rules, complex motifs can be built up from simple ones. For example, a series of β-α-β loops, arranged so that the β strands form a barrel, creates a particularly stable and common motif called the α/β barrel (Fig 1.3).

As we have seen, the complexities of tertiary structure are decreased by considering substructures. Taking this idea further, researchers have organized the complete contents of databases according to hierarchical levels of structure. The Structural Classification of Proteins (SCOP, [8]) database offers a good example of this very important trend in biochemistry. At the highest level of classification, the SCOP database (http://scop.mrc-lmb.cam.ac.uk/scop) borrows a scheme already in common use, in which protein structures are divided into four classes: all α, all β, α/β (in which the α and β segments are interspersed or alternate), and $\alpha + \beta$ (in which the α and β regions are somewhat segregated) (Fig 1.3). Within each class are tens to hundreds of different folding arrangements, built up from increasingly identifiable substructures. Some of the substructure arrangements are very common, others have been found in just one protein. A variety of motifs array among the four classes of protein structure. The number of folding patterns is not infinite, however. As the rate at which new protein structures are elucidated has increased, the fraction of those structures containing a new motif has steadily declined. Fewer than 1,000 different folds or motifs may exist in all proteins. In SCOP the top two levels of organization, class

and fold, are purely structural. Below the fold level, categorization is based on evolutionary relationships. Many examples of recurring domain or motif structures are available, and these reveal that protein tertiary structure is more reliably conserved than primary sequence. The comparison of protein structures can thus provide much information about evolution. Proteins with significant primary sequence similarity, and/or with demonstrably similar structure and function, are said to be in the same protein family. A strong evolutionary relationship is usually evident within a protein family. For example, the globin family has many different proteins with both structural and sequence similarity to myoglobin. Two or more families with little primary sequence similarity sometimes make use of the same major structural motif and have functional similarities; these families are grouped as superfamilies. An evolutionary relationship between the families in a superfamily is considered probable, even though time and functional distinctions – hence different adaptive pressures – may have erased many of the telltale sequence relationships. A protein family may be widespread in all three domains of cellular life, the Bacteria, Archaea, and Eukarya, suggesting a very ancient origin. Other families may be present in only a small group of organisms, indicating that the structure arose more recently. Tracing the natural history of structural motifs, using structural classifications in databases such as SCOP, provides a powerful complement to sequence analyses in tracing many evolutionary relationships. The SCOP database is curated manually, with the objective of placing proteins in the correct evolutionary framework based on conserved structural features.

Two similar enterprises, the CATH (class, architecture, topology, and homologous superfamily) and FSSP (fold classification based on structure-structure alignment of proteins) databases, make use of more automated methods and can provide additional information (Table 1.1). Structural motifs become especially important in defining protein families and superfamilies. Improved classification and comparison systems for proteins lead inevitably to the elucidation of new functional relationships. Given the central role of proteins in living systems, these structural comparisons can help illuminate every aspect of biochemistry, from the evolution of individual proteins to the evolutionary history of complete metabolic pathways.

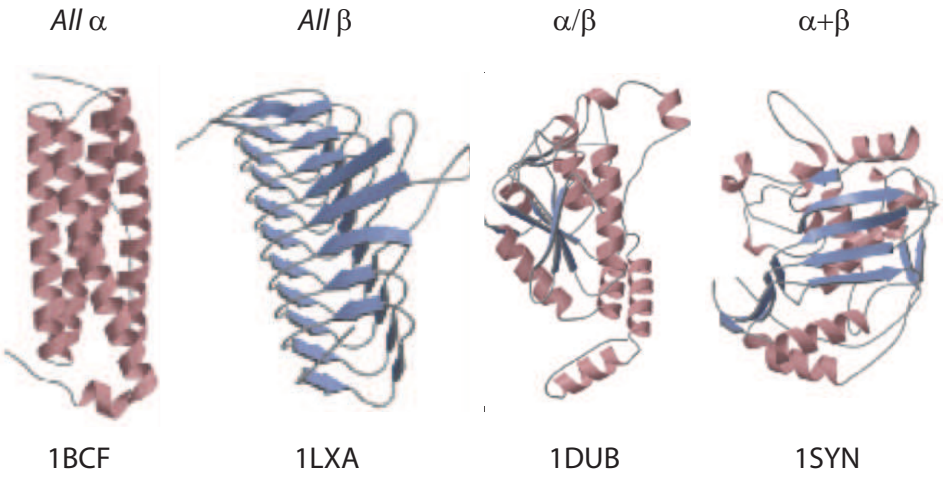


Figure 1.3: Organization of proteins based on motifs. Shown here are just a small number of the hundreds of known stable motifs. They are divided into four classes: all α, all β, α/β, and $\alpha + \beta$. Structural classification data from the SCOP (Structural Classification of Proteins) database (http://scop.mrc-lmb.cam.ac.uk/scop) are also provided. The PDB identifier is the unique number given to each structure archived in the Protein Data Bank (www.rcsb.org/pdb).

1.1.4 From dimers to large complexes

Many proteins have multiple polypeptide subunits. The association of polypeptide chains can serve a variety of functions. Many multisubunit proteins have regulatory roles; the binding of small molecules may affect the interaction between subunits, causing large changes in the protein's activity in response to small changes in the concentration of substrate or regulatory molecules. In other cases, separate subunits can take on separate but related functions, such as catalysis and regulation. Some associations, such as the fibrous proteins and the coat proteins of viruses, serve primarily structural roles. Some very large protein assemblies are the site of complex, multi-step reactions. One example is the ribosome, site of protein synthesis, which incorporates dozens of protein subunits along with a number of RNA molecules. A multisubunit protein is also referred to as a multimer. Multimeric proteins can have from two to hundreds of subunits. A multimer with just a few subunits is often called an oligomer. If a multimer is composed of a number of nonidentical subunits, the overall structure of the protein can be asymmetric and quite complicated. However, most multimers have identical subunits or repeating groups of nonidentical subunits, usu-

ally in symmetric arrangements. The repeating structural unit in such a multimeric protein, whether it is a single subunit or a group of subunits, is called a protomer.

1.2 Protein complexes

1.2.1 Protein interactions

Knowing the three-dimensional structure of a protein is an important part for understanding how the protein functions. Proteins are dynamic molecules whose functions almost invariably depend on interactions with other molecules, and these interactions are affected in physiologically important ways by sometimes subtle, sometimes striking changes in protein conformation. The importance of molecular interactions to a protein‘s function can hardly be overemphasized. Most of these interactions are fleeting, though they may be the basis of complex physiological processes such as oxygen transport, immune function, and muscle contraction. The functions of many proteins involve the reversible binding of other molecules. A molecule bound reversibly by a protein is called a ligand. A ligand may be any kind of molecule, including another protein. The transient nature of protein-ligand respond rapidly and reversibly to changing environmental and metabolic circumstances. A ligand binds at a site on the protein called the binding site, which is complementary to the ligand in size, shape, charge, and hydrophobic or hydrophilic character. Furthermore, the interaction is specific: the protein can discriminate among the thousands of different molecules in its environment and selectively bind only one or a few. A given protein may have separate binding sites for several different ligands. These specific molecular interactions are crucial in maintaining the high degree of order in a living system.

Proteins are flexible. Changes in conformation may be subtle, reflecting molecular vibrations and small movements of amino acid residues throughout the protein. A protein flexing in this way is sometimes said to "breathe". Changes in conformation may also be quite dramatic, with major segments of the protein structure moving as much as several nanometers. Specific conformational changes are frequently essential to a protein's function. The binding of a protein and ligand are often coupled to a conformational change in the protein that makes the binding site more complementary to the ligand, permitting tighter binding. The structural adaptation that occurs between protein and ligand is called

induced fit. In a multisubunit protein, a conformational change in one subunit often affects the conformation of other subunits. Interactions between ligands and proteins may be regulated, usually through specific interactions with one or more additional ligands. These other ligands may cause conformational changes in the protein that affect the binding of the first ligand.

Two properties of a protein characterize its interaction with ligands. Affinity refers to the strength of binding between a protein and ligand; the equilibrium constant Keq or the dissociation constant Kd for binding is a measure of affinity. Specificity refers to the ability of a protein to bind one molecule in preference to other molecules. Both properties depend on the structure of the ligand-binding site on a protein, which is designed to fit its partner like a mold. For high-affinity and highly specific interactions to occur, the shape and chemical surface of the binding site must be complementary to the ligand molecule.

Enzymes represent a special case of protein function. Enzymes bind and chemically transform other molecules' they catalyze reactions. The molecules acted upon by enzymes are called reaction substrates rather than ligands, and the ligand-binding site is called the catalytic site or active site.

1.2.2 Physicochemical properties of protein interactions

Covalent bonds, which hold the atoms within an individual molecule together, are formed by the sharing of electrons in the outer atomic orbitals. The distribution of shared as well as unshared electrons in outer orbitals is a major determinant of the three-dimensional shape and chemical reactivity of molecules.

Noncovalent bonds are critical in maintaining the three-dimensional structures of large molecules such as proteins and nucleic acids. Noncovalent bonds also enable one large molecule to bind specifically but transiently to another, making them the basis of many dynamic biological processes.

Weak interactions occur in an aqueous system. Water describes physical and chemical properties that are of crucial importance for the structure and function of biomolecules. Noncovalent interactions responsible of strength and specificity of recognition among biomolecules are decisively influenced by the solvent properties of water, including its ability to form hy-

drogen bonds with itself and with solutes. In the following sections, we describe the four types of noncovalent interactions among macromolecules in aqueous solvent.

The hydrogen bond

Normally, a hydrogen atom forms a covalent bond with only one other atom. However, a hydrogen atom covalently bonded to a donor atom, D, may form an additional weak association, the hydrogen bond, with an acceptor atom, A: In order for a hydrogen bond to form, the donor atom must be electronegative, so that the covalent D —H bond is polar. The acceptor atom must also be electronegative, and its outer shell must have at least one nonbonding pair of electrons that attracts the $\delta+$ charge of the hydrogen atom. In biological systems, both donors and acceptors are usually nitrogen or oxygen atoms, especially those atoms in amino (—NH_2) and hydroxyl (—OH) groups (Fig 1.4). Because all covalent N —H and O —H bonds are polar, their H atoms can participate in hydrogen bonds. By contrast, C —H bonds are nonpolar, so these H atoms are hardly never involved in a hydrogen bond.

Water molecules provide a classic example of hydrogen bonding. The hydrogen atom in one water molecule is attracted to a pair of electrons in the outer shell of an oxygen atom in an adjacent molecule. Not only do water molecules hydrogen-bond with one another, but also they form hydrogen bonds with other kinds of molecules. The presence of hydroxyl (—OH) or amino (—NH_2) groups makes many molecules soluble in water. In general, molecules with polar bonds that easily form hydrogen bonds with water can dissolve in water and are said to be hydrophilic (Greek, 'water-loving'). Most hydrogen bonds are 0.26 – 0.31 nm long, about twice the length of covalent bonds between the same atoms. In particular, the distance between the nuclei of the hydrogen and oxygen atoms of adjacent hydrogen-bonded molecules in water is approximately 0.27 nm, about twice the length of the covalent O —H bonds in water. The hydrogen atom is closer to the donor atom, D, to which it remains covalently bonded, than it is to the acceptor. The length of the covalent D —H bond is a bit longer than it would be if there were no hydrogen bond, because the acceptor 'pulls' the hydrogen away from the donor. The strength of a hydrogen bond in water ($\approx$ 5 kcal/mol) is much weaker than a covalent O —H bond ($\approx$ 110 kcal/mol).

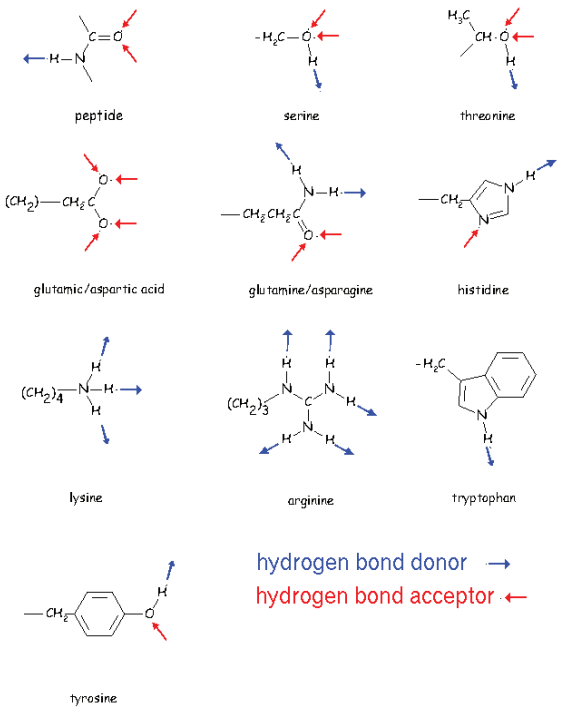


Figure 1.4: Amino acids' hydrogen bond properties.

An important feature of all hydrogen bonds is directionality. In the strongest hydrogen bonds, the donor atom, the hydrogen atom, and the acceptor atom all lie in a straight line. Nonlinear hydrogen bonds are weaker than linear ones; The strengths of the hydrogen bonds in proteins and nucleic acids are only 1 to 2 kcal/mol, considerably weaker than the hydrogen bonds between water molecules.

Electrostatic interactions and salt bridges

Electrostatic interactions are the forces that electric charges exert on each other, described by Coulomb's law. The magnitude of the electrostatic force (F) on a charge (q_1) due to the presence of a second charge (q_2), is given by:

$$F = k_e \frac{q_1 q_2}{r_2} \tag{1.1}$$

where r is the distance between the two charges and k_e a proportionality constant. A positive force implies a repulsive interaction, while a negative force implies an attractive interaction.

Salt bridges are close range electrostatic interaction between two side-chains oppositely charged in proteins. The anion is a carboxilate (from Asp or Glu) and the cation is an ammonium (RNH_3^+, from Lys) or a guanidinium ($RNHC(NH_2)_2^+$), from Arg). If a salt bridge is located on the surface of the protein, the dielectric constant should be close to that of pure water. Such an exposed salt bridge might contribute very little to protein stability. Again ammonium acetate is dissociated in water, so an Asp• • •Lys salt bridge should be weak or negligible in bulk water. Alternatively, the salt bridge might be found somewhat or completely buried in the interior of the protein. The consensus from a large number of studies of salt bridges is that they can contribute to protein stability, but there is considerable variation. Typically, a surface-exposed salt bridge is worth around anywhere from 0 to 2 kcal/mol, and a buried salt bridge can can be worth up to 3 kcal/mol, with some exceptional cases being worth more.

Van der Waals interactions

When any two atoms approach each other closely, they create a weak, nonspecific attractive force that produces a van der Waals interaction, named for Dutch physicist Johannes Diderik van der Waals (1837 – 1923), who first described it. These nonspecific interactions result from the momentary random fluctuations in the distribution of the electrons of any atom, which give rise to a transient unequal distribution of electrons, that is, a transient electric dipole. If two noncovalently bonded atoms are close enough together, the transient dipole in one atom will perturb the electron cloud of the other. This perturbation generates a transient dipole in the second atom, and the two dipoles will attract each other weakly. Similarly, a polar covalent bond in one molecule will attract an oppositely oriented dipole in another.

Van der Waals interactions, involving either transient induced or permanent electric dipoles, occur in all types of molecules, both polar and non-polar. The strength of van der

Waals interactions decreases rapidly with increasing distance; thus these noncovalent bonds can form only when atoms are quite close to one another. However, if atoms get too close together, they become repelled by the negative charges in their outer electron shells. When the van der Waals attraction between two atoms exactly balances the repulsion between their two electron clouds, the atoms are said to be in van der Waals contact (Fig 1.5). Each type of atom has a van der Waals radius at which it is in van der Waals contact with other atoms. The van der Waals radius of an H atom is 0.1 nm, and the radii of O, N, C, and S atoms are between 0.14 and 0.18 nm. Two covalently bonded atoms are closer together than two atoms that are merely in van der Waals contact. For a van der Waals interaction, the internuclear distance is approximately the sum of the corresponding radii for the two participating atoms. Thus the distance between a C atom and an H atom in van der Waals contact is 0.27 nm, and between two C atoms is 0.34 nm. In general, the van der Waals radius of an atom is about twice as long as its covalent radius. For example, a C —H covalent bond is about 0.107 nm long and a C —C covalent bond is about 0.154 nm long.

The energy of the van der Waals interaction is about 1 kcal/mol, only slightly higher than the average thermal energy of molecules at 25 °C. Thus the van der Waals interaction is even weaker than the hydrogen bond, which typically has an energy of 1 – 2 kcal/mol in aqueous solutions. The attraction between two large molecules can be appreciable, however, if they have precisely complementary shapes, so that they make many van der Waals contacts when they come into proximity.

Hydrophobic bonds

Nonpolar molecules do not contain ions, possess a dipole moment, or become hydrated. Because such molecules are insoluble or almost insoluble in water, they are said to be hydrophobic (Greek, 'water-fearing'). The covalent bonds between two carbon atoms and between carbon and hydrogen atoms are the most common nonpolar bonds in biological systems. The force that causes hydrophobic molecules or nonpolar portions of molecules to aggregate together rather than to dissolve in water is called the hydrophobic bond. This is not a separate bonding force; rather, it is the result of the energy required to insert a nonpolar molecule into water. A nonpolar molecule cannot form hydrogen bonds with wa-

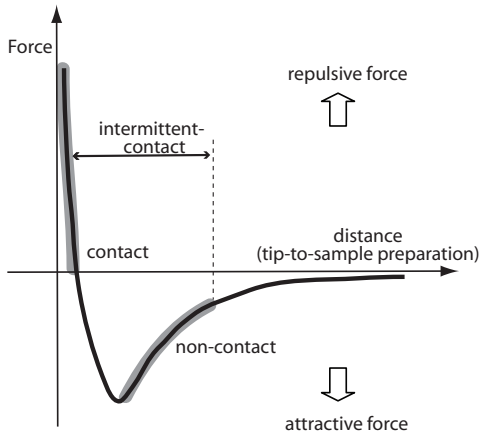


Figure 1.5: Van der Waals energy representation

ter molecules, so it distorts the usual water structure, forcing the water into a rigid cage of hydrogen-bonded molecules around it. Water molecules are normally in constant motion, and the formation of such cages restricts the motion of a number of water molecules; the effect is to increase the structural organization of water. This situation is energetically unfavorable because it decreases the randomness (entropy) of the population of water molecules.

The opposition of water molecules to having their motion restricted by forming cages around hydrophobic molecules or portions thereof is the major reason why hydrophobic molecules are essentially insoluble in water and interact mainly with other hydrophobic molecules. Nonpolar molecules can also bond together, albeit weakly, through van der Waals interactions. The net result of the hydrophobic and van der Waals interactions is a very powerful tendency for hydrophobic molecules to interact with one another, and not with water.

Weak interactions are crucial to macromolecular structure and function

The noncovalent interaction we have described are individually much weaker than the covalent bonds. Proteins contain so many sites of potential hydrogen bonding or ionic , van der Waals, or hydrophobic interactions that the cumulative effect of the many small binding

forces can be enormous. For macromolecules, the most stable complex is usually the one in which weak-bonding possibilities are maximized. Receptors extensive surfaces provide many opportunities for weak interactions.

Besides contributing to the stability of large biological molecules, multiple noncovalent bonds can also confer specificity by determining which regions of different molecules will bind together. All types of these weak interactions are effective only over a short range and require close contact between the reacting groups. For noncovalent bonds to form properly, there must be a complementarity between the sites on the two interacting surfaces, where almost any other arrangement of the same groups on the two surfaces would not allow the molecules to bind so tightly. Such multiple, specific interactions allow protein molecules to bind.

1.3 Computational biomolecular simulations: molecular dynamics

Molecular dynamics (MD) is a form of computer simulation that solves Newton's equations of motion for a system of N interacting atoms:

$$m_i \frac{\delta^2 \vec{r}_i}{\delta t^2} = \vec{F}_i \tag{1.2}$$

where forces are:

$$\vec{F}_i = \frac{-\delta V}{\delta \vec{r}_i} \tag{1.3}$$

where V is the potential.

The equations are solved simultaneously in small time steps, so that the temperature and pressure remain at the required values, and the coordinates are written to an output file at regular intervals. The coordinates as a function of time represent a trajectory of the system. After initial changes, the system usually reaches an equilibrium state. By averaging over an equilibrium trajectory many macroscopic properties can be extracted from the output file. The pipeline typical for energy minimization and MD in common is represented in

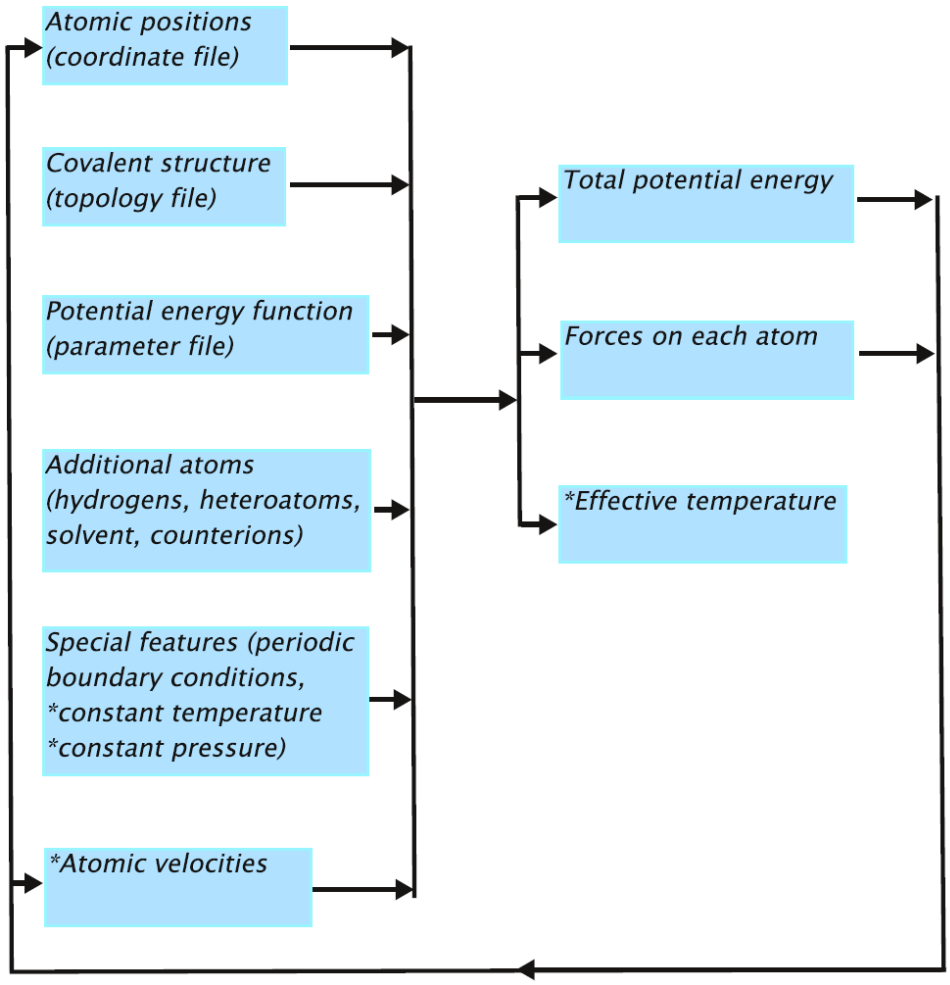


Figure 1.6: Schematic flow chart of algorithms for energy minimization and MD. Features which apply only to molecular dynamics are indicated with asterix. Each cycle of energy minimization represents a step in conformation space, while each cycle of molecular dynamics represents a step in time [11].

the Figure 1.6. Further on, we give some definitions and details on the stages and aspects important for MD runs.

1.3.1 Forcefield

Forcefield refers to the functional form and parameter sets used to describe the potential energy of a system of particles (typically but not necessarily atoms). Force field functions and parameter sets are derived from both experimental work and high-level quantum me-

chanic calculations. All-atom forcefields provide parameters for every atom in a system, including hydrogens, while united-atom forcefields could treat the hydrogen and carbon atoms in methyl and methylene groups, for example, as a single interaction center. Coarse-grained forcefields, which are frequently used in long-time simulations of proteins, provide even more abstracted representations for increased computational efficiency. More detailed description of the physical nature behind potential included in the force fields is given in the subsection 1.2.2. In comparison to analytical expression for the potentials mentioned before, bonded interactions are explicitly decomposed into bond, torsion angle and dihedral angle potentials:

$$V_{bonded} = V_{bond} + V_{angle} + V_{dihedral} \tag{1.4}$$

Electrostatic interaction usually contains explicitly only the component of the equation 1.1, which describes point-charge interactions, unless a force-field is polarizable and, therefore, has a different expression for the electrostatic potential. So, in general, the potential has the following form:

$$V(\vec{r}) = \sum_{bonds} K_r(r-r_{eq})^2 + \sum_{angles} K_\theta(\theta-\theta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2}(1+\cos[n\phi-\gamma]) + \sum_{i<j}^{atoms} (\frac{A_{ij}}{r_{ij}^2}) - \frac{B_{ij}}{r_{ij}^2}) + \sum_{i<j}^{atoms} \frac{q_i q_j}{\epsilon R_{ij}} \tag{1.5}$$

and polarization component is usually expressed as:

$$E_{pol} = -\frac{1}{2}\sum_{i}^{atom} \vec{\mu}_i \vec{E_i^{(0)}} \tag{1.6}$$

where μ_i is an induced atomic dipole and $E_i^{(0)}$ is an initial electric field causing this polarization. In addition, charges that are not centered on atoms, but are off-center (as for lone-pairs) that can be included in the forcefield.

1.3.2 Energy minimization

Energy minimization methods are common techniques to compute an equilibrium configuration of molecules, which should be a stable state and correspond to a local minimum of their potential energy. This kind of calculations generally start from an arbitrary state of molecules, then the mathematical procedure of optimization allows to move atoms (to vary coordinates) in a way to reduce the net forces (the gradients of potential energy) to nearly zero (or defined cut-off value). From a computational viewpoint, the problem of minimizing the energy of a model macromolecular system falls into the general area of nonlinear optimization problems. The functional $V(\vec{r})$ should be minimized in the multidimensional space $\vec{r}$. In case of a model with N atoms, there are 3Ndimensional space (since each atom has 3 cartesian coordinates), and V is the potential energy of the system. There are two most popular algorithms used for the minimization: steepest descent and conjugate gradient.

Steepest descent

Steepest descent is a first-order optimization algorithm. To find a local minimum of a function using steepest descent, one takes steps proportional to the negative of the gradient (or the approximate gradient) of the function at the current point. For each next iteration:

$$\vec{r}_k = \vec{r}_{k-1} + \lambda_k \frac{\vec{F}}{F} \tag{1.7}$$

where λ_k is a positive coefficient. Since the vector $\frac{\vec{F}}{F}$ is parallel to the negative gradient of the energy, it points straight downhill. The weaknesses of steepest descent are:

1. The algorithm can take many iterations to converge towards a local minimum, if the curvature in different directions is very different.

2. Finding the optimal λ per step can be a time-consuming task. Conversely, using a fixed λ can yield poor results.

Conjugate gradient

Because of these limitations, steepest descent is usually followed by conjugate gradient in the minimization procedure. This technique combines information on the current gradient with that based on the gradient at previous steps. Iteratively:

$$\vec{r}_k = \vec{r}_{k-1} + \lambda_k \left(\vec{F}_k + \frac{F_k^2}{F_{k-1}^2} \frac{\vec{F}_{k-1}}{F_{k-1}} \right) \tag{1.8}$$

It can be proven that for a quadratic surface, the search direction specified by the last equation passes through minimum on the Nth step for an N-dimensional surface, as long as the minimum along each successive search direction is found. Even if the step size is not optimal, the conjugate gradient method still yields a search direction that is superior to that of steepest descent.

1.3.3 Verlet integration

Verlet integration [12] is a numerical method frequently used to integrate Newton's equations of motion. The Verlet integrator offers greater stability than the much simpler Euler method, as well as other properties that are important in physical systems such as time-reversibility. Stability of the technique depends heavily upon either a uniform update rate, or the ability to accurately identify positions at a small time intervals into the past.

1.3.4 Periodic boundary conditions

Periodic boundary conditions (PBC) are a set of boundary conditions that are often used to simulate a large system by modeling a small part that is far from its edge. A unit cell of a certain geometry is defined, and when an object passes through one face of the unit cell, it reappears on the opposite face with the same velocity. The simulation is of an infinite perfect tiling of the system. The tiled copies of the unit cell are called images, of which there are infinitely many. During the simulation, only the properties of the unit cell need to be recorded and propagated. The minimum-image convention is a common form of PBC particle bookkeeping, where each individual particle in the simulation interacts with the closest image of the remaining particles in the system.

In MD PBC are usually applied to simulate bulk gases, liquids, crystals or mixtures. A common application used in this thesis is to use PBC to simulate solvated macromolecules in a bath of explicit solvent. Since MD contains electrostatic interactions, the net electrostatic charge of the system must be zero to avoid summing to an infinite charge when PBC is applied. However, there are still some artifacts originated from the correlations between unit cells and artificial interactions between 'heads' and 'tails' of different unit cells.

PBC requires the unit cell to be a shape that tiles perfectly into a three-dimensional crystal. Thus, a spherical or elliptical droplet cannot be used. A cube or rectangular prism is the most intuitive and common choice, but can be computationally expensive due to unnecessary amounts of solvent molecules in the corners, distant from the central macromolecules. A common alternative that requires less volume is the truncated octahedron.

1.3.5 Particle mesh Ewald

Particle Mesh Ewald (PME) method is utilized for electrostatic calculations in PBC. PME uses Ewald summation, which is a special case of the Poisson summation formula, replacing the summation of interaction energies in real space with an equivalent summation in Fourier space. The advantage of this approach is the rapid convergence of the Fourier-space summation compared to its real-space equivalent when the real-space interactions are long-ranged. Because electrostatic energies consist of both short- and long-range interactions, it is maximally efficient to decompose the interaction potential into a short-range component summed in real space and a long-range component summed in Fourier space [13]. In practice, the cut-off for PME is defined by a researcher, who uses MD, and depends on the size of a unit in PBC. The cut-off should be less or equal to the half of unit's dimension, to assure the convergence of electrostatic summation in the direct space.

1.3.6 Temperature coupling

Temperature coupling is a technique used for maintenance of constant temperature in PBC in NTP or NTV microcanonical ensembles. There are several ways to carry out temperature couplings. In the weak-coupling algorithm a single scaling factor is used for all atoms [14]. This algorithm ensures that the total kinetic energy is appropriate for the desired tem-

perature but does not control that the temperature is even over all parts of the molecule. Atomic collisions tend to ensure an even temperature distribution, but in reality this is not guaranteed, and there are many subtle problems that can arise with weak temperature coupling [15]. Andersen temperature coupling scheme [16] implies imaginary collisions, which randomize the velocities to a distribution corresponding to the fixed constant temperature. The dynamics process is Newtonian. Hence time correlation functions can be computed and the results averaged over an initial canonical distribution. Too high collision rate slows down the speed at which the molecules explore configuration space, whereas too low rate means that the canonical distribution of energies is sampled slowly [17]. Use of Langevin dynamics is an another approach for temperature coupling. There is a collision frequency parameter in this algorithm to be defined and a simple leapfrog integrator (a variant of Verlet integration) is used to propagate the dynamics, with the kinetic energy adjusted to be correct for the harmonic oscillator case [18, 19]. A collision frequency parameter is not necessary equal to the physical collision frequency. In fact, it is often advantageous, in terms of sampling or stability of integration, to use much smaller values for this parameters than the physically relevant ones.

1.3.7 Pressure coupling

Pressure coupling adjusts the volume of the unit cell (gradually on each step) to make the computed pressure approach the target pressure. Equilibration with NTP microcanonical ensemble is generally necessary to adjust the density of the system to appropriate values. Pressure coupling algorithms are often analogous to weak temperature coupling [14].

1.3.8 SHAKE algorithm

SHAKE algorithm is used to perform bond length constraints [20]. It is normally utilized for hydrogens in MD simulations. The size of the MD time step is determined by the fastest motions in the system. SHAKE removes the bond stretching freedom, which is the fastest motion, and consequently allows a larger time step to be used, resulting in speeding up the calculations. For water models, a special "three-point" algorithm is used [21]. Since SHAKE is an algorithm based on dynamics, the minimizer is not aware of what SHAKE is

doing; for this reason, minimizations should be carried out generally without SHAKE.

1.3.9 Counterions

Counterions are used in MD with PBC to make a charge of a unit neutral to avoid problems with electrostatics. For counterions Na+, K+ and Cl- are usually used. A study on counterions impact to MD results in AMBER shows that the simulations of solvated proteins are moderately sensitive to the presence of counterions. However, this sensitivity is highly dependent on the starting model and different procedures of equilibration used. The neutralized systems tend to evince smaller root mean square deviations regardless of the system investigated and the simulation procedure used. The results of parameterized fitting of the simulated structures to the crystallographic data, giving quantitative measure of the total charge influence on the stability of various elements of the secondary structure, revealed a clear scatter of different reactions of various systems' secondary structures to counterions addition: some systems apparently were stabilized when neutralized, while the others were not. Thus, one cannot unequivocally state, despite consideration of specific simulation conditions, whether protein secondary structures are more stable when they have neutralized charges. This suggests that caution should be taken when claiming the stabilizing effect of counterions in simulations involving small, unstable polypeptides or highly charged proteins [22].

1.3.10 MM-PBSA/MM-GBSA

The MM-PBSA/MM-GBSA (Molecular Mechanics-Poisson-Bolzmann Surface Area/Molecular Mechanics-Generalized Born Surface Area) approach represents the post-processing method to evaluate free energies of binding or to calculate absolute free energies of molecules in solution. The MM-PBSA/GBSA method combines molecular mechanical energies with the continuum solvent approaches. Often, the key quantity that needs to be computed is the total free energy of the molecule in the presence of solvent, which could be written as:

$$E_{tot} = E_{vac} + \delta G_{solv} \tag{1.9}$$

where E_{vac} represents a molecule's energy in vacuum (gas-phase), and δG_{solv} is the free energy of transferring the molecule from vacuum into solvent, i.e. solvation free energy. Usually it is assumed that E_{vac} is given by a classical potential function, or forcefield, that breaks the interaction down into various physical components, such as bond and angle stretching, torsional twist, and also VDW and Coulomb interactions between its atoms were assumed (subsection 1.2.2).

To estimate the total solvation free energy of a molecule δG_{solv} one typically assumes that it can be decomposed into the electrostatic and non-electrostatic components:

$$\delta G_{solv} = \delta G_{el} + \delta G_{nonel} \tag{1.10}$$

where δG_{nonel} is the free energy of solvating a molecule from which all charges have been removed (i.e. partial charges of every atom are set to zero), and δG_{el} is the free energy of first removing all charges in the vacuum, and then adding them back in the presence of a continuum solvent environment. The above decomposition, which is yet another approximation, is the basis of the widely used MM-PBSA scheme [23]. Generally speaking, δG_{nonel} comes from the combined effect of two types of interaction: the favorable van der Waals attraction between the solute and solvent molecules, and the unfavorable cost of breaking the structure of the solvent around the solute. δG_{nonel} is described in terms of solvent accessible surface area (ASA), the surface area of a biomolecule that is accessible to a solvent. The ASA is usually measured in $\mathring{A}^2$. ASA is typically calculated using the 'rolling ball' algorithm [24], which uses a sphere (of solvent) of a particular radius to probe the surface of the molecule. The choice of the probe radius does have an effect on the observed surface area, as using a smaller probe radius detects more surface details and, therefore, reports a larger surface. A typical value is 1.4 $\mathring{A}$ (also used as a default value in a popular NACCESS program [25]), which approximates the radius of a water molecule. Another factor that affects the results is the definition of the VDW radii of the atoms in the studied molecule. For example, the molecule may often lack hydrogen atoms which are implicit in the structure. The hydrogen

atoms may be implicitly included in the atomic radii of the heavy atoms, with a measure called the group radii.

The ASA is closely related to the concept of the solvent-excluded surface (also known as the molecular surface or Connolly surface), which is imagined as a cavity in bulk solvent (effectively the inverse of the solvent-accessible surface). In practice, it is also calculated via a rolling-ball algorithm [26] and independently implemented three-dimensionally in two studies [27, 28]. Connolly spent several more years perfecting the method [29] and it is thus sometimes called the Connolly surface.

The ASA can be used in protein interfaces characterization (difference of the ASA of complex and unbound components gives the size of an interface of the complex) and for empirical estimation of hydration energy, which is considered to be proportional to ASA. Within the PBSA, δG_{nonel} is supposed to be proportional to the total solvent ASA of the molecule, with a constant derived from experimental solvation energies of small non-polar molecules:

$$\delta G_{nonel} \sim ASA \tag{1.11}$$

which is an approximation, but arguably not the most critical one in the hierarchy of assumptions that form the foundation of the implicit solvent methodology [30]. In a model of continuous solvent the remained component δG_{el} can be easily calculated if the potential distribution $\varphi(r)$ in space is known. This distribution is described by Poisson-Bolzmann equation for the charge density $\rho(r)$ distribution in a dielectric with constant $\epsilon(r)$:

$$\nabla\epsilon(r)\nabla\varphi(r) = -4\pi\rho(r) + \kappa^2\epsilon(r)\varphi(r) \tag{1.12}$$

where κ is Debye-Huckel parameter. However, in molecular dynamics applications, the associated computational costs are often very high, as the Poisson-Boltzmann equation needs to be solved every time the conformation of the molecule changes. The Generalized Born (GB) model is an approximation to the exact (linearized) Poisson-Boltzmann (PB)

equation. The GB approach is computationally more effective than PB approach and it is an approximation to the exact PB equation. It is based on modeling a protein as a volume whose internal dielectric constant differs from the external solvent. The model has the following functional form:

$$G_{el,GB} = \frac{1}{8\pi}\left(\frac{1}{\epsilon_0} - \frac{1}{\epsilon})\right)\sum_{1<j}^{N}\frac{q_i q_j}{f_{GB}} \tag{1.13}$$

where:

$$f_{GB} = \sqrt{r_{ij}^2 + a_{ij}^2 e^{-D}} \tag{1.14}$$

and:

$$D = \left(\frac{r_{ij}}{2a_{ij}}\right)^2, a_{ij} = \sqrt{a_i a_j} \tag{1.15}$$

where, ϵ_0 is the dielectric constant *in vacuo*, q_i is the electrostatic charge on particle i, r_{ij} is the distance between particles i and j, and a_i is a quantity (with the dimension of length) known as the effective Born radius [31]. The effective Born radius of an atom characterizes its degree of burial inside the solute; qualitatively it can be thought of as the distance from the atom to the molecular surface. Accurate estimation of the effective Born radii is critical for the GB model [32]. Often, the challenge in these calculations is to extract frames from trajectory to essentially sample the conformational space. Depending on a system, there is a different number of randomly chosen frames to be enough for this purpose. Averaged structure of a system is usually not used since even for two equally probable rotamer states of a sidechain its average would not correspond to the physically relevant state of the system. The most important impact of conformational changes is on the electrostatic energy component, that is why at the first step, statistical sampling could be carried out only for less computationally expensive electrostatic component.

For calculations of entropies in the MM-PBSA method normal mode analysis is used. Nevertheless, entropy components of free energies are still the least accurate in MM-PBSA energy calculations [33].

1.4 Description of the work

My doctoral thesis is organized in the following five chapters:

- Chapter I (this chapter) contains basic concepts and techniques used during my doctoral thesis time. It first connects the concepts of protein sequence, structure and function to protein interactions. It also provides an overview of the main databases and explains the basics of the computational biomolecular simulation techniques used during the thesis. Finally, it describes the chapters contained in this dissertation.

- Chapter II. All available 3D structural information of proteins at atomic resolution is stored in the Protein Data Bank (PDB), and its growth is exponential with no sign of falling off. For this reason, there are already many databases that obtain the interacting information from the PDB, although most of them miss fundamental components for the interface definition and a classification of family binding regions. In this chapter, I describe the SCOWLP platform, a database and web application that aims to improve the other existing databases of protein interfaces by i) the inclusion of peptidic ligands, interfacial water molecules and more accurate interaction rules, and ii) the classification of protein binding regions by families. SCOWLP is presented by means of two peer-reviewed research articles published during my PhD in the BMC Bioinformatics Journal.

- Chapter III. Proteins exhibit an enormous diversity of structures, and different protein families, even different folds, may present non-obvious structural similarities. The detection of these similarities may lead to the discovery of evolutionary relationships and reveal common structural features important for function. In this chapter, I describe a methodology I developed to calculate binding regions conservation between structurally similar proteins. The non-obvious similarities are obtained performing

pair wise non-sequential structural alignments between all proteins independently of their fold classification. This work is under revision for publication in the Proteins Journal.

- Chapter IV. Protein interactions take place in aqueous solution, and water molecules can mediate residue-residue interactions complementing direct interactions. The specific role that water plays in binding affinity and specificity is still not well understood. This lack of knowledge makes solvent often ignored in many computational protein interaction analysis, rational drug design and protein docking studies. In this chapter, I make use of the novel interfacial information contained in the SCOWLP database (see Chapter 2) to study the characteristics of solvent and water-bridged residues in protein interfaces. In the first section, I carry out a descriptive analysis of interfacial solvent in a representative dataset of high-resolution protein complexes containing interfacial water molecules. In the second section, I studied the mobility and energetics of interfacial water molecules. (both sections correspond to articles published in the Proteins Journal).

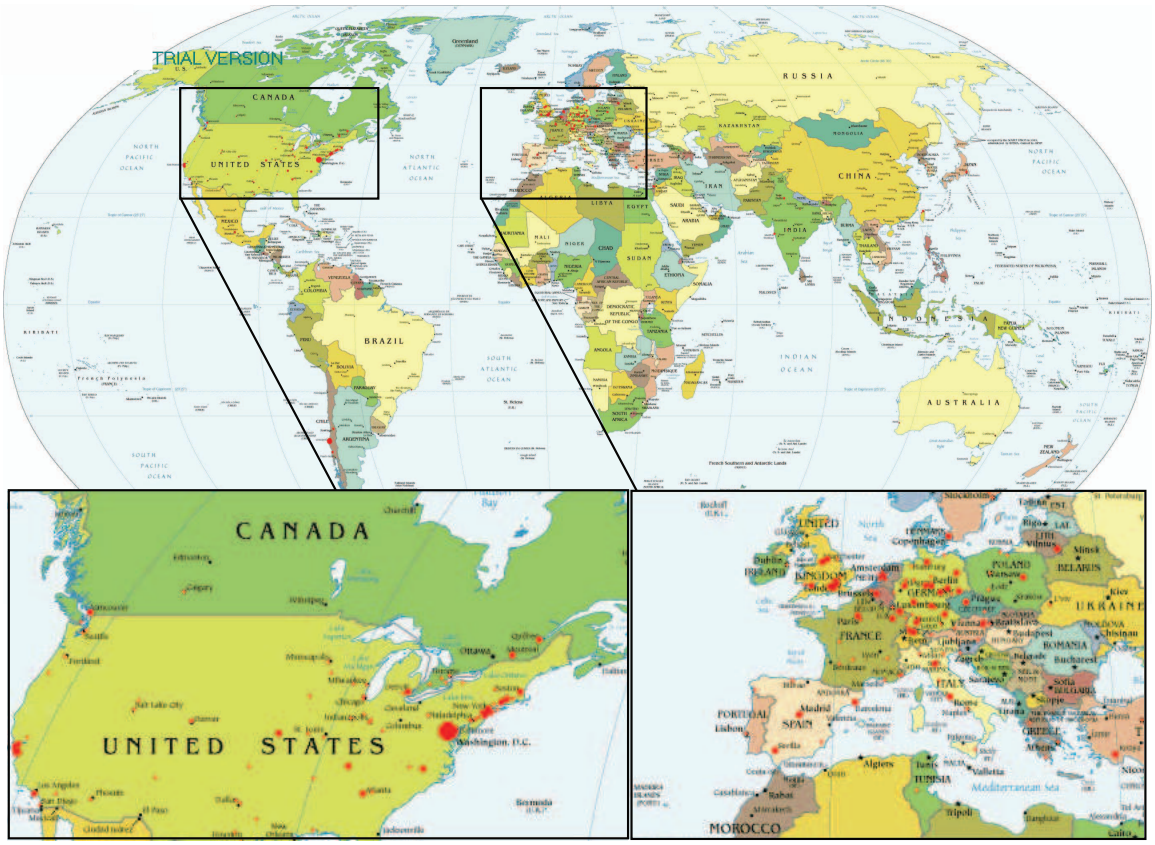
- Finally, chapter V outlines the main achievements of my thesis work and relates them to the most recent research performed in the area. Future directions and possible implications of my results are also discussed.

Chapters 2, 3 and 4 are a compendium of published research articles in peer reviewed international journals. They all follow an standard research article structure: Abstract, Introduction, Methodology, Results and Discussion and Conclusions. The bibliography is collected at the end of the thesis to avoid redundancy among chapters.

Chapter 2

THE SCOWLP DATABASE

All available 3D structural information of proteins at atomic resolution is stored in the Protein Data Bank (PDB), and its growth is exponential with no sign of falling off. For this reason, there are already many databases that obtain the interacting information from the PDB, although most of them miss fundamental components for the interface definition and a classification of family binding regions. In this chapter, I describe the SCOWLP platform, a database and web application that aims to improve the other existing databases of protein interfaces by i) the inclusion of peptidic ligands, interfacial water molecules and more accurate interaction rules, and ii) the classification of protein binding regions by families. SCOWLP is presented by means of two peer-reviewed research articles published during my PhD in the BMC Bioinformatics Journal.



[1]The image in the cover illustrates the 'worldwide' usage of the SCOWLP web-application. The red dots represent the IP location of the user and the size of the dot the amount of connections. The USA and Europa are zoomed for better appreciation. The image was generated using the Alien IP program. A total of 2,426 people visited the www.scowlp.org during the first two years, where 661 were unique IPs. The users visited 966 domain families (restricting the visit of one IP to a family once). A total of 424 domain families are visited from the 1,250 existing ones. The most dominant ones are SH3, SH2 and PDZ domains.

2.1 SCOWLP: a web-based database for detailed characterization and visualization of protein interfaces

by Joan Teyra, Andreas Doms, Michael Schröder and M. Teresa Pisabarro
in BMC Bioinformatics, 2006, 7:104

2.1.1 Abstract

Currently there is a strong need for methods that help to obtain an accurate description of protein interfaces in order to be able to understand the principles that govern molecular recognition and protein function. Many of the recent efforts to computationally identify and characterize protein networks extract protein interaction information at atomic resolution from the PDB. However, they pay none or little attention to small protein ligands and solvent. They are key components and mediators of protein interactions and fundamental for a complete description of protein interfaces. Interactome profiling requires the development of computational tools to extract and analyze protein-protein, protein-ligand and detailed solvent interaction information from the PDB in an automatic and comparative fashion. Adding this information to the existing one on protein-protein interactions will allow us to better understand protein interaction networks and protein function. SCOWLP (Structural Characterization Of Water, Ligands and Proteins) is a user-friendly and publicly accessible web-based relational database for detailed characterization and visualization of the PDB protein interfaces. The SCOWLP database includes proteins, peptidic-ligands and interface water molecules as descriptors of protein interfaces. It contains currently 74,907 protein interfaces and 2,093,976 residue-residue interactions formed by 60,664 structural units (protein domains and peptidic-ligands) and their interacting solvent.

The SCOWLP web-server allows detailed structural analysis and comparisons of protein interfaces at atomic level by text query of PDB codes and/or by navigating a SCOP-based tree. It includes a visualization tool to interactively display the interfaces and label interacting residues and interface solvent by atomic physicochemical properties. SCOWLP is automatically updated with every SCOP release. SCOWLP enriches substantially the description of protein interfaces by adding detailed interface information of peptidic-ligands and solvent to the existing protein-protein interaction databases. SCOWLP may be of

interest to many structural bioinformaticians. It provides a platform for automatic global mapping of protein interfaces at atomic level, representing a useful tool for classification of protein interfaces, protein binding comparative studies, reconstruction of protein complexes and understanding protein networks. The web-server with the database and its additional summary tables used for our analysis are available at http://www.scowlp.org

2.1.2 Introduction

One of the most interesting and important challenges in the so-called "Post-genomic Era" is the understanding of protein networks. Protein-protein interactions have been extensively investigated using a variety of methods [34], and many databases have been built becoming very helpful tools for the analysis of protein networks [35–37].

Protein interfaces have long been studied at protein chain and domain interface levels [38–46]. Furthermore, numerous analyses have used datasets of protein chain interfaces to investigate residue type propensities, sequence and structure conservation at protein interfaces [41, 44, 46–49]. Databases containing structural domain-domain interactions have also been recently created: 3did [50], PiBase [51], iPfam [52], PSIbase [53], InterPare [54], PRISM [55]. However, in these methods still many protein residues are not taken into account as "interfacial" or "interacting" because of peptidic-ligands and also solvent being frequently ignored from the protein interaction analysis.

Peptidic-ligands and solvent mediate protein interactions and are fundamental components for a complete description of protein interfaces. Proteins can interact with peptides to perform their biological function. Besides, peptides have been used to mimic protein binding interfaces, and their complexes with proteins have been used to study protein binding affinity/specificity properties in a simplified way [56–58]. For these reasons, many protein-peptide complexes have been experimentally studied by X-ray crystallography and/or NMR studies, providing additional information on protein interfaces [58]. Moreover, protein interactions take place in an aqueous solution. Solvent molecules can bridge binding partners via hydrogen bonds contributing significantly to molecular recognition and function [56, 59–64].

Most current methods do not provide an accurate description of protein interfaces, which is required to be able to establish the bases for understanding the principles that govern

molecular recognition and protein function.

Here we present SCOWLP (Structural Characterization Of Water, Ligands and Proteins), a platform for complete and detailed characterization and visualization of protein interfaces. Our database includes all protein-interacting components of the PDB including peptides and solvent, which until now have been excluded from systematic protein interface analysis and databases. In our database all interface interactions are described at atom, residue and domain level by using interacting rules based on atomic physicochemical criteria. This complete characterization makes SCOWLP useful for comparative structural analysis of molecular interfaces. The web application allows the user to access all the atomic interaction information by querying the PDB or the SCOP hierarchy. All interface information characterized by different interaction descriptors can be interactively visualized by using a Jmol 3D applet [65].

2.1.3 Methodology

SCOWLP is a web-based relational database formed by eleven tables describing PDB interface interactions at atom, residue and domain level. The database contains 74,907 protein interfaces and 2,093,976 residue-residue interactions formed by 60,664 structural units and interacting solvent. For the creation of the SCOWLP, we extract 3D data of protein domains, peptidic-ligands and interface solvent from the PDB [10], and we define protein domains from the SCOP 1.69 [66]. We compute protein interactions at atom, residue and domain level by using bounding shape-based algorithms [67]. We also have developed a web application to handle and navigate through the interfacial data in an automatic and user-friendly fashion. We designed the SCOWLP methodology based on the following steps:

SCOL-Ligand (Structural Characterization Of Peptidic-Ligands)

The first step of our methodology consists of creating the SCOL table. Each structural unit in a PDB file is represented by a different chain name. We extract all structural units of the PDB and compare them with the domain definitions of SCOP. Although SCOP has a "Peptide" class containing functional peptides, it does not contain all peptidic-ligands complexed in the PDB. For this reason, structural units bigger than two and smaller than

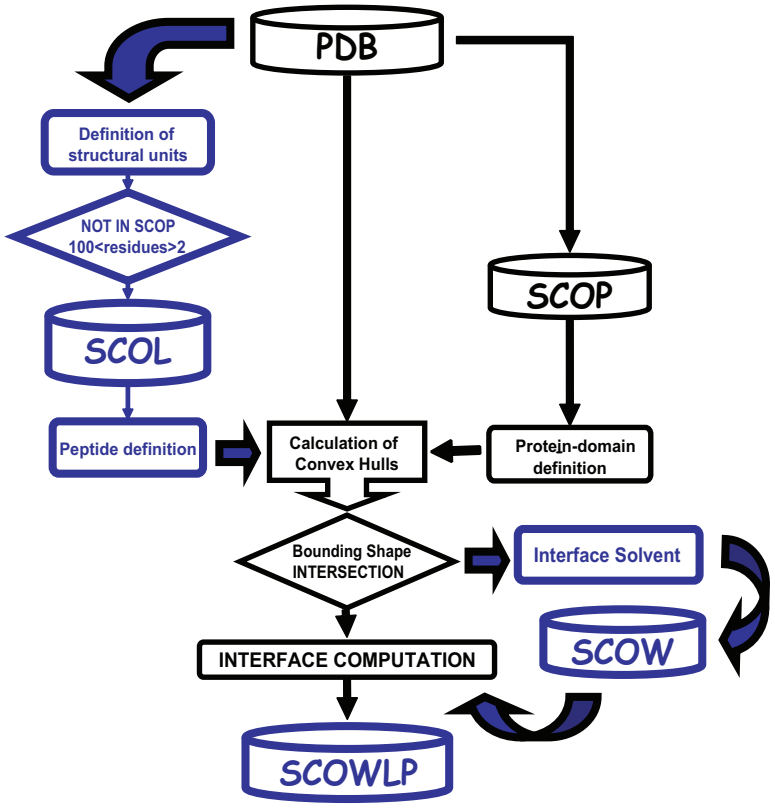


Figure 2.1: Schematic overview of the methodology. SCOWLP uses information from PDB, SCOP, SCOL and SCOW for the computation of atomic interface interactions.

one hundred residues not defined in SCOP are considered peptidic-ligands. We stored this information in the SCOL table (Fig 2.1). Heteroatoms and modified residues that form part of the same polypeptide chain are included, and DNA residues are excluded. We characterize each SCOL peptidic-ligand by resolution, sequence length and secondary structure. SCOWLP contains 2,739 peptidic-ligands, which add 3,413 new interfaces (Fig 2.2).

Interacting structural unit pairs

We label all structural units of the PDB with the SCOL-peptide and the SCOP-domain definitions in order to compute their interactions. We consider a contact distance cut-off of 9 $\AA$ between two residues in order to allow up to two bridging water molecules in the shortest axes defining the interface. We use bounding shape-based algorithms to compute a

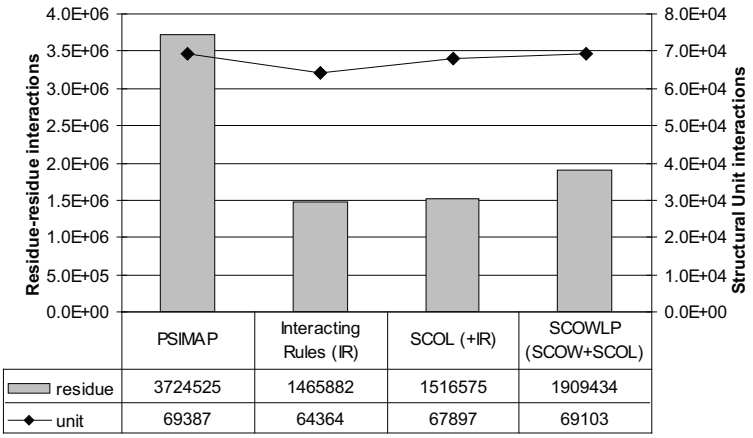


Figure 2.2: Comparative histogram of SCOWLP vs. PSIMAP database. Representation of the number of residue-residue (left y axis) and structural unit interactions (right y axis) contained in SCOWLP and comparison with PSIMAP.

9 $\mathring{A}$ convex hull (the smallest convex set containing all atoms at 9 $\mathring{A}$) for each structural unit of each PDB entry. Convex hull algorithms have been proved to reduce the computational time required for an interface calculation by both, reducing the search space to decrease the number of residues checked for the calculation and allowing distributed computations [67]. Structural units with intersecting shapes and having at least one residue-residue interaction are considered interacting pairs (Fig 2.1).

SCOW-Water (Structural Characterization Of Water)

We consider a water molecule as part of an interface when it is located in the shape intersection of two interacting structural units. All interface water molecules are stored in the SCOW table and are then included in the atomic interface computation. We also consider an interaction when two residues are bridging through one or two water molecules. Residue contacts are defined as only water-mediated (OWM), non water-mediated or direct (D), and mixed (M). Residues that only interact through water are defined as wet spots (Fig 2.3). SCOWLP contains 435,086 new water-mediated interactions thanks to the implementation of SCOWL. This represents 20% of the SCOWLP database (Fig 2.2).

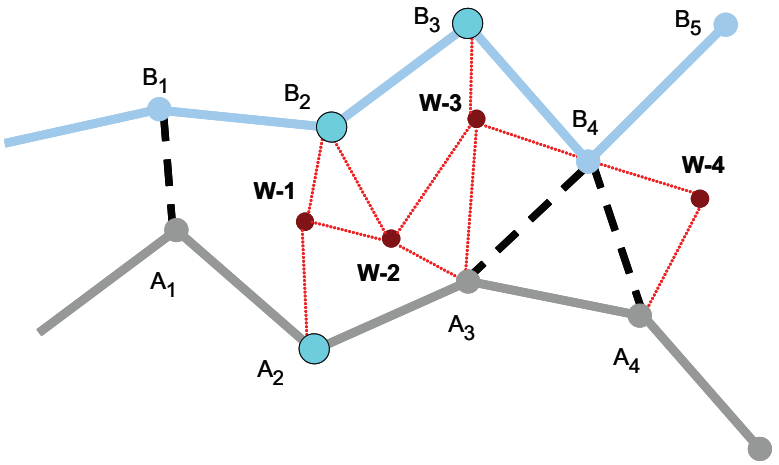


Figure 2.3: Schematic representation of the interface interaction of two molecules and definition of wet spots. Molecules A and B form an interface. Interacting residues and water molecules are represented as black and open circles, respectively.

Interaction rules for interface computation

Only amino acid residues and water molecules placed in the intersection of structural unit shapes are potential interactors. We apply atom type and distance criteria to compute interactions between structural unit pairs at physicochemical level. For hydrogen bonds we apply a ≤ 3.2 $\AA$ donor-acceptor distance. For salt bridges, we apply a ≤ 4 $\AA$ distance criteria. Van der Waals energies are defined by hydrophobic atoms at van der Waals radii distance. At atomic level, we characterize the interactions by: i) nature: hydrophilic, hydrophobic; ii) contact type: main chain, side chain, mixed; iii) number of bridging water molecules. At residue level, we characterize the interactions by: i) nature: hydrophilic, hydrophobic, dual; ii) contact type: main chain, side chain, mixed; iii) number of bridging water molecules; iv) total number of atoms contacting. At structural unit level, we characterize the interactions by: i) contact volume; ii) surface area from convex hull surface; ii) number of interacting atoms/residues per unit; iv) type of interaction: intra-/inter-molecular. All interfacial interaction information is stored in the SCOWLP database (Fig 2.1).

Summary tables

We have created the following additional tables for the filtering and comparative analysis of the information contained in the database:

- Interface description; This table summarizes all interfaces of the SCOWLP database. It contains 74,907 interfaces constituted by SCOP domains labelled with the attributes: PDB Id code, atomic resolution, contact type (intra-/inter-molecular) and SCOP Id code. All interfaces are also labelled by number of interactions (total, all water-mediated and only water-mediated) and number of interacting residues per binding partner. Each interaction is classified by type (side-/main-chain or both) and by number of bridging water molecules.

- Wet interfaces selection; This table stores interfaces of complexes at resolution ≤ 2.5 $\mathring{A}$ from the Interface description table for interfacial solvent analysis. This table does not include homodimer interfaces because of their patchy, poorly packed and highly hydrated nature [68]. With the resultant dataset, we create three tables:

  - Content; This table can be used to rank superfamilies based on their content in water mediating interface interactions. For each interface, it contains the average of total interactions, all water-mediated interactions and the ratio from the percentage of water-mediated interactions at superfamily level.
  - Morphology; This table can be used to rank the interfaces by number of wet spots. In this table each family is represented by the complex with the highest number of wet spots, labelled with the total number of interacting residues and wet spots.
  - Comparative; This table can be used to monitor solvent variations in interfaces and compare them at family level. It contains interfaces sorted out by domain, and then by their respective ligands (protein or peptide). Because a protein-ligand interface can be found in different PDBs, we select the interfaces that appear more than once and contain wet spots. When the same PDB file contains

a repeated interface of two binding partners, we select as a representative the one with more wet spots.

Implementation

We used MySQL and the Java programming language to generate and analyze the SCOWLP database. Interface calculations are performed on a 2.6 GHz Pentium IV in approximately 36 hours. SCOWLP is automatically updated with every SCOP release.

2.1.4 Results and discussion

SCOWLP database contains detailed information of protein interfaces including peptidic-ligands and solvent in the PDBs, and classifies protein interfaces by using specific physicochemical atomic criteria. The database can be accessed through a user-friendly web application.

Interaction rules

The use of atom type and distance rules allows us to characterize and classify interactions at physicochemical level. Other existing methods adopt exclusively a general distance criterion. PSIMAP [67], for example, considers as an interacting pair any atom distance at ≤ 5 $\mathring{A}$. For this reason, the total number of residue-residue and structural unit interactions we obtain by applying our interaction rules is reduced in comparison to PSIMAP (Fig 2.2). This reduction translates into more accurate interface definitions.

Peptidic-ligand contribution

Some proteins have been subject of many structural studies complexed with peptides (e.g. Proteases, b.47.1). Besides, the superfamilies that have the higher occurrence of peptides are not necessarily those with higher domain-domain representation (e.g. Cyclophilin, b.62.1). By taking into account information about protein-peptide complexes SCOWLP contributes interfacial information of 8 SCOP superfamilies uniquely represented by protein-peptide complexes (a.23.4, a.50.1, d.76.1, a.8.5, d195.1, g.33.1, a.144.1, a.12.1). In addition, it contributes with more than 50% of the interacting information in other superfamilies. Our

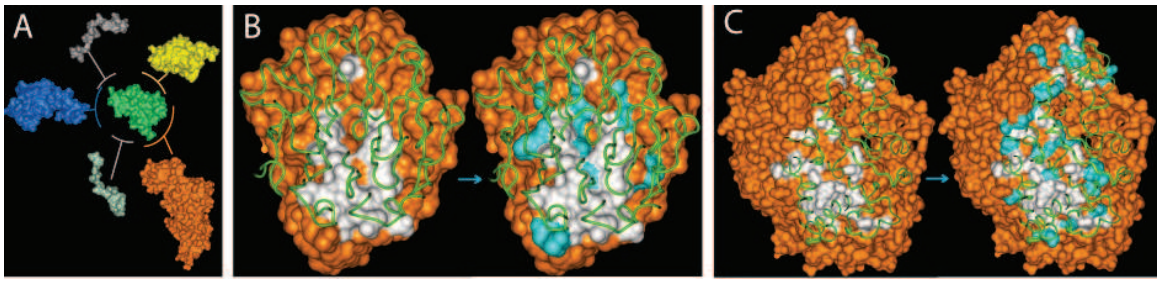


Figure 2.4: Enrichment of the interface definitions by peptidic-ligands and solvent. A) Enrichment in the description of protein interfaces by peptidic-ligands. The molecular recognition features of the BTB/POZ domain family are summarized. A representative POZ domain

results show the importance of including protein-peptide interfacial information in order to enrich considerably the description of protein interfaces. Proteins can bind to peptides in places that do not exactly correspond to binding sites in their known protein-protein complexes. As an example, we show the BTB/POZ (Poxvirus and Zinc finger) family. The twelve BTB/POZ complexes in the PDB present five domain-binding regions, two of them described by the protein-peptide complexes (Fig 2.4.A). The POZ-peptide interfacial information is functionally relevant. It may help to propose new POZ contacts when reconstructing multi-protein complexes and modelling signalling pathways where the POZ domain-containing proteins are involved. Our results show that the addition of peptidic information can help to complete the view on how a protein recognizes its binding partners.

Solvent contribution

All superfamiles of the Content table contain solvent mediating interactions. Furthermore, in some of these superfamilies water-mediated interactions represent up to 75% of the total interfacial interactions (e.g. d.250.1). Relating to the "only water-mediated" interactions, we observe from the Morphology table that 43 is the maximum number of wet spots found. Figures 2.4.B and 2.4.C illustrate how solvent, in particular wet spots, may play an important role in the morphological description of protein interfaces (shape and size). Considering the solvent, a discontinuous surface formed by several small isolated patches changes to a bigger and rounded patch. These observations show that we can enrich the description of

protein interfaces by considering interfacial solvent.

Although solvent molecules mediating protein interactions can be conserved in a protein family, variations may occur due to different facts: i) atomic resolution and/or quality of the structural data, ii) conformational changes upon ligand binding, iii) protein flexibility, iv) new interacting regions (e.g. loop insertions and deletions), v) residue mimicry. Wet spots variations may be used as indicators in these cases. The Comparative table allows us to compare the interfaces of 127 families in 751 complexes based on wet spots variations.

Solvent molecules play an important role in the replacement of residues in protein interfaces. Sometimes the atomic resolution, the existence of different rotamers or even small differences in contact distances defining the interaction may influence the number of wet spots. Nevertheless, small variations of wet spots in complexes of the same family that do not present changes in total number of interactions can be used to locate residue mimicry cases (e.g. Lys+H2O≈Arg). Making use of this information may be very useful in analysis of protein interfacial evolution and in protein engineering/rational design when designing affinity and specificity of a protein for its ligands.

Web application

SCOWLP contains atomic interfacial information of all the PDB entries structured by the SCOP hierarchy. There are two ways to query our database: SCOP or PDB. The user can query SCOP by keywords, SCOP/PDB Ids, or by simply navigating the SCOP hierarchical tree (Fig 2.5.1). When the user selects a family from the tree (labelled as FA), SCOWLP retrieves a list of the PDBs containing interfaces of that family in one frame. A second frame shows a summary table listing all the interfaces of that family with PDB id, type of contact, superfamily description of binding partners, interfacial area, total interacting residues and number of wet spots. This summary table gives a good overview over the interacting partners and interfacial variations at family level. By selecting any of the PDB IDs in this table, the user retrieves a list of all the interfaces of that PDB organized in two interactive tables: Interfaces and Interactions. We obtain the same tables querying SCOWLP by PDB ID (Fig 2.5.2). The "Interfaces" table shows binding partners, interfacial area, total number of interfacial residues and wet spots. The Interaction Types table classifies the interactions

based on their water mediation, nature and type. The user can select the interfaces in a master/slave way to display a 3D molecular viewer and the selected domain contacts. We have implemented Jmol scripts [65] to allow the user to display and interactively analyze interfaces by using two control panels (Fig 2.5.3). The first one (on the right; Fig 2.5.3a; Domain Contact Selection) controls the interface display in the 3D viewer, allowing the user to highlight the residues forming part of each interface. The second panel (bottom left; Fig 2.5.3c) controls: Molecule View: ON/OFF residue labelling, water mediators and spinning; Interacting Descriptions: interfacial residues colouring based on wet spots, nature and type. Fig 2.5.3 shows a protein domain (red) interacting with a peptidic-ligand (yellow) and their respective interacting residues (wet spots in blue).

By using SCOWLP, the user can achieve specific queries, SCOP family analysis, interface comparisons and a detailed 3D display of the atomic interaction data contained in PDBs.

2.1.5 Conclusion

Detailed analysis of the interfacial information contained in the PDB is very useful to obtain more accurate descriptions of protein interfaces. We have created SCOWLP to have a platform for the characterization and 3D visualization of protein interfaces. SCOWLP enlarges the available information on protein-protein interactions by introducing 3,413 new protein-peptide interfaces and 435,086 additional water-mediated interactions. All interactions contained in SCOWLP are characterized and classified at physicochemical level instead of using general distance criteria. This allows a more appropriate definition and enhanced comparison of the interfaces contained in our database.

As the origin of specificity and affinity in molecular recognition can be partially explained in terms of solvent's contribution to the interaction, our database constitutes a very useful tool to facilitate rational ligand design. In particular wet spots can be used as indicators of interfacial solvent variations, being helpful in comparison of protein family interfaces, and perhaps guiding docking experiments.

SCOWLP may be of interest to many structural bioinformaticians, representing a useful tool for classification of protein interfaces, protein binding comparative studies, reconstruction of protein complexes and understanding protein networks.

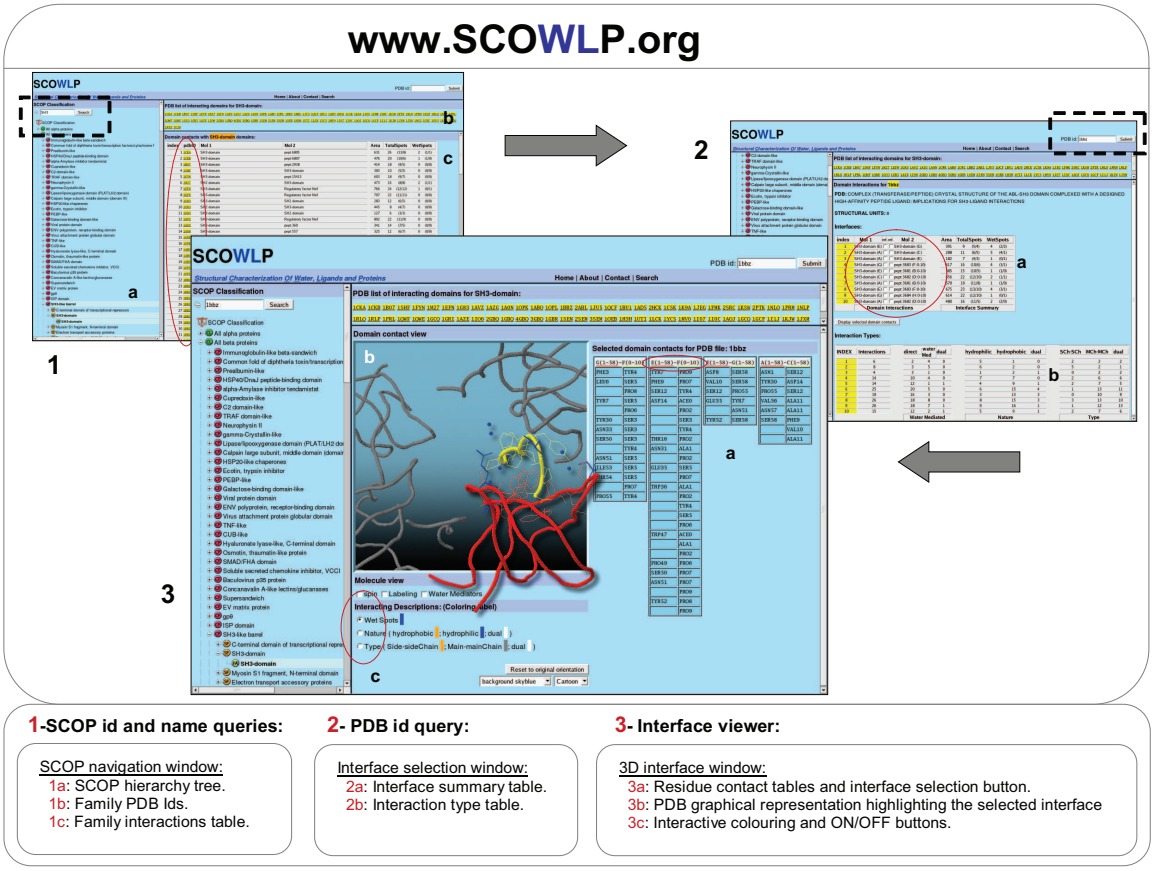


Figure 2.5: Screenshots and legends showing the structure of the SCOWLP website.

2.2 SCOWLP classification: Structural comparison and analysis of protein binding regions

by Joan Teyra, Maciej Paszkowski-Rogacz, Gerd Anders and M. Teresa Pisabarro
in BMC Bioinformatics, 2008, 9:9

2.2.1 Abstract

Detailed information about protein interactions is critical for our understanding of the principles governing protein recognition mechanisms. The structures of many proteins have been experimentally determined in complex with different ligands bound either in the same or different binding regions. Thus, the structural interactome requires the development of tools to classify protein binding regions. A proper classification may provide a general view of the regions that a protein uses to bind others and also facilitate a detailed comparative analysis of the interacting information for specific protein binding regions at atomic level. Such classification might be of potential use for deciphering protein interaction networks, understanding protein function, rational engineering and design. Protein binding regions (PBRs) might be ideally described as well-defined separated regions that share no interacting residues one another. However, PBRs are often irregular, discontinuous and can share a wide range of interacting residues among them. The criteria to define an individual binding region can be often arbitrary and may differ from other binding regions within a protein family. Therefore, the rational behind protein interface classification should aim to fulfil the requirements of the analysis to be performed.

We extract detailed interaction information of protein domains, peptides and interfacial solvent from the SCOWLP database and we classify the PBRs of each domain family. For this purpose, we define a similarity index based on the overlapping of interacting residues mapped in pair-wise structural alignments. We perform our classification with agglomerative hierarchical clustering using the complete-linkage method. Our classification is calculated at different similarity cut-offs to allow flexibility in the analysis of PBRs, feature especially interesting for those protein families with controversial binding regions.

The hierarchical classification of PBRs is implemented into the SCOWLP database and extends the SCOP classification with three additional family sub-levels: Binding Region,

Interface and Contacting Domains. SCOWLP contains 9,334 binding regions distributed within 2,561 families. In 65% of the cases we observe families containing more than one binding region. Besides, 22% of the regions are forming complex with more than one different protein family. The current SCOWLP classification and its web application represent a framework for the study of protein interfaces and comparative analysis of protein family binding regions. This comparison can be performed at atomic level and allows the user to study interactome conservation and variability. The new SCOWLP classification may be of great utility for reconstruction of protein complexes, understanding protein networks and ligand design. SCOWLP will be updated with every SCOP release. The web application is available at http://www.scowlp.org

2.2.2 Introduction

Protein interactions are essential for intra-cellular communication in biological processes. Proteins are composed of small units or domains that can physically interact together forming multi-domain protein complexes. A single protein can have several binding regions, and each region can engage distinct ligands, either simultaneously or at successive stages of signalling [69].

In our previous work we developed the SCOWLP database [70], which contains detailed interfacial information of structurally known protein complexes, peptide complexes and water molecules as mediators of interactions. SCOWLP and other existing protein interaction databases [50, 51, 71] contain lists of interfaces for SCOP protein families and, therefore, they are only able to perform individual interface analysis. A classification of protein binding regions (PBRs) is essential in order to characterize all protein regions participating in the binding and to be able to compare protein complexes sharing the same binding region. At the same time, such a classification should provide some insights into the interacting properties preserved by members of a protein family. However, the criteria to delineate PBRs can be difficult to assess, and often arbitrary and controversial.

Binding regions in protein domains can form separated patches, but also some protein families bind through multiple binding regions with different ranges of residue overlapping. Furthermore, some observed protein interfaces are the result of non-biological artefacts (i.e.

crystal packing) and are often difficult to distinguish from the biological ones, creating discrepancy among the current resources [72, 73]. Some of these interfaces can connect binding regions or can be included into existing ones, introducing noise quite difficult to handle for clustering algorithms. As different clustering algorithms can vary the grouping completely, an advantageous classification of PBRs should contain a proper measurement of similarity and a flexible clustering algorithm to cover the requirements of the analysis to be performed.

Hierarchical clustering comprises a whole family of clustering methods differing only on the manner inter-cluster distance is defined (the linkage function). The more common aggregation methods are single-, complete- and average-linkage. Complete and single-linkage are extreme procedures with completely different properties. Complete-linkage uses the similarity between the furthest pair of objects from two clusters. In contrast to these requirements, single-linkage only uses the nearest pair of objects from each cluster. Both methods have an extreme conception of homogeneity of a cluster. Single-linkage leads to grouping and may result in a few large and heterogeneous clusters [74]. Complete-linkage results in dilatation and may produce many clusters, being more suitable for isolating poorly separated clusters [75]. Average-linkage tries to avoid these effects by computing the average. This method is used by two different computational approaches for protein interface classification described so far. Nussinov and colleagues pioneered interface classification based on common structural features shared among the interfaces from various folds and considering full interfaces at chain level [45, 76]. More recently Kim and colleagues [77, 78], instead of classifying interfaces as a whole, classified the domain faces forming an interface by SCOP families. Their classification uses interfaces defined as biological in the PQS database [72] and does not include peptidic and solvent interaction data.

We present a classification of PBRs from all existing contacting domains from the SCOWLP database, which includes detailed information about proteins, peptides and solvent interaction. Peptide and solvent interactions are highly represented in the PDB and are highly informative in protein interactions [79]. For our classification we use hierarchical clustering with the complete-linkage method. The similarity measure used is obtained based on the overlapping of interacting residues mapped in pair-wise structural alignments

and exclusion of gap regions. We explain and discuss the methodology used to classify PBRs and the rational behind applying flexible similarity cut-offs. Our PBRs classification is implemented in SCOWLP and extends its usage from individual analysis of protein interfaces to comparative structural analysis of specific family binding regions. We describe the SCOWLP web application and its utilities for PBRs analysis.

2.2.3 Methodology

The PBRs classification extends the SCOWLP relational database by four additional tables describing the hierarchical classification at binding region, interface and contacting domain level. The content of the classification is given at similarity zero, which offers a general view of the regions that protein families use for recognition. In addition, our classification offers different similarity cut-offs to allow flexibility in the analysis of the PBRs. The classification of PBRs was performed as follows (Fig 2.6.A):

Extraction of interfaces and contacting domains

An accurate definition of the interacting residues is crucial to have a proper clustering of a family PBR. We extracted all protein interfaces from the SCOWLP database, in which the interactions are defined at atomic level and based on their physiochemical properties [70, 79]. Protein domains interacting with peptidic ligands and residues interacting through a water molecule (wet spots) are also taken into account. We consider "interface" all domain-domain interactions; that means those belonging to the same protein and also to different proteins. SCOWLP contains 79,803 interfaces contained in 2,561 SCOP families. We grouped the domains participating in each interface by SCOP families, obtaining for each family a list of contacting domains with the residues forming part of the binding region.

Pair-wise structural alignments (PSAs)

A reliable alignment is indispensable to calculate the similarities among binding regions. For this purpose we used MAMMOTH, which has shown proved accuracy to structurally align protein families [80]. We performed all-against-all PSAs of the contacting domains for each family to be able to measure the similarity among binding regions. SCOWLP

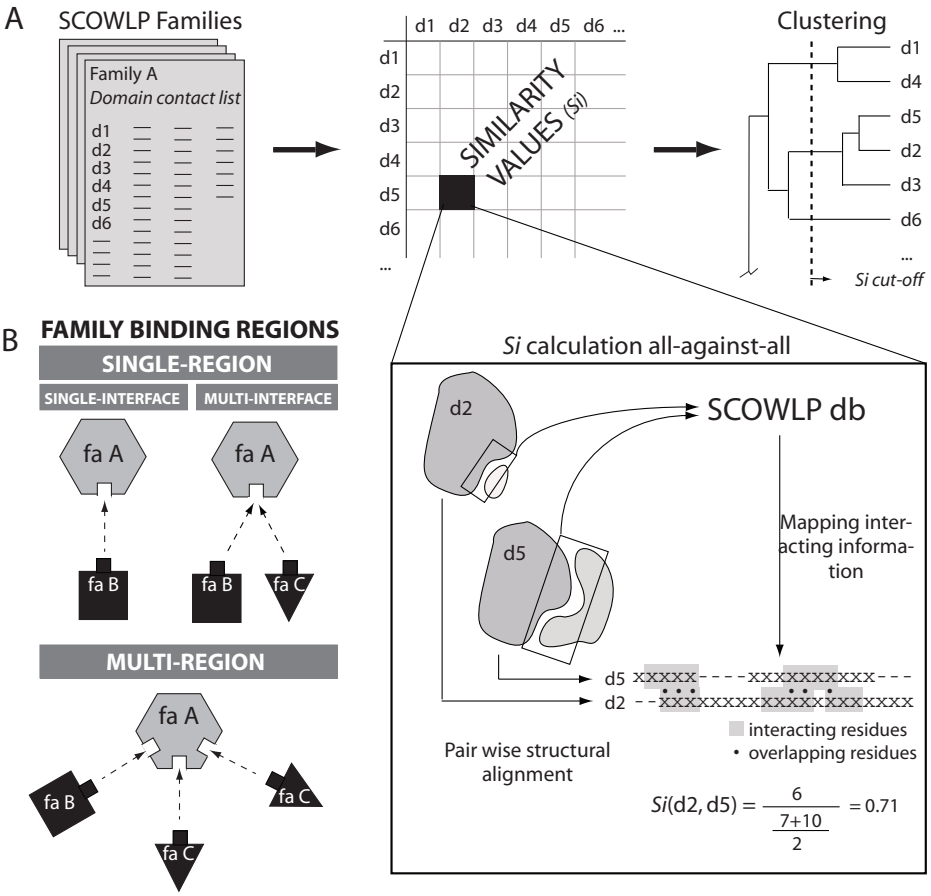


Figure 2.6: Schematic overview of the methodology. A) The contacting domains per family are extracted from the SCOWLP database. The Si are calculated for all-against-all contacting domains and used for clustering. Results are displayed in a dendrogram. B) Classification of protein binding regions (PBRs). A protein family can recognize other proteins and ligands by single- or multi-binding regions, and for each binding region single- or multi-interfaces may exist depending on the number of partners that have been structurally observed interacting within a specific binding region.

contains about 160,000 contacting domains uneven distributed by families. This represents 276 million PSAs performed in a cluster of five Pentium IV 2.6 GHz. The alignments were performed taking the Cα atoms into account and using a gap penalty function for opening and extension [81]. The root-mean-squared deviation (RMSD) was not considered for measuring the similarity between two interfaces, as the superimposed members of the same family share a common structure.

Similarity Index (Si)

The residues described in SCOWLP to be forming and interface were mapped onto the domain-pair structural alignment. We calculated a similarity index (Si) based on the number of interacting residues that overlap and the length of both interacting regions by (Fig 2.6):

$$Si(a,b) = \frac{IR_{overlap(a,b)}}{\frac{\left(IR_{length(a)} - IR_{gaps(a)}\right) + \left(IR_{length(b)} - IR_{gaps(b)}\right)}{2}} \tag{2.1}$$

where a and b represent the two domain structures aligned. The number of interacting residues that match in the PSA is represented by $\mathrm{IR}_{overlap(a,b)}$. This value is divided by the average number of the interacting residues in both domains excluding the interacting residues located in gap regions in the structural alignment (IR_{gaps}).

Clustering binding regions

Based on the calculated Si, we clustered the binding regions of each SCOP family using the agglomerative hierarchical algorithm [74] following several steps (Fig 2.6):

1) Define as a cluster each contacting domain.
2) Find the closest pair of clusters and merge them into a single cluster.
3) Re-compute the distances between the new cluster and each of the remaining clusters.
4) Repeat steps 2 and 3 until all contacting domains are clustered into a single cluster.

To re-compute the distances we used the complete-linkage method [75], which considers the distance between two clusters to be equal to the minimum similarity of the two members.

Binding region definition by Si cut-offs

The result of the clustering can be represented in an intuitive tree or dendrogram, which shows how the individual contacting domains are successively merged at greater distances into larger and fewer clusters. The final PBRs depend on the Si cut-off that is set up. We can observe in Fig 2.7.A that the total number of binding regions for all the SCOP families grows exponentially as the Si cut-off increases. Based on our observations of a representative group of families we set up an empirical maximum similarity cut-off value of

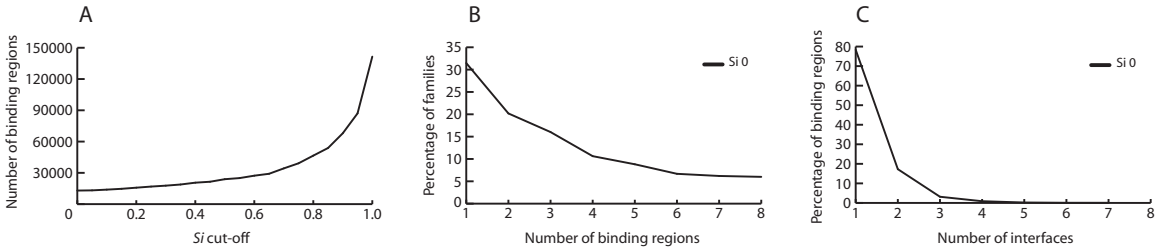


Figure 2.7: PBR analysis. A) Representation of the number of binding regions obtained using different Si cut-offs. B) Representation of the relative percentage of SCOP families with different number of binding regions at zero Si cut-off. C) Representation of the relative percentage of binding regions depending on the number of interfaces at zero Si cut-off. (x-axis in B and C are limited to 8 for simplicity).

0.4. We pre-calculated the results for Si cut-offs at 0, 0.1, 0.2, 0.3 and 0.4 to offer a range of values that allow flexibility in the final analysis of PBRs. The SCOWLP web application offers the possibility to display the classification at any of these cut-off values.

Our classification clustered 160,000 contacting domains from 2,561 families in 9,334 binding regions. About 65% of the families contain more than one binding region (Fig 2.7.B). These values are obtained for similarity zero and may vary depending on the similarity cut-off applied.

Interface definitions

In order to differentiate binding regions having single-interfaces from multi-interfaces, we identified in each binding region the partner for each contacting domain (Fig 2.6.B). Each binding region was divided in sub-clusters when there were different domain families interacting in the same binding region. This resulted in a total of 10,300 interfaces. The classification shows a 78% of the binding regions having a single-interface and the rest having mainly 2 or 3 interfaces per region (Fig 2.6.C). These numbers have to be carefully interpreted by taking into account the limitation of the structural information contained in the PDB (i.e. 1,715 binding regions contain a unique member in the PDB and therefore only one known interface per binding region).

Implementation

We used MySQL and Java programming language to generate the classification of PBRs. Calculations were performed on a cluster of five Pentium IV 2.6 GHz. The PBRs classification has been included into the SCOWLP database. SCOWLP will be updated with every SCOP release.

2.2.4 Results and discussion

In this section we first discuss the methodology used for the classification of PBRs. Besides, we describe the utility of the SCOWLP web application.

Extraction of similarities

The classification of PBRs requires a proper definition of the similarity between binding regions. For this purpose it is essential to have (a) a reliable source of interface definitions, (b) high quality alignments, and (c) a adequate similarity function:

(a) *Interface definitions.*

Our work includes detailed atomic interfacial information from the SCOWLP database, which comprises protein-protein complexes, protein-peptides complexes and solvent-mediated interactions [70, 79]. The contained interacting information at physico-chemical level is very useful to study and compare conservation/variability among complexes even at low sequence similarity.

(b) *Domain PSAs of a family are computationally efficient and give reliable Si.*

The two partners forming an interface do not have to be aligned in order to extract a Si. For each interface, only the partner belonging to the family in study is structurally aligned with the rest of the members of the family. This procedure has two clear advantages: 1) an increased computational speed for each PSA as we overlook one of the partners, reducing the amount of residues to align; 2) the good quality of the family domain alignments, as family domains are structurally conserved. Protein binding regions are often irregular, discontinuous and difficult to compare. Therefore, a good alignment is critical to calculate the range of overlap between two regions.

Our classification method is exclusively based on structural alignments, which makes the methodology computationally expensive but gives better accuracy than sequence alignments at family level.

(c) *The similarity index penalizes gap regions.*

The Si reflects the overlap of interacting residues between two binding regions in domains belonging to the same protein family. It is important to consider the number of interacting residues per domain, which allows us to obtain the percentage of interacting residues that is overlapping over the total (see Methods). This helps to distinguish whether a binding region is identical, different or included into another one.

Ligands and proteins possess internal degrees of freedom and can adopt various conformational states. Furthermore, many family members often contain sequence inclusions/deletions in loops or C-/N-termi, or even additional secondary structure elements, which are often involved in protein interactions. For these reasons, we calculate the Si without considering the interacting residues belonging to gap regions in the PSA. This is graphically illustrated in Figure 2.8.A. Two proteins belonging to the same family differ in an insertion of 55 residues, which creates a gap region in the PSA. This additional region is involved in binding and, therefore, increases the number of interacting residues for the protein containing it. In general, dismissing interacting residues belonging to gap regions in PSAs produces a condensation effect on the clusters at high level of similarity. Additionally, it can also cause reorganization of cluster members at lower levels of similarity. Ignoring gaps for Si calculation and applying flexible similarity cut-offs might help in the final clustering and consequent analysis. This is illustrated in Figure 2.8.B, where two contacting domains of the same family presenting two different overlapping binding regions (peptide-binding and crystal packing) are clustered differently depending on considering or excluding gap regions and by applying flexible Si cut-offs. As an example, applying a 0.2 cut-off when excluding gaps clusters all peptide-binding interfaces separated from the crystal packing interface. This clustering may facilitate further analysis of these different binding regions and their properties.

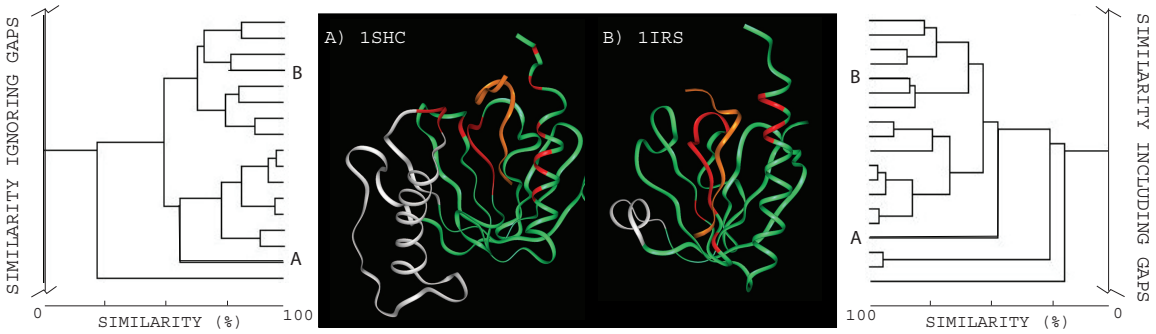


Figure 2.8: Effects of gap regions for Si calculation and clustering. A) Two structures of the PTB domain differing in an insertion/deletion are displayed as green ribbons, and their respective ligands in orange (PDB entry codes: 1SHC and 1IRS). Interacting residues are coloured in red. The insertion/deletion is shown in white. Note that some of the red regions may be included in white ones. B) A section of the dendrogram obtained from the clustering of the PTB domain binding regions (more detailed in figure 2.10) is shown for both cases, excluding and including gap regions. Two members of these clusters presenting different (peptide-binding, 1NMB:AB and crystal packing, 1QQG:AB) but overlapping binding regions are highlighted in blue and pink, respectively. A Si cut-off of 0.2 (dashed line) is shown for comparison.

Aggregation using the complete-linkage method

Some protein families bind through multiple binding regions with different ranges of residue overlapping. This produces extensions of the binding region definitions and association of two clearly defined regions by a third into a bigger single one. To cope with these usual situations, instead of using the average-linkage used by other authors [76, 78], we have rather applied the complete-linkage [75] due to two main properties:

- *Property 1: Complete-linkage is sensitive to zero similarity.* This method defines at similarity zero all binding regions that do not share interacting residues. Besides, it also assumes that in the same binding region all the members must have some range of similarity among them; otherwise they are split in two separate clusters. This is illustrated in Figure 2.9 (left panel), where a binding region of domain X might appear as a single one due to the overlapping of several interfaces (A to G). The handling of the three "connector interfaces" (C, D, G) will be responsible of the definition of the final clusters at similarity zero. The clustering is decided based on the higher similarity; C is more similar to B than to D and, on the other hand, G is more similar to F than to D. Therefore, the connectors G and C will be part of the cluster EF

and AB respectively, whereas D will belong to a separate cluster. At no similarity, complete-linkage differentiates three binding regions, whereas single-linkage offers only one cluster containing all interfaces. In single-linkage the members having no direct similarity (D, F) are included in the same cluster if there is a "connector interface" (G) having some similarity with both. This enables progressive extensions of a binding region depending on the Si cut-off applied. The average-linkage method would have intermediate properties.

- *Property 2: Complete-linkage expands the differences between clusters.* Complete-linkage always takes the member with less similarity to join clusters. Domain Y in Figure 2.9 (right panel) is an illustrative example of binding regions included into others (EFG included in ABCD). The dendrogram shows how the complete-linkage enlarges the differences between both groups more than the single-linkage. The average-linkage would have intermediate values.

These two properties of the complete-linkage method may be very useful for clustering of PBRs. Figure 2.10 represents a specific example of these properties for all the structurally known binding regions of the PTB (phospho-tyrosine-binding domain) domain (see below).

Threshold values define the final PBRs

The clustering process can be represented by a dendrogram, which shows how the individual objects are successively merged at greater distances into larger and fewer clusters. The branches are proportional in length to the estimated similarity of each binding region with the others. The final clusters depend on the similarity cut-off that is set up.

Binding regions of a family can often present overlapping residues, which makes their definition to be sometimes unclear and arbitrary. Sometimes there is no unique criteria to adopt in order to define clear PBRs and, in these cases, an appropriate classification may depend on user-based considerations. Illustrative examples are: i) being able to distinguish multi-interfaces versus multi-regions (Fig 2.6.B) in a protein family, ii) distinction of domain-domain versus domain-peptide interfaces, and iii) being able to separate and analyze "non-biological" interfaces.

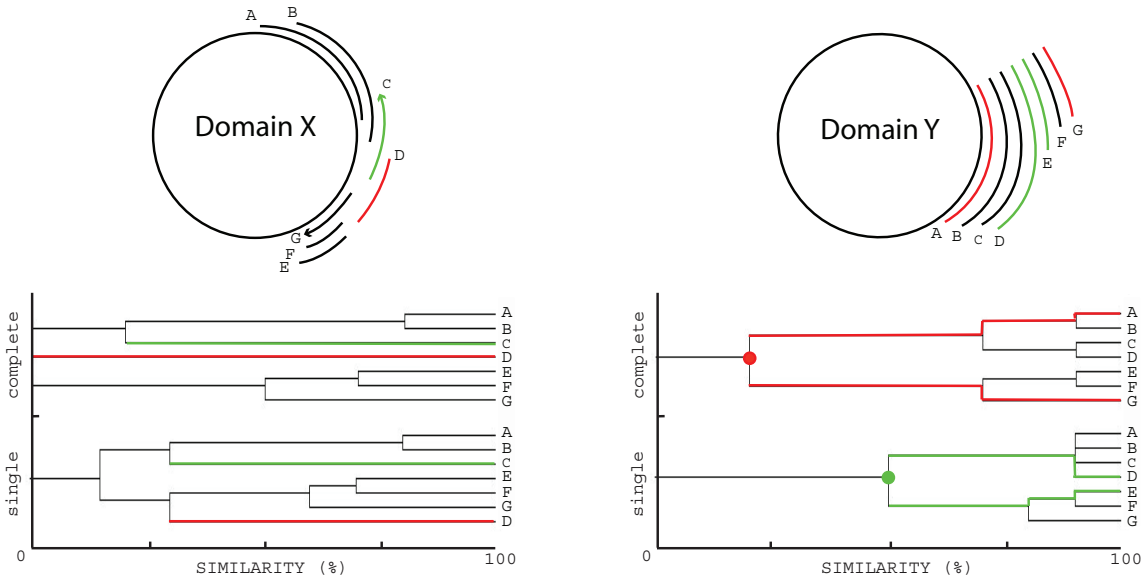


Figure 2.9: Aggregation methods for clustering. Schematic comparison between complete- and single-linkage method properties. Two domains (X and Y) with their respective binding regions (A to G) are schematized. Dendrograms obtained from the clustering of PBRs using complete- and single-linkage methods are shown at the right of each domain scheme for comparison.

This panorama encouraged us to proceed with the application of several cut-offs within an empirical range of similarities by taking advantage of the clustering properties of the complete-linkage method. The minimum Si cut-off value was fixed to zero to give a general view of the binding regions used by a family (property 1). The maximum value was fixed to 0.4 based on our observations (see Si cutoff and Definition section). We also pre-calculated the results for 0.1, 0.2, 0.3 Si cut-offs to allow flexibility in the analysis of PBRs. Figure 2.10 shows all the structurally known binding regions of the PTB domain and the clusters for different Si cut-offs for complete- and average-linkage. It can be appreciated that the slope is not so drastic in complete as it is in the average-linkage method. Although offering a similar grouping of elements, the complete-linkage method produces dilatation of the differences among the elements (property 2) and assists in the application of different cut-offs for separation of clusters. As an example, a cut in a specific point (highlighted in yellow bars) gives a wider similarity range for complete than for average-linkage. The introduced flexibility for choosing cut-offs offers, for example, the possibility to differentiate sub-clusters

(i.e. 2NMB:AB and 1XR0:BA in Figure 2.10) and decide to include or exclude them in a specific binding region for comparative analysis.

Binding regions vs. interfaces clustering

In this section we compare SCOWLP with a different method, PRISM [55], to give insights to users into the utilization of our approach and its biological applications compared to other strategies to classify protein interactions. Whereas SCOWLP compares and classifies interfaces based on defined binding regions in the fold of each counterpart (at family level), PRISM compares full interfaces (both partners) in a sequence position independent manner. By using a geometric hashing algorithm it groups interfaces by similarities of the space distribution of interacting residues independently of the fold. Although being two different approaches, both methods can provide a similar number and composition of clusters for a specific protein family; however, differences may also exist in other cases. The following examples are intended to illustrate it (0.2 similarity cut-off used):

1) If a protein family interacts with two different proteins using the same binding region (Single region-multi-interface; Fig 2.6.B), SCOWLP would always include both interfaces in the same cluster, whereas PRISM would do it only in case it considers similar the distribution of the interfacial residues. This is exemplified in Figure 2.10. SCOWLP includes 1j0w:AB in the same binding region cluster as 1m7e:BA and 1p3r:BA, whereas PRISM classifies 1j0w:AB unaccompanied in an only-one-member interface cluster. The same applies to the case of classification of protein-peptide interfaces, where conformational differences of the short peptidic sequences may cause a different PRISM-architecture and, therefore, a separate classification. SCOWLP groups several peptides binding to the same binding region of the PTB domain in one single cluster of 16 members (Figure 2.10, cluster 1aqc:AC to 1shc:AB); however, PRISM groups these interfaces in two different clusters of six and seven members. For this specific example, the difference in overall numbers of interfaces is due to the fact that some of the protein-peptide interfaces obtained with SCOWLP are missing in the PRISM clustering (1uef:BD, 1m7e:CF and 1oqn:BD).
2) In the case of structural symmetry (i.e. symmetrical protein assemblies and crystal

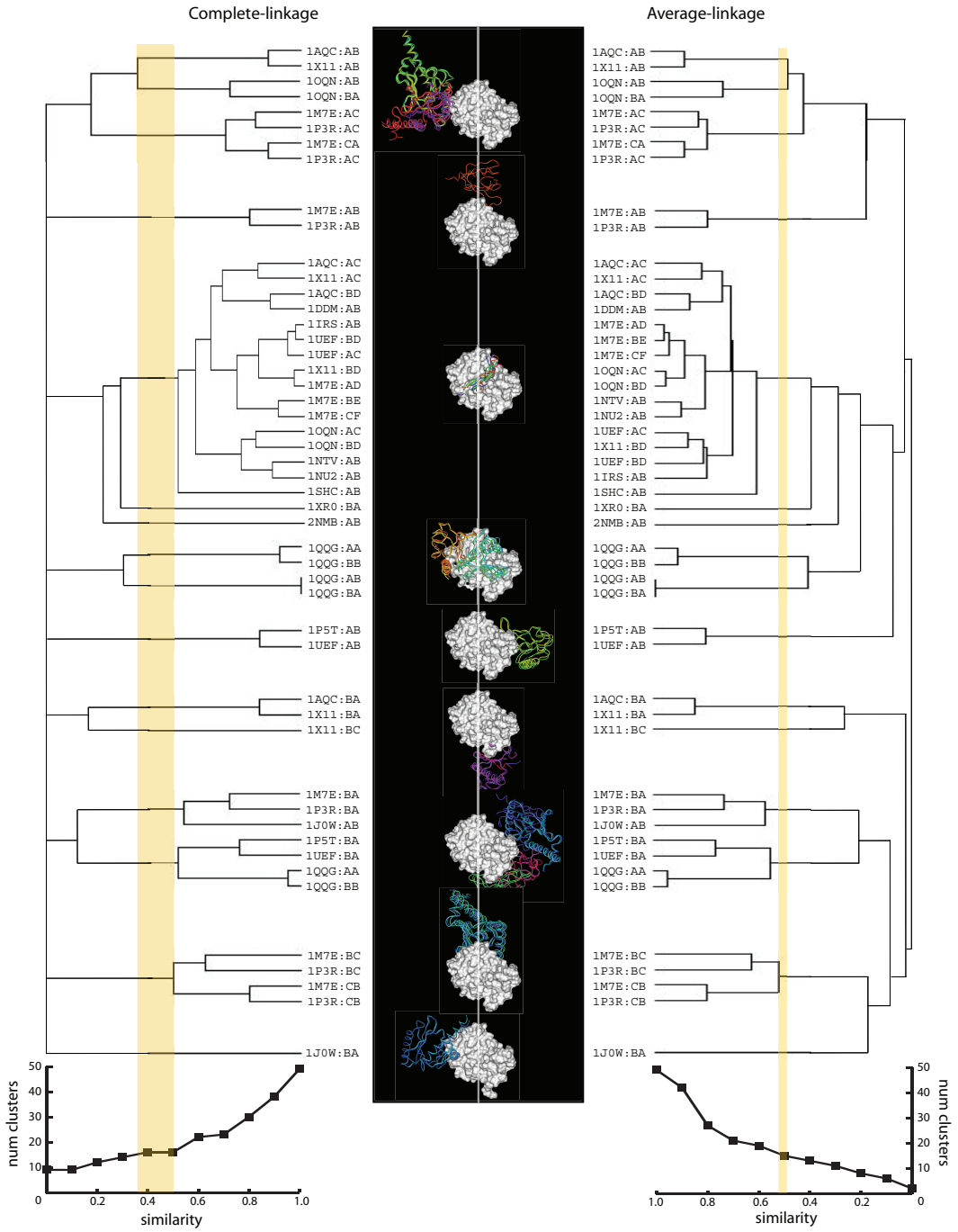


Figure 2.10: Aggregation methods for clustering. Clustering of the PBRs of the PTB domain. The dendrograms derived from the clustering using complete- and average-linkage are shown in the left and right panels, respectively. An example to illustrate the range of similarity that could be used to produce a specific cutting point is represented by the thickness of a yellow line for both methods. The centre of the figure contains the graphical representation of the PTB binding regions obtained using the complete-linkage at zero Si cut-off. The PTB domain is represented as a grey surface, and the corresponding binding partners in coloured ribbons. Two graphs representing the number of clusters at different Si cut-offs are shown at the bottom.

packing), PRISM would include all interfaces in a cluster, whereas SCOWLP would have separated clusters for each binding region.

3) PRISM takes protein chains as a domain unit and therefore does not consider intra-interacting domains, which are considered in SCOWLP.

Web application

We implemented the hierarchical classification of PBRs into the SCOWLP web application. Based on a selected SCOP family, SCOWLP retrieves its binding regions and a summary of the interacting information. The results are generated based on a user-selected similarity cut-off. The analysis of the binding regions can be performed in three different ways (Fig 2.11):

a) visualizing the spatial location of each binding region on a representative family structure by using Jmol plug-in [65],
b) keyword search for PDB ids and chains to identify specific complexes, or
c) visualizing the structure-based aligned representative sequences for a binding region with highlighted interacting residues.

Once the binding region of interest is localized, a tree-based structure shows three additional classification levels (Fig 2.11.d): binding region (BR), interface (IF) and contacting domain (DC). All domains in a family that contain interacting information are structurally aligned and their sequences are displayed. Upon selection, the interacting residues can be coloured based on their physico-chemical properties (hydrophobic, hydrophilic or both), and also by the water contribution to the interfacial interactions (dry, wet or dual interaction). A label with the interacting correspondences will appear on each interacting residue when pointed with the mouse. The physico-chemical properties allow the user to distinguish conserved vs. variable interactions. In Figure 2.11, the PTB domain is used as an example of the utility of the SCOWLP database for analysis of PBRs. In this example, the clustering is selected for similarity cut-off value 0.4 (corresponding dendrogram shown in Figure 2.10). A structure-based alignment of the PBRs is obtained, and all interacting residue patterns are highlighted (panel c). A specific binding region is expanded to display all interfaces; in this case corresponding to PTB binding to phospho-tyrosine peptidic ligands. This binding

region gets automatically displayed in the 3D viewer for graphical inspection (panel a). This interface is expanded to obtain a structure-based alignment of all PTB domains that use this binding region for recognition. The secondary structure of the domain is displayed at the top of the alignment to help with interpretation of interacting information. The interacting residues are highlighted with different colouring; in this case based on the water contribution to their interfacial interactions (panel d). This information allows comparative analysis of the interfaces, including conservation vs. variation of the interactions. In this example we easily are able to analyze (at structure and sequence level) all the interfaces of the PTB domain with different phospho-tyrosine peptides and their interaction patterns. In the example, the three main recognition regions described for the X11 PTB and a peptide motif from the Alzeimer's amyloid precursor protein (APP; PDB entry 1AQC) are displayed and structurally aligned with the recognition regions of other peptides known to bind PTBs in this region. Also, specific differences in the interaction pattern can be further analyzed individually by clicking on each PDB entry code. Analysis of the conservation/variability of the interactions describing an interface may be of great utility for understanding energetic and evolutionary aspects of protein interactions and for helping in rational engineering and design.

2.2.5 Conclusions

Classification of the regions that a protein family uses to recognize binding partners is important for understanding the interactome. Protein binding regions are often irregular, discontinuous and can share interacting residues among them, making their clustering difficult, and arbitrary. A suitable classification requires proper measurements of similarity between protein binding regions and an appropriate clustering approach. Our approach consists of hierarchical clustering of the PBRs included in the SCOWLP database, which contains detailed interfacial information of proteins, peptides and solvent. We use the complete-linkage method and a similarity index obtained by mapping interacting residues in non-gap regions of pair-wise structural alignments. This approach provides a dilatation of the differences among clusters, making it suitable to isolate poorly separated clusters. In addition, we introduce flexibility in the usage of different similarity cut-offs for PBRs

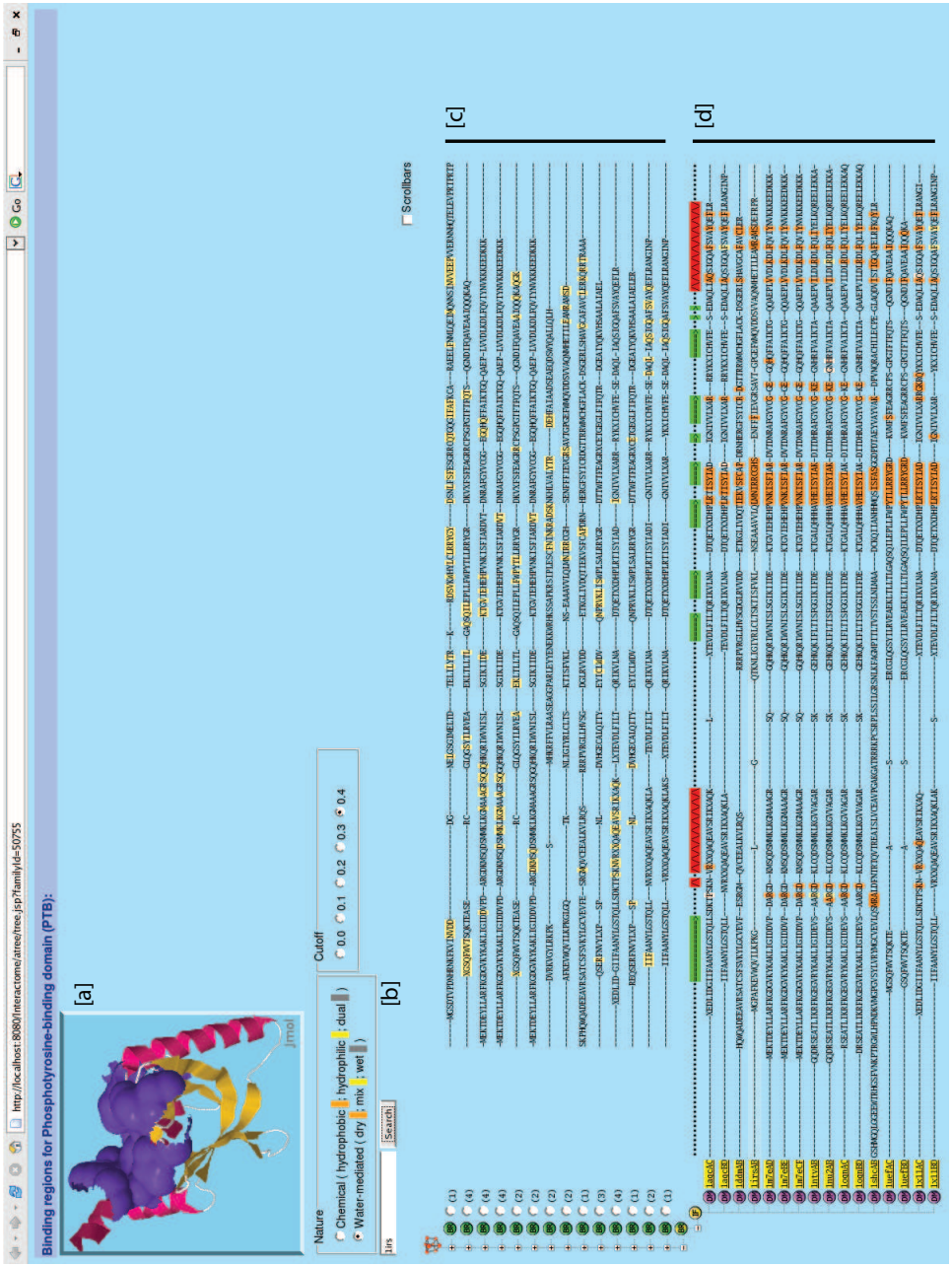


Figure 2.11: SCOWLP web application screenshot and utilities. PTB domain used as an example of the utilities of the SCOWLP database for analysis of PBRs. [a] 3D viewer allows structural analysis of binding regions the PTB family. [b] The Manager Box allows selection of Si cut-offs, display of interacting properties and keyword search. [c] Multiple structure alignment of representative PBRs of the PTB domain is provided together with the highlighting of the interacting residues, including solvent-mediated interactions. From the PBR tree, each PBR can be selected to be displayed in the 3D viewer. The PBR tree can be expanded by clicking on its branches to display all interfaces belonging to a particular PBR. [d]. Visualization of the contacting domains. Secondary structure of the domain is displayed, and physico-chemical properties of the interacting residues can be highlighted with the help of the "manager box" (see text for more details about SCOWLP utilities for analysis of PBRs)

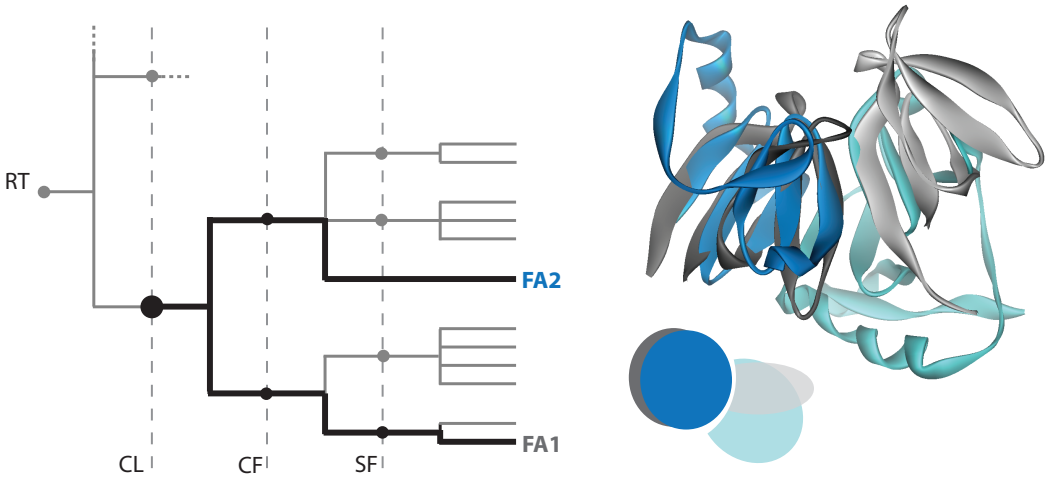
analysis. Our results show that, from 2,561 families containing binding regions, 65% use more than one binding region to interact. Furthermore, from all existing binding regions in SCOWLP, 22% are interacting with more than one protein family. In order to be able to analyze all family binding regions of the PDB in a detailed and comparative fashion we have implemented our PBR classification into the SCOWLP web application.

The current SCOWLP classification and its web application represent a complete framework for the study of protein interfaces and comparative analysis of protein family binding regions. Mining of this information may be of great utility for understanding energetic and evolutionary aspects of protein interactions, reconstruction of protein complexes, understanding protein networks and rational ligand design.

Chapter 3

BINDING INFERENCES ACROSS FOLD SPACE

Proteins exhibit an enormous diversity of structures, and different protein families, even different folds, may present non-obvious structural similarities. The detection of these similarities may lead to the discovery of evolutionary relationships and reveal common structural features important for function. In this chapter, I describe a methodology I developed to calculate binding regions conservation between structurally similar proteins. The non-obvious similarities are obtained performing pair wise non-sequential structural alignments between all proteins independently of their fold classification. This work is under revision for publication in the Proteins Journal.



[1]Scheme representing the methodology performed in this chapter to infer binding regions between structurally similar proteins classified in different families. The structure consists in the alignment of YgdR homodimer and Neurophysin II homodimer (2ra2:A,C and 1jk6:F,A) that belong to different folds (CF, class fold)

3.1 Studies on the inference of protein binding regions across fold space based on structural similarities

by Joan Teyra, John Hawkins and M. Teresa Pisabarro

in Proteins, 2010, submited

3.1.1 Abstract

The emerging picture of a continuous protein fold space highlights the existence of non-obvious structural similarities between proteins with apparent different topologies. The identification of structure resemblances across fold space and the analysis of similar recognition regions may be a valuable source of information towards protein structure-based functional characterization. In this work, we use non-sequential structural alignment methods (ns-SAs) to identify structural similarities between protein pairs independently of their SCOP hierarchy, and we calculate the significance of binding region conservation using the interacting residues overlap in the ns-SA. We cluster the binding inferences for each family to distinguish already known family binding regions from putative new ones. Our results indicate that binding region conservation across fold space occurs systematically and, therefore, there is a plethora of molecular recognition information that could be potentially inferred among proteins, independently of current fold classifications. We obtain a 6 to 8 fold enrichment of binding regions, and identify binding inferences for 728 protein families that lack binding information. We also investigate the presence of recurrent recognition structural features among protein families, and possible binding mode analogies between ligands from commonly clustered binding regions. The identification of similar recognition regions in proteins together with detailed analysis of their atomic and physico-chemical properties may constitute a first approximation to guide protein interactions prediction. The data obtained in this work is available in the download link at www.scowlp.org.

3.1.2 Introduction

Currently the limitations of protein structure classification approaches and the nature of the protein fold space are under debate. An emerging view is that the fold space may not be so discretely organized as thought [82–86]. Proteins exhibit an enormous diver-

sity of structures, and detection of similarities between them independently of current fold classifications may lead to the discovery of evolutionary relationships and reveal common structural features important for function [87]. Many of the current protein structure classifications are based on conventional structural alignment methods, which in their majority assume sequentiallity of the protein secondary structure elements (SSEs) favoring alignments of homologous proteins [7–9]. However, there is a growing occurrence of protein folds with common three-dimensional architectures but different topology due to duplication, swapping and/or deletion genetic events. In addition, protein structures presenting variations in SSEs composition (insertions/deletions) may still share a common structural core. Sequential structural alignment methods have difficulties to draw definite conclusions on structural resemblance in these cases [88, 89]. By contrast, non-sequential structural alignment algorithms (ns-SAs) overcome these limitations and have the ability to reveal unsuspected similarities between proteins belonging to different hierarchical levels in the current protein structure classifications [90–93].

The molecular function of a protein is defined on the basis of its interactions with other molecules, which take place through specific sites or binding regions [94]. The PDB contains atomic information about proteins and their complexes [10], but it is estimated that it is still far from having representative structures of all protein-protein interactions [95]. Despite the fact that current structural genomics efforts are progressively making such data available, to cover this knowledge gap and be able to understand the principles of protein recognition there is a need for approaches to facilitate comparative analysis of protein binding regions in known protein complexes.

Protein family members, which are homologous in sequence and structure, recognize other molecules through conserved binding regions and modes [44, 96]. This observation has recently inspired the development of several binding region classification techniques [55, 70, 71, 97] based on conventional structural alignments of the members. Analysis of this data has revealed an abundant redundancy of protein complexes, including homodimer products of crystallization [72]. These and other recent studies [97, 98] have reported that most protein families interact through many different regions, and that a single binding region may recognize multiple distinct partners. So far, protein binding region classification

methodologies have been restricted to comparing binding regions within family levels. The sequence identity between families is generally lower than within families, but they are still structurally related to each other and they might share similar binding properties.

In the late 90s, Russell et al. [94] performed a pioneer analysis of common binding regions of families within the same SCOP fold interacting with short polypeptides and small molecules. The authors reported nine examples of binding region concurrence in the apparent absence of sequence homology, proving that the methodology could be extended to protein-protein interactions.

It has also been observed that proteins with different folds and functions often interact with others through interfaces containing similar local structural features or interacting motifs, the number of which has been observed to be limited, highlighting the analogy between binding and folding [99]. So far, studies on the comparison of protein interfaces have been based on geometric algorithms that compare local atomic properties [100, 101] or non-covalent interactions [102] independently of the overall structure. The studies conducted by Nussinov and colleagues used geometric hashing techniques to perform several comparative interfacial studies [99], and Wolfson and coworkers use the same techniques to predict binding regions on protein surfaces [103]. In spite of their utility, structural alignment algorithms that do not require atomic connectivity have some intrinsic challenges that make them too sensitive for comparing binding regions: they are in some cases too rigid to allow certain structural flexibility and variability of the physicochemical properties.

Aytuna et al. [101] and Guenter et al. [104] have used ns-SAs to predict protein interactions when two structures contain regions in their surfaces that resemble the complementary partners of a known interface. To predict interactions these authors use template interface libraries to align both sides of an interface against unbound protein structures. All these studies have been useful for identifying similarities between protein binding regions and have revealed that relationships may appear among different folds.

Since it has already been observed that both, overall structural similarities between proteins and local binding region similarities also exist across folds, the aim of this work has been to structurally relate protein families independently of their fold classification, and to investigate protein binding regions inferences. We use ns-SAs and calculate the

significance of binding region conservation between protein pairs using the residues overlap of their binding regions in the alignment. We cluster known family binding regions together with the inferred ones derived beyond family level in order to define new putative recognition regions, p-sites. Our results indicate that there is a plethora of molecular recognition information that could be potentially inferred among proteins, independently of current fold classifications. We investigate binding mode analogies between ligands from known and inferred binding regions, and analyze the appearance of recurrent structural binding motifs contained in many protein families. Our approach provides a broad view of putative protein-family binding regions derived beyond family level, which may have important practical applications in guiding protein-protein docking experiments and rational engineering.

3.1.3 Methodology

SCOWLP database and binding region selection

We use our in-house protein binding regions classification scheme, SCOWLP [70, 97] to extract all existing protein-protein and protein-peptide complexes from the PDB. In the SCOWLP classification all protein complexes for each SCOP family are selected, and pairwise superimpositions of all members are performed. The similarity index for a binding region pair is then calculated based on the number of interacting residues overlapping in the structure-based sequence alignment of the proteins, as follows:

$$Si = \frac{IR_{overlap}}{\frac{IR_1 + IR_2}{2}} \tag{3.1}$$

where IR_1 and IR_2 are the interacting residues of each protein family member with the respective ligand, and IRoverlap represents the interacting residues overlap in the alignment. The collection of binding regions is then clustered to identify the non-redundant set of shared binding regions within each family. We use the agglomerative hierarchical algorithm [74] and the complete-linkage method, which considers distances between two clusters as their minimum similarity [75].

In this work, we extract a set of SCOWLP binding regions such that their clusters have zero similarity to each other [97]. This wide threshold allows binding regions that may include different ligand binding modes. We consider binding mode the relative orientation how structurally similar ligands are recognized by a protein receptor. All protein binding regions are mapped onto a single-family representative, which is the family member having more interacting residues at non-gap regions in a multiple structure alignment of the whole family. The representative includes all binding regions known for the family that it represents.

Protein family dataset

All protein families belonging to the first five classes of SCOP v 1.75 (all α , all β, α/β, $\alpha+\beta$ and multi-domain α/β) were included in the study. If the family contained binding information, the SCOWLP representative was selected, otherwise the first member in the SCOP family. This way, a total of 3,551 representatives were selected, where 2,532 of them contained binding information. Crystal packing contacts are filtered out using a support vector machine-based program, NOXclass [73] (cut off 70%), which takes into account their distinctive interface properties [105]. A total of 5,826 binding regions containing ≥ 5 residues were extracted from SCOWLP. Smaller binding regions were ignored since we observed that they were highly likely to be crystal packing artifacts that had passed the NOXclass filter.

Non-sequential structural alignments

The SSM algorithm [106] was used for the non-sequential structural alignments (ns-SAs). The method first matches graphs built on the secondary structure elements (SSEs) of the proteins, followed by an iterative 3D alignment of Cα atoms. SSM has the ability to find structural similarities between proteins independently of their topology allowing both, opposite directionality and different sequential order of SSEs. We performed all-against-all ns-SAs for all protein family representatives. This resulted in either sequential (SSEs consecutive in sequence) or non-sequential alignments. The following function, Q-score [106], was used to evaluate their structural similarity:

$$Q = \frac{N_{aln}^2}{\left[1 + \left(\frac{RMSD}{R_0}\right)^2\right] N_1 N_2} \tag{3.2}$$

where N_{aln} corresponds to the number of aligned residues, N_1 and N_2 are the total numbers of residues in the aligned structures, and R_0 is an empirical parameter that balances the relative significance of the root mean square deviation (RMSD) and N_{aln}. We have set R_0 to $3\mathring{A}$ following previous studies by Krissinel et al. [106]. Q-score reaches a value of 1 only for identical structures, and it decreases with similarity. The Q-score has been found to perform uniformly well in a broad similarity range of conditions and to be specially good in identifying low structural similarities [106], which is the requirement of our study. We define two structures to be similar enough to asses binding region conservation when the following conditions are satisfied: Q-score ≥ 0.15, number of aligned SSEs ≥ 3 and SSE coverage $\geq 40\%$ of at least one of the aligned proteins.

Binding regions conservation

We extract all the binding regions associated to the two representatives aligned. For each binding region, the interacting residues are mapped onto the structure-based sequence alignment. We also calculate the solvent accessibility for both representative proteins to distinguish the residues located in the core region from the solvent exposed ones, since core residues can not participate in recognition. We used NACCESS [25] to calculate the solvent accessibility of each residue using a probe sphere of radius $1.4\mathring{A}$. A residue is considered accessible if its total relative accessible surface area (RSA) is more than 5% [107]. We calculate the binding region conservation (BRC) as the ratio between the number of interacting residues located in structurally aligned regions that are also solvent exposed (IRaln) and the total number of interacting residues (IRtot):

$$BRC = \frac{IR_{aln}}{IR_{tot}} \tag{3.3}$$

P-value estimation

We assess the statistical significance of the BRC by estimating the p-values under the null hypothesis that two random protein families do not contain conserved binding regions. The estimation was carried out by calculating the BRC of $10\hat{5}$ randomly selected samples of protein representative pairs and a binding region for each pair. The distribution of these scores was used to estimate the p-values obtained as (r+1)/(n+1), where n is the number of samples that have been simulated (10^5), and r is the number of these replicates that have a score greater than or equal to the BRC value for which we are estimating the p-value [108]. Note that, as the sampling procedure can possibly contain undetermined cases of similar binding regions from the alternate distribution, these p-values are likely to be an underestimate of true significance, i.e. in some instances the real p-values will be much more significant. In a pairwise ns-SA, a binding region is inferred from one to another protein family if the conservation significance has a p-value ≤ 0.05.

Clustering of inferred binding regions

The inferred binding regions (iBR) and the known family binding regions (kBR) were collected for each family. Since binding inferences may occupy equivalent surface regions in the family, we re-clustered the binding regions in a similar way as described above for SCOWLP (see SCOWLP binding regions selection). To make sure that the obtained kBR clusters from SCOWLP are not modified in this process, we set the similarity between these initial kBR to zero. Three distinguishable cluster types were obtained: 1) those that only contained one kBR (family sites; f-sites), 2) those that only contained iBR (putative sites; p-sites), 3) those that contained both (mixed sites; m-sites).

Normalization

Each protein in SCOP is classified at four hierarchical levels below the root (RT): class (CL), class fold (CF), superfamily (SF) and family (FA). In this scheme, any pair of proteins necessarily has a single common ancestor level that relates them. For each of the all-against-all pairwise ns-SA we obtain the closest ancestor level (Table 3.1). The number of aligned pairs obtained per ancestor level is normalized by the number of pairwise alignments that

can be performed. At SF level the possible combinations are:

$$SF = \sum_{i=1}^{n} \frac{m_i \left(m_i - 1\right)}{2} \tag{3.4}$$

and at the upper levels:

$$CF, CL, RT = \sum_{i,j=1}^{n} \frac{m_i m_j}{2} \tag{3.5}$$

where n is the number of SF (or CF, CL, RT); i (and j) represents each SF (or CF, CL, RT); mi is the number of families for each SF (or CF, CL, RT). In a similar way, the number of significant binding regions is normalized based on the number of calculations performed for each level (SF, CF, CL, RT).

3.1.4 Results and discussion

All-against-all non-sequential structural alignments

We selected a representative protein for each of the 3,551 SCOP families and we performed all-against-all pairwise non-sequential structural alignments (ns-SAs) with them. These resulted in 6x10$\hat{6}$ ns-SAs that took about two weeks of calculations using 9 compute nodes. From the obtained ns-SAs, a total of 87,154 fulfilled the structural similarity constraints (see Methods: All-against-all non-sequential structural alignments).

The results obtained support previous observations of the existence of structural similarity among proteins independently of their SCOP hierarchy [90–93]. They also show a high percentage of similarities at inter-fold level (above fold level), and are shown in Table 3.1 (see total column for CL and RT). Once the data is normalized to the occurrence in a particular hierarchical level, we observe that 61.9% of all pairwise ns-SA performed at superfamily level (SF) present structural similarities. This high structural relationship between members of the same SF is to be expected, since proteins are grouped together at SF level when they have sufficient functional or structural similarity to infer a common origin. Although proportionally lower, the percentage found for the other levels is quantitatively significant,

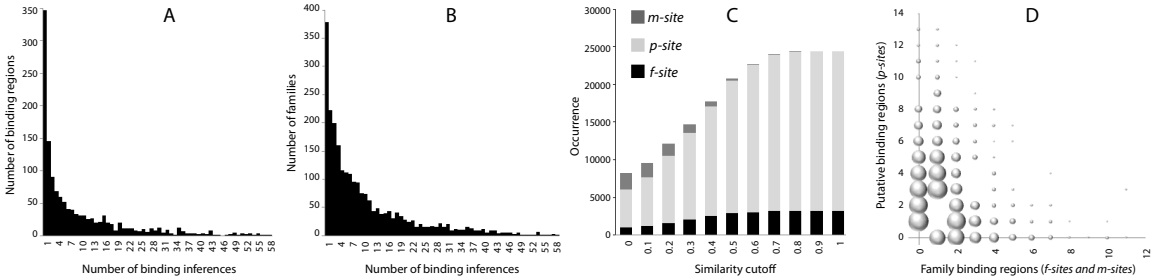


Figure 3.1: A) Number of binding inferences per binding region. B) Number of binding inferences per protein family. C) Histogram representation of number of binding regions after clustering depending on the similarity cutoff. D) Bubble representation of the number of predicted against the known binding regions per family, where the size of the bubble represents the amount of families.

as it represents an important count for structural similarities dispersed across folds. Interestingly, the high number of structural similarities found non-sequentially, specially at inter-fold level (see columns sq and ns), highlights the need of ns-SA approaches to relate protein structures and supports the current view of a continuous fold space [82, 84, 85].

Inference of protein binding regions

We mapped the binding regions of each protein on the pairwise ns-SA to assess their binding region conservation (BRC; see Methods). We calculated 206,889 BRC scores using the 87,154 pairwise alignments and 6,254 known family binding regions (kBRs). Each BRC score has associated a p-value representing the probability that a better score can occur by chance (see Methods). A total of 26,893 BRC scores were considered statistically significant, and the corresponding binding regions were then inferred. This result represents a 2.6 fold enrichment since the expected number of results at p-value ≤ 0.05 from the 206,889 BRC calculations is only 10,300. Based on this, and the fact that our p-values are an underestimate, we expect the false positive rate to be at most 38%.

The distribution of the data by ancestor level reveals that most of the inferences are inter-fold (see in Table 3.1 total column in Inferences for CL and RT). Once the data is normalized and in contrast with the structural alignment results, we interestingly observe that the percentage of significant inferences is comparable per ancestor level (norm column in Inferences in Table 3.1), which indicates BRC independently of the SCOP level.

The analysis of the number of inferences per binding region shows that most of the binding regions present a reduced number of inferences to other protein families, although in some cases may be up to 57 (Fig 3.1.A). We plot the binding inferences in a network in which the nodes represent protein families within the five SCOP classes (all α, all β, α/β, $\alpha+\beta$ and multi-domain α/β in Fig 3.2.A) and the edges represent inferences. The nodes are clustered based on the number of binding inferences (node connectivity). In this representation, as each class includes different folds, two proteins from the same SCOP class can be either connected by inter- or intra-fold binding inferences (edges in the plot). We observe from the network that within the classes there are sub-clusters of proteins presenting high intra-fold inference, which are connected by inter-fold inferences. This is better observed for all α and $\alpha+\beta$ classes. The distribution of the proteins in the network is driven by higher number of inter-fold inferences that connect the different classes (Table 3.1), where all-β acts as the pivotal class linking the others due to the common SSEs. The multi-domain class has a reduced number of families that are spread along the network and lowly connected with other families, showing no class preference. We also observe that all-β and α/β contain many protein families with low number of inferences, which in many cases are inter-fold.

Inspection of network sub-clusters populated with proteins from different classes, reveals the presence of inferred regions with slight differences in topology, orientation or number of SSEs, highlighting the existence of common architectures or binding motifs such as Rossmann fold, helix-turn-helix, β-sandwich and β-barrel (see Fig 3.2.A). Our observations confirm that proteins with different folds interact with others through regions containing similar structural features.

An example of binding inference based on structural similarity is shown for the Papain binding region of the protease inhibitor Staphostatin (Fig 3.2.B) [109]. The inhibitor interacts with the L- and R-protease domains through a "β-hairpin binding loop", which spans the active site of the protease. Staphostatin consists on a mixed eight-stranded β-barrel structure, where four strands define a binding region found structurally conserved in seven proteins (at inter- and intra-fold level). Our methodology identifies significant conservation in the strands forming the binding region. Due to the broad definition taken in our approach, some binding inferences may show a wide range of additional structural features:

Table 3.1: Distribution of the pairwise structural similarities and significant binding region inferences. The information is organized by ancestor level based in the SCOP hierarchy: root (RT), class (CL), class fold (CF) and superfamily (SF). The results of ns-SAs can be sequential (sq) or non-sequential (ns). Normalized values (N) and averaged Q-score (Q_{avg}) are shown (see Methods for details)

Ancestor level	Structurally similar protein pairs					Inferences	
	sq	ns	total	N(%)	Qavg	total	norm
RT	705	12,357	13,062	0.3	0.17	3,571	14.0
CL	5,230	49,153	54,383	0.9	0.19	16,522	14.5
CF	4,401	6,389	10,790	0.1	0.23	4,050	16.0
SF	4,489	4,430	8,919	61.9	0.26	2,750	18.0

longer loops connecting the β-strands defining the BRC (Fig 3.2.B:5, 7), additional SSEs (Fig 3.2.B:1, 2), or different connectivity between β-strands. Other inferred binding regions may even not contain the "β-hairpin binding loop" (Fig 3.2.B:1, 3) or present different β-hairpin length or conformations. Interestingly, it has been observed that the "β-hairpin binding loop" appears unstructured in the Staphostatin unbound form [55], suggesting that a range of flexibility in the search is required to describe these structural variations.

This example illustrates the usefulness of our approach to obtain a broad view of possible binding inferences that may help to identify putative ligands, which may present similar or different binding modes. Our approach may be combined with detailed atomic inspection for obtaining more precise binding information.

Classification of binding inferences

The analysis of the number of binding regions inferred per protein family for the 2,435 protein families taken in the study shows a wide variation, decreasing logarithmically the higher the binding inferences per family are (Fig 3.1.B). Some of this information is redundant, since different inferred binding regions (iBRs) may converge to equivalent regions or to known family binding regions (kBRs). Therefore, to reasonably classify the binding inferences, we performed a hierarchical clustering of the kBRs together with the iBRs. The kBRs and iBRs can either be clustered together (mixed sites; m-sites) or separately (being family sites; f-sites for kBRs, and predicted sites; p-sites for iBRs).

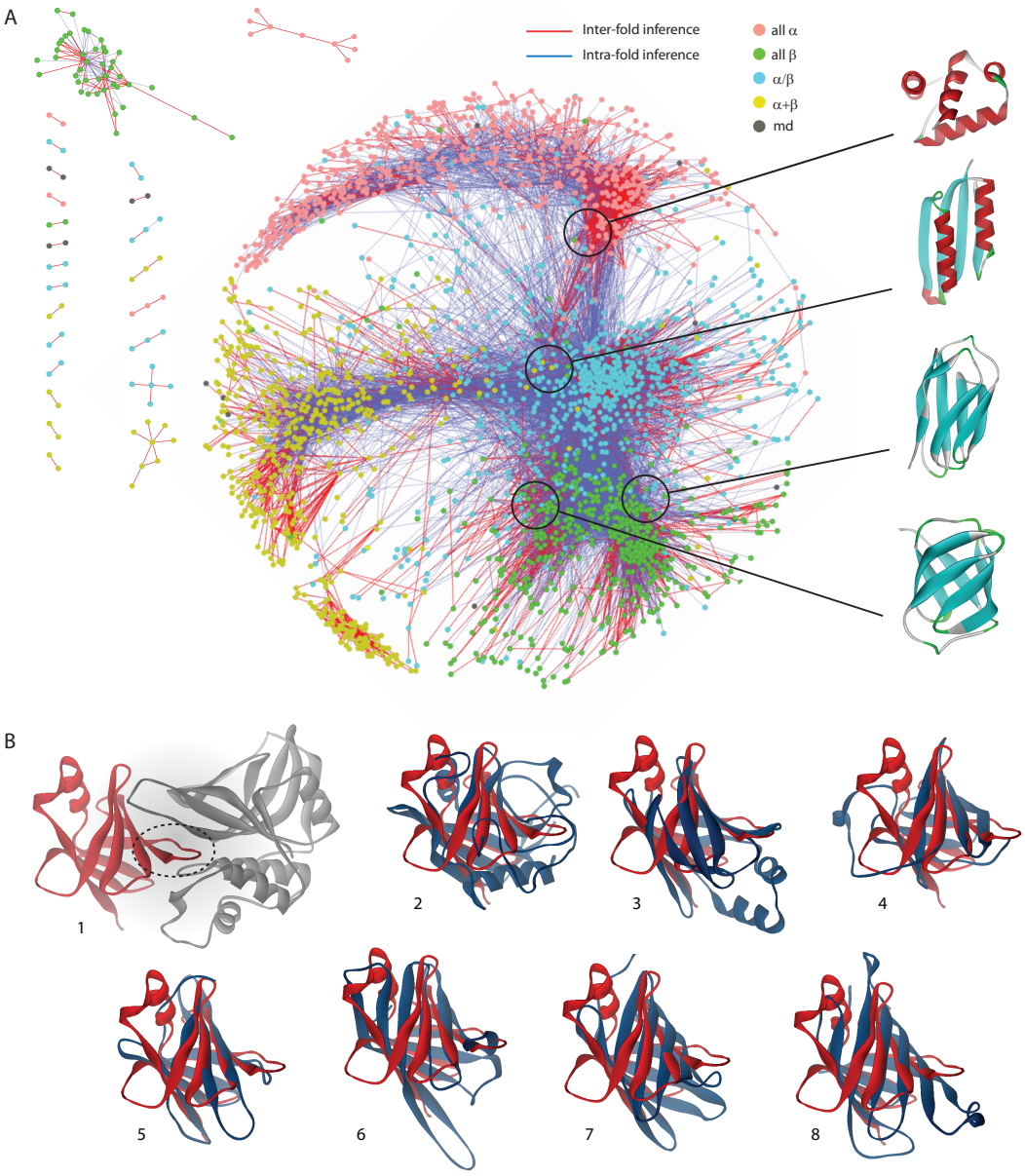


Figure 3.2: A) Connectivity network of the binding inferences. Network representation based on the binding inference connectivity (hedges) of the protein families (nodes) is shown using Cytoscape with the spring embedded layout. Representative examples of recurrent binding motifs are shown. B) Staphostatin protease inhibitor case. Inferences of binding regions from the Staphostatin protease inhibitor to the Papain protein family are shown. Panel 1) Structure of Streptavidin-like Staphostatin (red ribbon) in complex with the cysteine protease Papain (grey ribbon) is shown. (1pxv; chains C and A respectively). The "binding loop" is highlighted with a dashed ellipse. Panels 2 to 8 show the structure pairwise alignment of Staphostatin with seven proteins (blue ribbons) that present significant conservation in the corresponding Papain binding region. Panels 2 to 4: Proteins belonging to different SCOP folds than Staphostatin: 2) Cyclophilin-like (2esl:F), 3) Lipocalins (1xca:B), 4) Oncogene product p14-TCL1 (1jnp:B). Panels 5 to 8: Proteins belonging to the same Streptavidin-like fold: 5) Extracellular hemoglobin linker subunit (2gtl:O), 6) Avidin/streptavidin (2cam:B), 7) D-aminopeptidase (1ei5:A), 8) Quinohemoprotein amine dehydrogenase (1pby:A)

The clustering outcome is dependent on the applied similarity cutoff (Fig 3.1.C) and, therefore, the more restrictive the cutoff the more regions are obtained. For the initial 26,893 regions, for similarity cutoff 1.0 we obtain a total of 24,386 non-redundant clusters, whereas for cutoff 0 we get 8,259. We observe that the number of p-sites is 6 to 8 fold higher than the number of f-sites independently of the cutoff applied, which is indicative of a considerable enrichment of putative binding regions for protein families.

Proteomics studies have already shown that proteins may interact with many partners,[110, 111] and that timing and location play a key role in these interactions. In addition, protein interaction conservation analysis for protein families have also revealed that when a single binding region is taken into account [48] the results are not as conclusive as when multiple interfaces are considered [112], in which case it is observed that the interacting residues are significantly more conserved than those in the rest of the surface. Our methodology may help to add insights to the current structural protein recognition knowledge by suggesting putative protein recognition regions.

An illustrative example could be the Ornithine Decarboxilase (OD) family belonging to the CheY-like fold (Fig 3.3.A). There is no structure of OD in complex with other proteins and, although some inferences could be made from members of the same fold (CheY-like), with our approach we identify six structurally related protein families to OD, from which five are inter-fold related. Once their corresponding conserved binding regions are inferred to OD, it can be interestingly observed that they cover a large proportion of the solvent accessible surface of OD. The binding inferences located in equivalent regions are clustered together regardless of the different ligand recognition mode, and this information collected across fold space might give clues about potential recognition properties of the Ornithine Decarboxilase CheY-like family.

Ligand binding modes in m-sites

The ligands corresponding to complexes, where the binding regions of the receptors cluster together, may either present analogous or different binding modes (i.e. the relative orientation the ligand adopts when recognized by the receptor). Previous studies at family level revealed that only 4.2% of all interfaces clustered in the same binding region have analogous

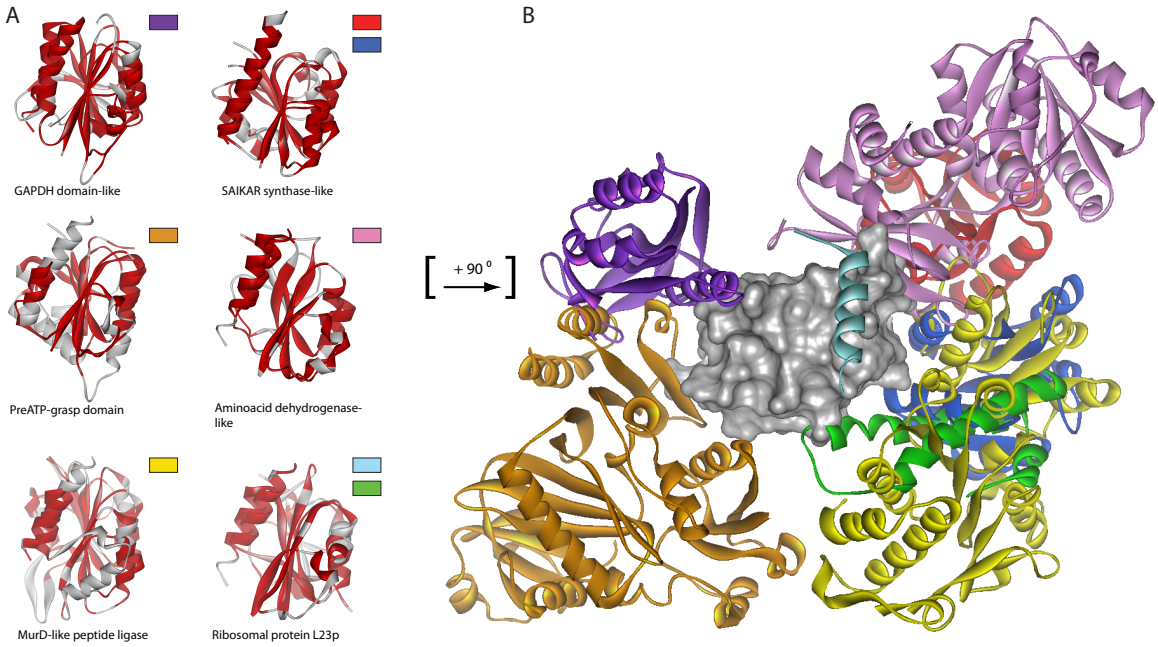


Figure 3.3: Binding regions inferred to Ornithine Decarboxilase (OD). A) Pairwise ns-SAs of the OD domain (1c4k:A) with six structurally related proteins. The proteins are shown as white ribbons, and the aligned regions are colored in red. Only SAIKAR synthase-like (top right) is intra-fold related to OD. The color code of the boxes on the right top of the ns-SAs corresponds to the protein structures of the respective ligands shown in panel B. B) The OD structure (gray surface) is shown with the respective binding partners in the inferred binding regions from the proteins in panel A (rotated +90° for better visualization). The structures were obtained from PDB: 1xbt:A,B (pink), 1zgz:A,C,D (red, blue), 1b6s:B,A (orange), 1npd:A,B (pink), 1eeh:A,A (yellow), 1jj2:R,1 (clear blue), 2qa4:S,V (green). In the PDB codes, the first chain listed corresponds to the protein with observed similarity to OD and the second to the ligand.

binding modes [98]. We analyzed the m-sites to find out binding mode analogies between ligands corresponding to inferred and known family binding regions (0.6 similarity cutoff; Table 3.2). We performed a detailed visual inspection of the results obtained. Of these results we identified 46 out of 81 analogous binding modes at intra-fold level, and 9 out of 65 at inter-fold level. The most frequent cases were homodimers of structurally similar proteins (double-homodimer in Table 3.2 and Fig 3.4.C, D).

Two cases correspond to a complex of two proteins that are structurally similar one another, allowing the alignment in reverse order (reverse-similarity in Table 3.2). At intra-fold level we found four cases of similar enzymatic subunits, where the structurally similar receptors were classified in different SCOP families and their ligands belonging to the same

Table 3.2: Distribution of ligand binding mode analogies in m-sites. The m-sites are obtained at 0.6 similarity cutoff. The similarity of the ligands is assessed by structural similarity, and the binding mode is inspected manually.

	Same fold	**Cross fold**
Different binding mode	38	56
Same binding mode:		
≫ same ligand	4	1
≫ similar ligand:		
≫ ≫ *double homodimer*	35	7
≫ ≫ *cross-similarity*	2	0
≫ different ligand	2	1
Total	81	65

family (Fig 3.4.E). Our method could also identify an interesting case with the same ligand binding mode at inter-fold level, where one of the receptors is structurally similar to a part of the other (Fig 3.4A). The last three are quite striking since correspond to structurally different ligands but sharing similar binding modes (Table 3.2 and Fig 3.4B, F, G).

Our methodology is able to identify new analogies between protein binding modes both within fold families and between protein folds. The number of analogies that we could identify is limited by the availability of representative data about protein complexes in the PDB [95]. We predict that particular structural motifs may recognize ligands at similar locations but in different modes, which agrees with previous studies performed at family level [98]. Some of this binding plasticity may be promoted by solvent, which plays an important role in the definition of protein interfaces [68, 113–115]. The analysis of ligand binding modes performed here with the m-sites should be extended to all binding region clusters, as their comparison can be of high relevance in studies of protein-protein interactions and might have important implications in rational engineering and design. For such a purpose, we foresee the potential of our methodology assisted by complementary techniques, such as algorithms aligning non-connected atoms with physicochemical features [116].

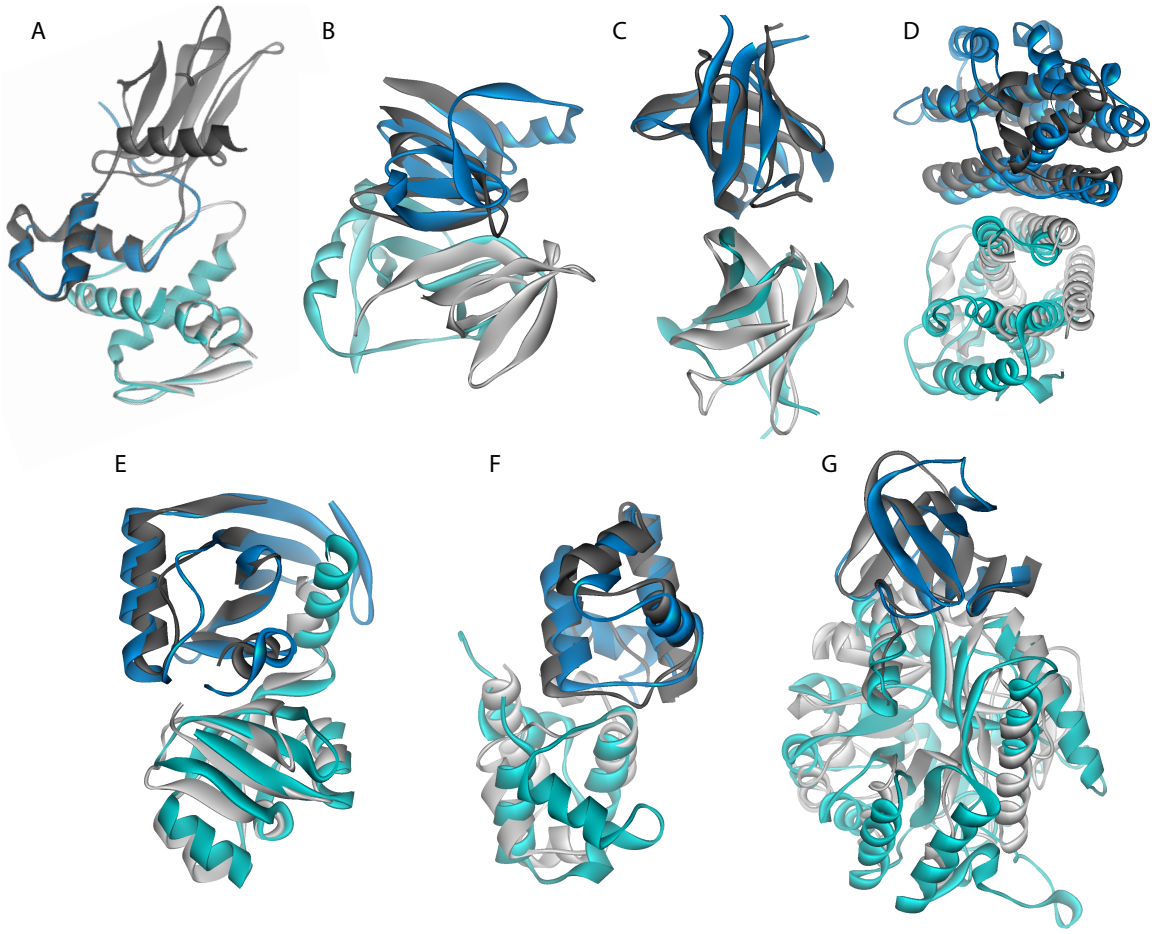


Figure 3.4: Illustrative examples of binding inferences with analogous ligand binding modes. Examples of inter-fold (A-D) and intra-fold (E-H) related binding inferences (see Table 3.2). The ns-SAs of structurally similar protein pairs are shown as ribbons in dark blue and gray and their respective ligands are shown in light blue and gray. A) Alignment of SOCS and VHC domains in complex with POZ domains (2c9w:A,C in blue and 1lm8:V,C in gray); B) Alignment of YgdR homodimer and Neurophysin II homodimer (2ra2:A,C and 1jk6:F,A); C) Alignment of YtmB homodimer and F1 ATP synthase α+β subunits homodimer (2nwa:A,B and 1bmf:A,D). D) Alignment of MiaE homodimer and long-chain cytokine homodimer (2itb:A,B and 1rhg:A,C); E) Alignment of APE0525 and Ta1423 domains in complex with PUA domains (1zs7:A,A and 1q7h:A,A); F) Alignment of PrgX C-terminal and SinR domains in complex with PrgX N-terminal and TM1459-like domains (2aw6:A,B and 1zz6:A,B); G) Alignment of Zn-dependent carboxipeptidase and NagA domains with their catalytic domains (2qs8:A,A and 1o12:A,A).

Enrichment of family binding regions

Some protein families contain much structural information about their complexes, whereas others may not have available information at all [78, 97]. Likewise, we observed the same trend for the putative binding regions described in this work, p-sites (Fig 3.1.D). It is noteworthy that a large number of the families, for which binding information is not known, are predicted to have binding regions by our methodology. Interestingly, most of these protein structures have been obtained from structural genomics initiatives [117], and most of them represent new folds or families for which no functional or binding data is available. These initiatives are increasing the number of SCOP folds and superfamilies containing a single family [118]. Out of 1,019 families without any known binding information, we find 728 with at least one p-site. Most of the binding inferences are found at inter-fold level, highlighting the importance of considering structural resemblances across folds to locate putative binding regions for uncharacterized proteins.

3.1.5 Conclusion

Many computational efforts have been directed towards the comparison and classification of binding regions of known protein complexes of the same family by making use of their high structure and sequence similarity. Also local comparative studies of protein binding regions and interfaces independently of the protein fold have been performed. However, to our knowledge no systematic comparison of protein binding regions beyond family level has been performed in order to obtain binding inferences using non-obvious structural resemblances across all folds.

We use non-sequential structural alignment algorithms to be able to reveal unexpected resemblances among three-dimensional protein architectures independently of their topology. The methodology we use assesses the significance of binding region conservation across all superimposed pairs of proteins independently of their family SCOP classification. Analysis of our results reveal that binding region conservation across fold space occurs systematically, and that it is not limited to homologous but extends to proteins considered having different folds in the current protein topology classifications. We find a high number of structural relationships beyond family level and detect nearly 27,000 putative binding in-

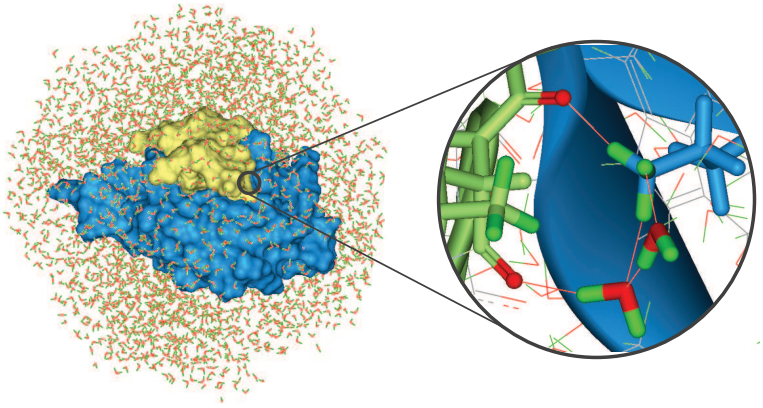
ferences among 3,551 SCOP families, from which the majority are derived from a different fold. After hierarchical clustering of the inferred binding regions we obtain a 6 to 8 fold enrichment of the known binding regions at family level. We obtain putative binding inferences (p-sites) for 728 protein families that have no binding information known, which are mostly derived from current structural genomic initiatives.

Current structural genomics efforts are aiming to experimentally solve representative protein structures and their complexes to complete the protein fold space and the structural interactome, and are creating a big demand for adequate automatic tools to sort, study and compare the existing and future structural data in order to understand and be able to predict protein-protein interactions. In these lines, our approach provides a broad view of protein binding regions across fold space and it is helpful for deriving putative binding inferences, which in combination with other complementary techniques may enhance the current picture of the structural interactome and assist protein-protein docking experiments and rational engineering.

Chapter 4

WATER IN PROTEIN INTERFACES

Protein interactions take place in aqueous solution, and water molecules can mediate residue-residue interactions complementing direct interactions. The specific role that water plays in binding affinity and specificity is still not well understood. This lack of knowledge makes solvent often ignored in many computational protein interaction analysis, rational drug design and protein docking studies. In this chapter, I make use of the novel interfacial information contained in the SCOWLP database (see Chapter 2) to study the characteristics of solvent and water-bridged residues in protein interfaces. In the first section, I carry out a descriptive analysis of interfacial solvent in a representative dataset of high-resolution protein complexes containing interfacial water molecules. In the second section, I studied the mobility and energetics of interfacial water molecules. (both sections correspond to articles published in the Proteins Journal).



[1]The image in the cover represents a protein complex (blue-yellow) surrounded of water molecules (green-red) simulating biological conditions

4.1 Characterization of interfacial solvent in protein complexes and contribution of wet spots to the interface description

by Joan Teyra and M. Teresa Pisabarro

in Proteins 2007, 67(4):1087-1095

4.1.1 Abstract

Water networks in protein interfaces can complement direct interactions contributing significantly to molecular recognition, function, and stability of protein association. Thus, water can be seen as an extension or addition of protein structural features, which may add plenty of information to protein interfacial definition. However, solvent is frequently neglected in protein interaction studies. Analysis of the interfacial information contained in the PDB is essential to achieve more accurate descriptions of protein interfaces. With this aim, we have used the SCOWLP database (http://www.scowlp.org) and applied computational geometry methods to extract and analyze interfacial information of a high-resolution non-redundant dataset of 176 protein complexes containing obligate and transient interfaces. We have identified all interfacial residues and characterized them in terms of temperature factors, secondary structure, residue composition, and pairing preferences to understand their contribution to the interface description. We have paid special attention to water-bridged residues; focusing on those that interact only mediated by a water molecule called wet spots. Our results show that 40.1% of the interfacial residues are interacting through water and that wet spots represent a 14.5% of the total, emphasizing the importance of the inclusion of solvent in protein interaction studies, and the contribution of wet spots to interfacial description. Wet spots present similar characteristics to residues binding buried water molecules in the core or cavities of proteins; being preferably located in non-regular secondary structures and establishing hydrogen bonds by their main-chains. We observe that obligate and transient interfaces present a comparable amount of solvent. Moreover, the role of solvent in both complex types differs according to the different nature of their interfaces. The information obtained in our studies will assist in the process of accomplishing more accurate descriptions of protein interfaces and may be helpful to improve comparison of protein family interfaces, to facilitate rational ligand design, and to guide protein docking.

4.1.2 Introduction

Understanding the binding properties of proteins constitutes the essence of functional genomics. To understand what a protein does, information is needed about what it binds, and even more important, how. Non-covalent contacts between amino acids are the basis of protein-protein interactions and they are responsible for affinity and specificity in biological processes. Furthermore, protein interactions take place in aqueous solution, reason why buried solvent molecules are common in protein interfaces [61, 63, 115]. Water networks bridging binding partners can complement interfacial direct interactions and contribute significantly to molecular recognition, function, and stability of protein association [60, 64, 119–121]. However, solvent is frequently ignored in protein interaction analysis and, for this reason, many protein residues are not taken into account in the definition of interfaces.

The structural data contained in the PDB is often used to study protein interfaces. In our previous work, we developed the SCOWLP database and a web-based visualization tool to extract and analyze at physicochemical level the atomic interactions of all protein interfaces of the PDB [70]. It describes protein interfaces including solvent as a mediator of protein interactions and considering all residues of the PDB that interact only through water, which we call wet spots.

Protein complexes have been classified according to their lifetime (transient, permanent), the possibility of finding them stable in vivo (obligate, non-obligate), and their composition (homo-, hetero-oligomeric). However, it remains arbitrary to classify protein complexes due to the overlap between types and the limitation on their biological knowledge (i.e. localization, co-expression, or binding energies) [122].

Several recent studies have investigated protein interfaces with the aim of characterizing protein complexes [47, 123–125]. All these studies have used different datasets to investigate and compare the properties of the different types of protein complexes. However, they all overlook interfacial solvent and its importance to protein binding and stability. Few studies have paid attention at water-mediated protein interactions [42, 68]. They have showed that crystal packing interfaces are more polar (often 50% more solvated) and that homo-

oligomers have slightly more hydrated interfaces than hetero-complexes; although presenting high heterogeneity within groups.

In the present work, we consider obligate and transient protein complexes using folding considerations [126]. Interfaces formed by protein-chains where the process of folding and binding is essentially inseparable are obligated to interact or in permanent contact (i.e. multi-subunit enzymes). On the other hand, interfaces formed by proteins that fold independently and then associate to carry out a particular biological task are transient [41, 127]. These are usually the key players in the regulation of biochemical pathways and signal transduction cascades. Permanent complexes are functionally obligate and only exist in their complexed form, while transient complexes exist independently and associate/disassociate in vivo. The comparison of interfaces of relatively higher hydrophilic nature (formed when two-folded monomeric proteins associate) with those more related to a protein core [41, 59, 128] (formed in coupled folding-binding process) may suggest a different contribution and role of solvent.

For our work, we have used a curated dataset consisting of obligate and transient high-resolution protein complexes to study their solvent contributions. We have extracted from the SCOWLP database all detailed information about water-mediating interfacial interactions and the protein residues involved. We have carried out a detailed analysis of protein interfaces looking at water-bridged residues, and identifying and characterizing wet spots.

Our studies identify interesting specific characteristics of water-bridged residues in terms of secondary structure, temperature factors, residue composition, and pairing preferences, and emphasize the role of wet spots as important contributors to protein interfaces. Interestingly, we observe that obligate and transient interfaces present a comparable amount of solvent, although each contains particular features.

4.1.3 Methodology

Dataset

For our studies, we created a curated dataset consisting of 104 obligate and 72 transient high-resolution protein complexes (Table 4.1). We started with a nonredundant dataset from Weng and Mintseris [126] and we updated and modified it as follows. For each complex of

the initial dataset, we used the SCOP classification [8] to find all complexes of the PDB with the same family-family composition. We restricted the search to crystallographic structures, and we used a threshold of resolution 2 $\AA$ to account for reliability of water positions. A manual check-step was necessary to be able to automatically identify the correct binding site in multichain complexes. Then, we selected as a representative of the family-family pair the complex with the highest amount of interacting waters and containing wet spots. This selection criterion was established to circumvent the intrinsic underestimation of water molecules in crystallographic structures.

Interface definition and characterization

We retrieved all the interfacial information used in this study from the SCOWLP database (http://www.scowlp.org), which describes all domain, residue, and atom interfacial interactions of the PDB [70]. SCOWLP considers interactions based on atom type and inter-atom distance criteria, including all interfacial solvent as an additional interface descriptor. We used a geometrical algorithm (Convex Hull [67]) to obtain a good approximation to the protein domains' shape, and we expanded it for both interacting domains to define their intersection as the interface, taking the atoms in the intersection for a pairwise Euclidean distance calculation. We used the same principles to finely capture the geometrical area of the interfaces. In addition, we characterized the interface size by the number of residues forming it.

We defined the interactions based on distance criteria between atom types. For hydrogen bonds, we considered a donor-acceptor distance 3.2 $\AA$, for salt bridges 4 $\AA$, and for van der Waals energies the van der Waals radii distance. Interfacial residues were then classified according to the interaction type as dry (direct interaction), dual (direct and water-mediated interactions), and wet spots, which are residues interacting only through one water molecule (Fig 4.1). For each interface of our dataset, we extracted the total interfacial residues, water-bridged and wet spots, and we used this information for comparative analysis. After the interfaces were defined for both, obligate and transient protein complexes, we analyzed their temperature factors, residue propensities, pairing preferences, and secondary structure (α-helix, β-sheet, and unstructured regions). Based on the nature

Table 4.1: Non-redundant dataset

OBLIGATE protein complexes (104)

1a9x C:D	1fm0 D:E	1i9c A:B	1kz8 A:F	1nzi A:B	1uc4 A:G
1ade A:B	1g72 D:C	1iak A:B	1l8a A:B	1o7n A:B	1umd A:D
1b25 C:D	1g8k E:F	1ir1 D:V	1li1 DE:F	1o97 C:D	1uzb A:B
1bbh A:B	1g8t A:B	1ird A:B	1lm8 B:C	1oao AB:C	1xgs A:B
1cmc A:B	1go3 F:E	1j5w A:B	1luc A:B	1oaq H:L	2ahj A:B
1d7w A:B	1got G:B	1jb7 A:B	1m1n A:B	1og8 A:B	2aps A:B
1dce A:B	1gt3 A:B	1jbo A:B	1m4r A:B	1ooy A:B	2bbk H:L
1dj7 A:B	1h32 A:B	1jcr A:B	1mka A:B	1owf A:B	2cst A:B
1dxr C:HLM	1h6k B:Y	1jnr A:B	1mro A:B	1p7g Q:T	2ubp A:C
1e3d A:B	1h8e A:D	1jsd A:B	1mty B:D	1pby A:C	3daa A:B
1e9g A:B	1h8e B:E	1k5n A:B	1my7 A:B	1pby AC:B	3lyn A:B
1ec3 A:B	1hbn D:E	1k8k B:F	1mzn A:C	1qgw A:C	3pcc A:M
1ed9 A:B	1hbn D:F	1k8k C:F	1n5w A:B	1qip A:B	8gss A:B
1eex A:B	1hbn E:F	1k8k D:F	1n60 A:B	1qq5 A:B	9wga A:B
1eto A:B	1hfe T:M	1kfc A:B	1n62 D:F	1r4p A:BCF	
1f0v A:B	1hj5 A:B	1kqf A:B	1nh2 BC:D	1req C:D	
1f3u A:B	1hsb A:B	1kqf B:C	1nms A:B	1sox A:B	
1fj2 A:B	1i7q C:D	1kqp A:B	1nr4 E:F	1ubp B:C	

TRANSIENT protein complexes (72)

1a4y D:E	1dtd A:B	1gcq B:C	1jiw I:P	1m5a A:D	1r0t A:B
1acb E:I	1eay A:C	1gl4 A:B	1jps HL:T	1m9y E:H	1r3j AB:C
1ava A:C	1eer A:B	1got A:B	1jtg C:D	1mz8 C:D	1rew AC:B
1b2s B:E	1euv A:B	1gvn C:D	1jw9 B:D	1nci A:B	1s6v B:C
1blx A:B	1f3v A:B	1h1r A:B	1kli H:L	1nmm C:D	1ugh E:I
1clv A:I	1f60 A:B	1he1 B:D	1klu A:D	1o6s A:B	1wej F:HL
1d2z A:B	1f83 A:CB	1hx1 A:B	1ksh A:B	1o94 AB:CD	2btc E:I
1d4x A:G	1fle E:I	1i2m C:D	1kxv B:D	1oga AC:DE	2sge E:I
1dan LH:U	1flt VW:X	1icf D:J	1l6x A:B	1on1 A:B	3btq E:I
1dhk A:B	1fns A:HL	1iqd AB:C	1lk3 A:HL	1osp HL:O	3gal A:B
1dpj A:B	1fs1 C:D	1j34 AB:C	1lw6 E:I	1qav A:B	3lyn A:B
1dqj AB:C	1g4y B:R	1jdh A:B	1m48 A:B	1qkz A:HL	3sic E:I

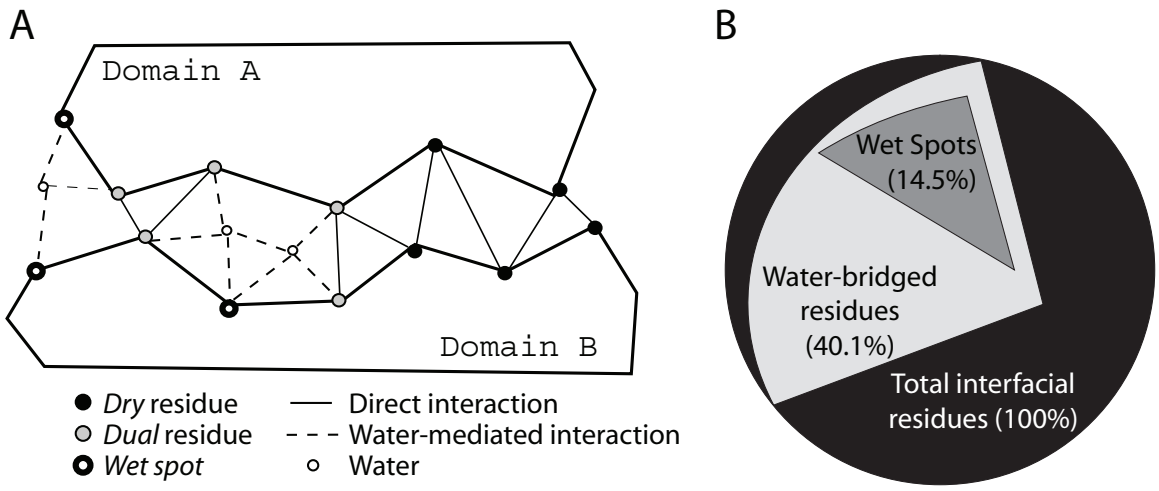


Figure 4.1: Residue interaction types. (A) Definition of residue interactions: Interface between domains A and B is formed by 13 residues (five dry, five dual, and three wet spots). (B) Partition of residue interactions (percentages taken from Table 4.2).

of the contact, we differentiated between main- and side-chain interactions.

Normalizations

The information given for all the interfaces in Table 4.2 was normalized against the number of complexes per type. We also normalized the amino acid pairing preferences and secondary structure composition against the frequency of appearance of each residue or secondary structure element in the complex as follows:

$$S_{ij} = \frac{b_{ij}}{\frac{c_i}{n_c} \times \frac{c_j}{n_c}} \tag{4.1}$$

where S_{ij} is the score for the pairing preference between amino acids i and j or the secondary structure elements i and j. The value of b_{ij} is the number of binding pairs between i and j that occur at the interfaces of the dataset. The denominator is the product of the relative frequencies of appearance at the interface of the amino acids or secondary structure elements i and j. The relative frequencies are derived by the quotient of the number of the amino

Table 4.2: Summary table.

	Obligate	**Transient**	**All**[a]
Number of Complexes in the dataset	104	72	176
Mean number per complex:			
Area ($\AA^2$)	3362	1394	2557
Total interfacial residues	85	39	66
Water-bridged residues	35	15	27
Wet spots	12	6	10
Interacting waters	17	8	13
Waters mediating wet spots	10	4	8
Mean number per 1000 $\AA^2$:			
Total interfacial residues	25	28	26
Water-bridged residues	10	11	10
Wet spots	4	4	4
Interacting waters	5	6	5
Waters mediating wet spots	3	3	3
Percentage (%):			
Water-bridged/total residues	41.6	37.9	40.1
Wet spots/total residues	14.8	14.2	14.5
Wet spots/Water-bridged	36.1	37.2	36.4

[a] The averages are normalized against the number of complexes per type

acids or secondary structure elements i or j occurring at the interface (c_i or c_j) and the total number of amino acids or secondary structure elements at the interfaces in the whole dataset (n_c).

4.1.4 Results and discussion

Our curated dataset consists of 176 protein-protein complexes: 104 obligate and 72 transient (Table 4.1). The average number of wet spots and the average number of interacting water molecules in our dataset are similarly distributed, showing their nondependence on resolution (Fig 4.2). We studied the contribution of solvent to the interfacial characteristics of obligate and transient protein complexes in our dataset by analyzing and comparing the

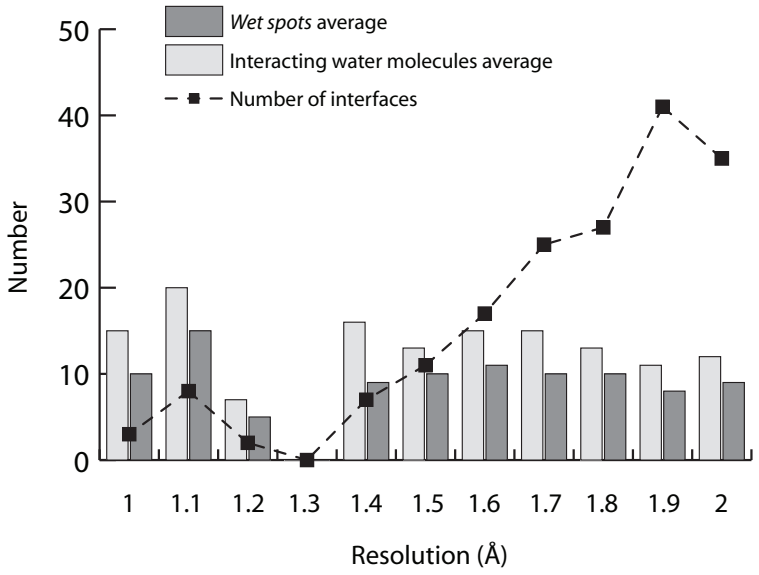


Figure 4.2: Interface solvent distribution of our dataset by resolution. The dataset of protein interfaces used is given in Table 4.1.

temperature factors, secondary structure, residue propensities, and pairing preferences of all interfacial residues (dry, dual, and wet spots).

Participation of solvent in protein interfaces

We defined the interface size based on both, the number of interacting residues and the geometrical area (see Materials and Methods). Table 4.2 summarizes the average number of residue interactions per complex, per 1000 $\AA^2$, and in percentages for obligate, transient, and both protein complexes. The number of water-bridged residues per interface is on average 27. This means that 40.1% of the interfacial residues interact through water, either complementing direct interactions (63.6%) or forming wet spots (36.4%) (Fig 4.1.B). The fact that water-mediated interactions at protein interfaces are almost as numerous as direct residue interactions clearly stresses the relevance of solvent in protein interfaces. Rodier et al. also observed a similar abundance of water-mediated interactions and direct hydrogen bonds in their interfacial hydration analysis of homo- and hetero-complexes [68]. Moreover, we find that the average of wet spots in protein interfaces is 14.5% of the total interfacial residues, and that more than half of the interacting waters in the interface are bridging wet

spots (Table 4.2).

In terms of interfacial area, there are on average five interacting water molecules and 26 interfacial residues per 1000 $\mathring{A}^2$ in protein complexes. From those, 10 are water-bridged; four of them wet spots. We appreciate that the existing differences between obligate and transient complexes in terms of hydration vanish when we consider them relative to the area (per 1000 $\mathring{A}^2$). The interface size in obligate complexes is twice as high as in transient complexes [122], thus obligate complexes appear slightly more hydrated (41.6% vs. 37.9%) and less dense (25 vs. 28 residues/1000 $\mathring{A}^2$) than transient. The correlation between water-bridged residues and area is much stronger in obligate ($R^2 = 0.76$) than in transient ($R^2 = 0.36$). However, the correlation between wet spots and area is weak in both cases; although obligate ($R^2 = 0.59$) is stronger than transient ($R^2 = 0.29$). These discrete correlations reflect a high heterogeneity within obligate and, even more, within transient complexes in terms of solvent. Rodier et al. also observed different degrees of solvation among each complex type in their studies of a homo-/hetero-complexes dataset [68].

Role of interfacial solvent

To analyze the role that the solvent plays in protein interfaces, we plotted the percentage of wet spots of the total water-bridged residues against the percentage of water-bridged residues of the total number of interfacial residues (Fig 4.3). While the percentage of water-bridged residues represents the level of the interface's hydration, the percentage of wet spots reflects the role of solvent either complementing direct interactions (low %) or adding new interacting residues (high %). Looking at Figure 4.3 divided into four quadrants at 50% hydration and 50% complementation/addition, we can appreciate a central area (quadrant I) where the majority of the interfaces are placed with no complex type distinction. Of special interest are quadrants II and III, where the most hydrated interfaces having up to 70% water-bridged residues can be found; showing again no difference by complex type. In quadrants III and IV, we find those interfaces with higher percentages of wet spots. There are some interfaces where more than 25% of the interfacial residues are wet spots, highlighting the importance of the inclusion of wet spots in the description of protein interfaces.

Figure 4.3 shows an overlapping of both complex types, illustrating the heterogeneity

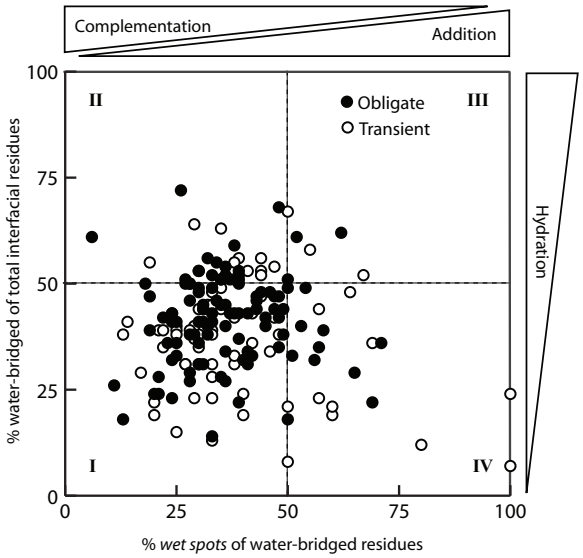


Figure 4.3: Role of interfacial solvent. The level of hydration is given according to the percentage of water-bridged residues of the total number of residues forming part of the interface (y-axis). The percentage of wet spots of the total water-bridged residues indicates the contribution of wet spots to complementation of direct interactions or addition of interacting residues to the definition of the interface (x-axis). The figure is divided into four quadrants for discussion.

of obligate and transient protein complexes with respect to water-bridged residues and wet spots. This suggests that the role of interfacial solvent is not dependent on the type of complex but on intrinsic properties.

Interfacial distribution of wet spots

Protein topology determines many aspects of the molecular recognition mechanism. However, because proteins do not always entirely fit in their binding modes, they use solvent as an important contributor to their shape complementarities. Water molecules act as molecular glue between two binding proteins, providing polar interactions even between residues that are too far to interact directly. Wet spots in protein interfaces can be located both externally and internally, enlarging and/or enriching the interface definition, respectively. To analyze the distribution of wet spots in interfaces, we calculated the interfacial area without considering solvent (A_{dry}) and the interfacial area considering interactions mediated by one water molecule (A_{solv}). Our aim was to use the difference of the calculated geometrical

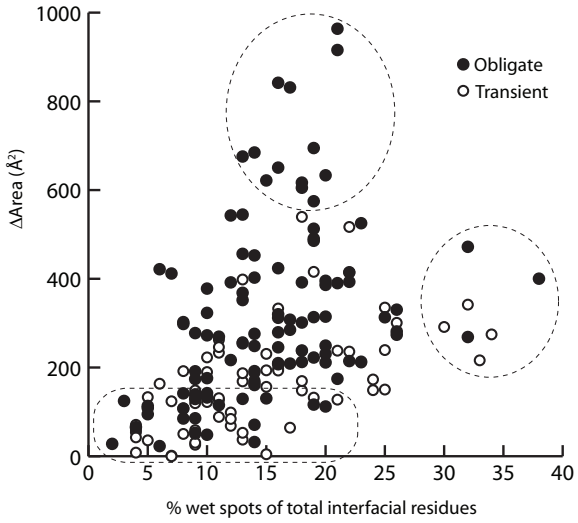


Figure 4.4: Contribution of wet spots to the enlargement or enrichment of protein interfaces. The Area is the difference of the interfacial area with and without solvent (A_{solv} - A_{dry}).

areas (Area = A_{solv} - A_{dry}) as an indicator of interface enlargement or enrichment due to the addition of wet spots. While for a large area we should find interfaces containing many wet spots distributed surrounding the interfacial area, for a low area wet spots should most often be internally located. In Figure 4.4, we can appreciate that for a high number of wet spots only obligate complexes emerge due to their larger interface size. However, it is difficult to differentiate between obligate and transient, and exceptions exist in both categories due to the wide variety of interface morphologies [68] (i.e. well-defined interacting cores surrounded by wet spots enlarging the interface, wet spots connecting dry multipatch interfaces, wet interfaces due to many cavities, etc.).

Mobility of the interaction

The flexibility of the residues at the protein surface together with the participation of water molecules help to define the shape and physicochemical complementarities of the binding partners, which determine the specificity and stability of the association. In proteins, the accessibility to solvent generally correlates with the mobility of atoms expressed in crystallographic temperature B-factors [129].

We calculated the temperature B-factor's mean value for all interacting residues (dry, dual, and wet spots; other authors have studied B-factors of interfacial waters [68]) and we compared obligate and transient complexes to account for systematic differences (Fig 4.5). We can appreciate that interfacial residues of obligate complexes have lower temperature B-factor averages than transient. A general observation for both complex types is that dual residues are less mobile than the dry ones. This observation is in agreement with molecular dynamics studies of buried waters in globular proteins [130].

Wet spots are comparable to dry residues in obligate complexes in terms of mobility, although in transient complexes wet spots are more similar to dual residues.

Interacting waters are likely to form multiple hydrogen bonds with the neighboring atoms. We checked the interacting behavior of the waters forming wet spots and compared it with the behavior of the other interacting waters of the interface. We computed the number of bridges that any water is doing, and grouped waters according to this number. In Figure 4.6, we can appreciate that waters mediating wet spots form preferably one but also two bridges, which is a similar behavior to the observed for the other interacting waters (data not shown).

Structural composition of interfacial residues

We analyzed the secondary structure composition of all interfacial residues (dry, dual, and wet spots) of our dataset. Based on our analysis, we could observe that in general interfacial residues are located predominantly in unstructured regions and prefer side-chains to interact (Fig 4.7). These suggestions have also been proposed in previous studies [131]. α-Helices and β-sheets have been described to occur infrequently at interfaces by Ansari and Helms in their study of transient complexes [124]. We observe that both, obligate and transient complexes have a comparable preference for α-helix and unstructured side-chain interactions. On the other hand, in transient complexes we observe a higher participation of the β-sheet side-chain interactions than in obligate. Dry and dual residues show similar frequencies in both complexes and interaction types. However, wet spots interact more frequently by the main-chain. While wet spots interacting by the main-chain tend to be located in unstructured regions, we do not observe such strong nonstructured preferences for side-

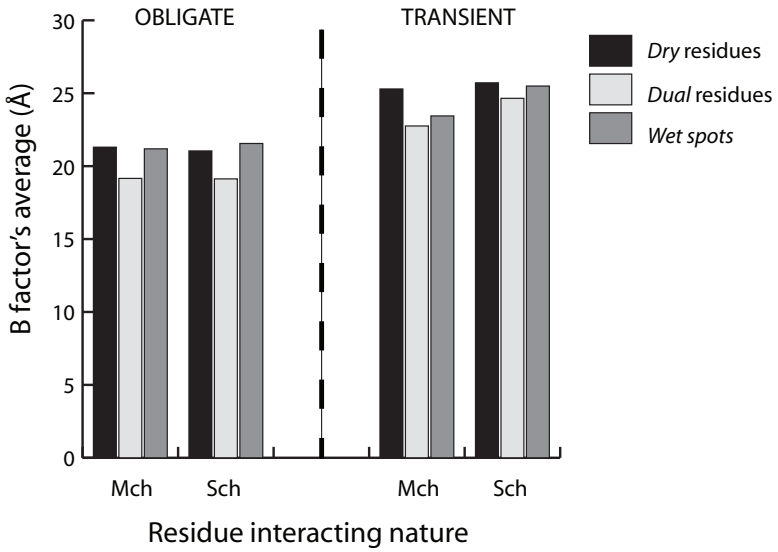


Figure 4.5: B-factors of interfacial protein residues. Average temperature B-factors are shown for dry, dual, and wet spots. Residues are grouped according to their interacting nature (Mch, main-chain; Sch, side-chain).

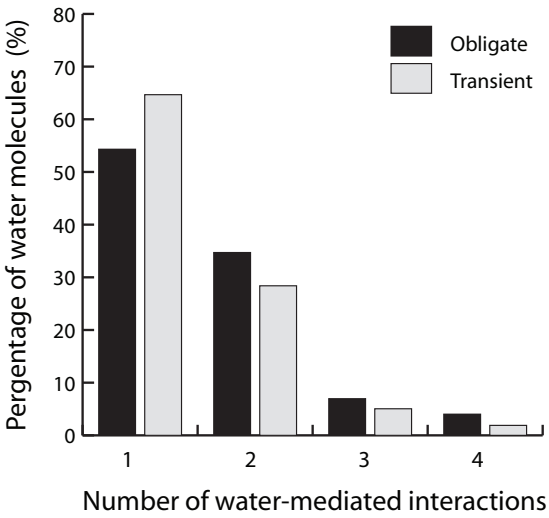


Figure 4.6: Multiplicity of waters forming wet spots. The number of interactions that waters mediating wet spots are forming is represented for transient and obligate complexes.

chain interactions. The active participation of unstructured regions in the formation of water-mediated interactions can be explained because residues in α-helix and β-sheets form systematic main-chain hydrogen bonds. Main-chain atoms in unstructured regions often fail to form direct intramolecular hydrogen bonds, being available to interact with water molecules.

Interestingly, Park et al. have recently described that residues interacting with buried waters in the core of globular proteins are predominantly located in unstructured regions and prefer main-chains to interact [130]. The fact that we observe similar features in wet spots makes us to suggest that they might play similar structural roles as the residues interacting with waters in protein cores and cavities. This applies for interfaces of obligate complexes, which mimic protein cores due to their coupled fold-binding process; whereas in transient complexes wet spots features could be explained by the involvement of unstructured regions (i.e. loops) in protein recognition.

Amino acid composition of interfaces

We examined the amino acid composition in all interfacial interactions of our dataset (Fig 4.8.A-C). While dry are interacting mainly by the side-chains of hydrophobic residues (obligate: 72.2%, transient: 74.6%), dual prefer the side-chains of hydrophilic residues (obligate: 61.8%, transient: 67.3%). Moreover, the frequencies of the main-chains in dual are increased in comparison to the dry residues. These observations are due to the fact that residues with capabilities to form multiple contacts are more likely to complement direct with water-mediated interactions. We can appreciate that for dry and dual residues the composition slightly differs between transient and obligate complexes. In both, dry and dual, hydrophobic side-chains are preferred by obligate, whereas transient prefer hydrophilic (Fig 4.8.A,B). Obligate complexes have been reported to be more hydrophobic than transient [41, 42, 47]. Many transient complexes involved in signal transduction need to bind quickly and specifically and do not need to be stable over long periods and thus, to achieve specificity they make use of a higher rate of hydrophilic residues.

Wet spots show a different composition from dual residues (Fig 4.8.B,C), having a higher participation of the main-chains (dual: 38.2%, 32.7%, wet: 55.8%, 50.9%, for obligate and

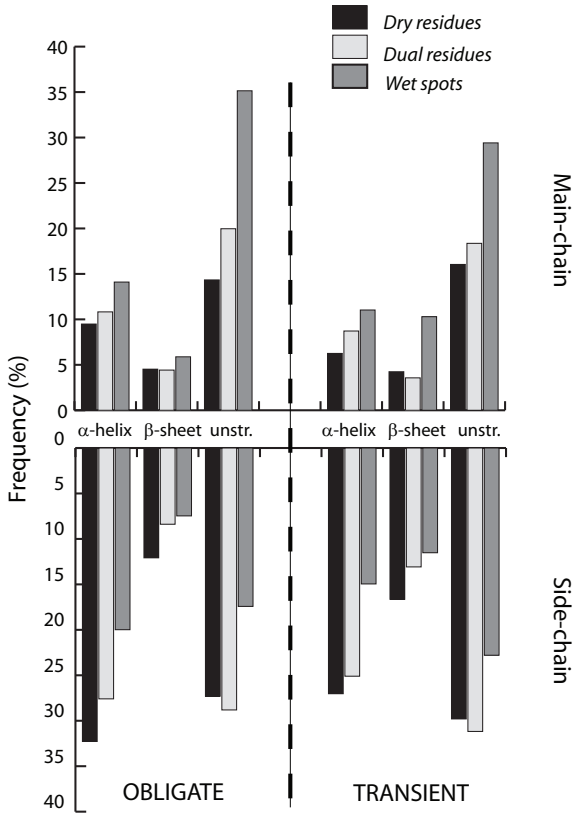


Figure 4.7: Secondary structure composition of interfaces. Distribution of the interacting residues in terms of secondary structure (α-helix, β-sheet, and unstructured). Information is shown for transient and obligate complexes, and divided into main- and side-chain.

transient, respectively). Wet spots exhibit a small preference for some short side-chains (i.e. Ser/Thr preferred to Tyr, and Asn preferred to Gln), opposite to dual residues. Moreover, Arg frequencies show quite a substantial decrease in wet spots. These differences between dual and wet spots in the residue length preference are consistent with the fact wet spots only interact mediated by water. Dual and wet spots have in common a remarkable preference for Gly, as well as a strong preference for Ala main-chain in obligate complexes.

Interfacial pairing preferences

We examined all residue-residue interfacial interactions of our dataset to analyze the pairing preferences of obligate and transient complexes, and we generated pair matrices based on the following interaction types: direct, dual, and only water-mediated (Fig 4.8.D-F). For direct

interactions, we appreciate in both complex types a preference for hydrophobic-hydrophobic and charged-charged pairing. Obligate complexes contain high frequencies of hydrophobic interaction pairs, mostly with Leu and Phe; whereas transient complexes have a more spread range of interacting pairs.

When we analyze the dual interacting pairs, we observe high frequency pairs with Arg, not only with hydrophilic but also with hydrophobic residues. Transient complexes present high interacting frequencies between charged residues due to their electrostatic component and their capabilities to form water hydrogen bonds. Arg also shows pairing preferences to hydrophobic residues, as it can form not only water hydrogen bonds with the main-chain of hydrophobic residues but also hydrophobic interactions with the aliphatic carbons of its side-chain. Obligate complexes have more diversity of residue pairs than transient (clearly reflected by much less white in the matrix).

Wet spots show more wide range of residue pairs and more diversity of high frequencies than the dual interactions. In the only water-mediated matrix, transient complexes show the highest peaks with Gly, Tyr, Asn, Phe, Ile, and charged residues. On the other hand, obligate residues contain a homogeneous frequency distribution only altered by the high interacting frequencies that Glu and Tyr have.

This statistical analysis of residue composition and pairing preferences by interaction type shows very interesting particularities about the role of water in protein interfaces. It comes into sight that obligate complexes tend to establish a broad range of main-chain interactions with solvent to complement direct hydrophobic interactions. This may give a more hydrophilic nature to obligate interfaces than initially thought when solvent is not included.

4.1.5 Conclusions

We present a detailed analysis and characterization of the interfacial solvent of a high-resolution dataset containing obligate and transient interfaces, carried out with the aim of assessing the role of the, often overlooked, water-mediated interactions in protein interfaces. Our analysis identifies 40.1% of the interfacial residues interacting through water and an average of five water molecules per 1000 $\mathring{A}^2$ mediating interactions in interfaces. Further-

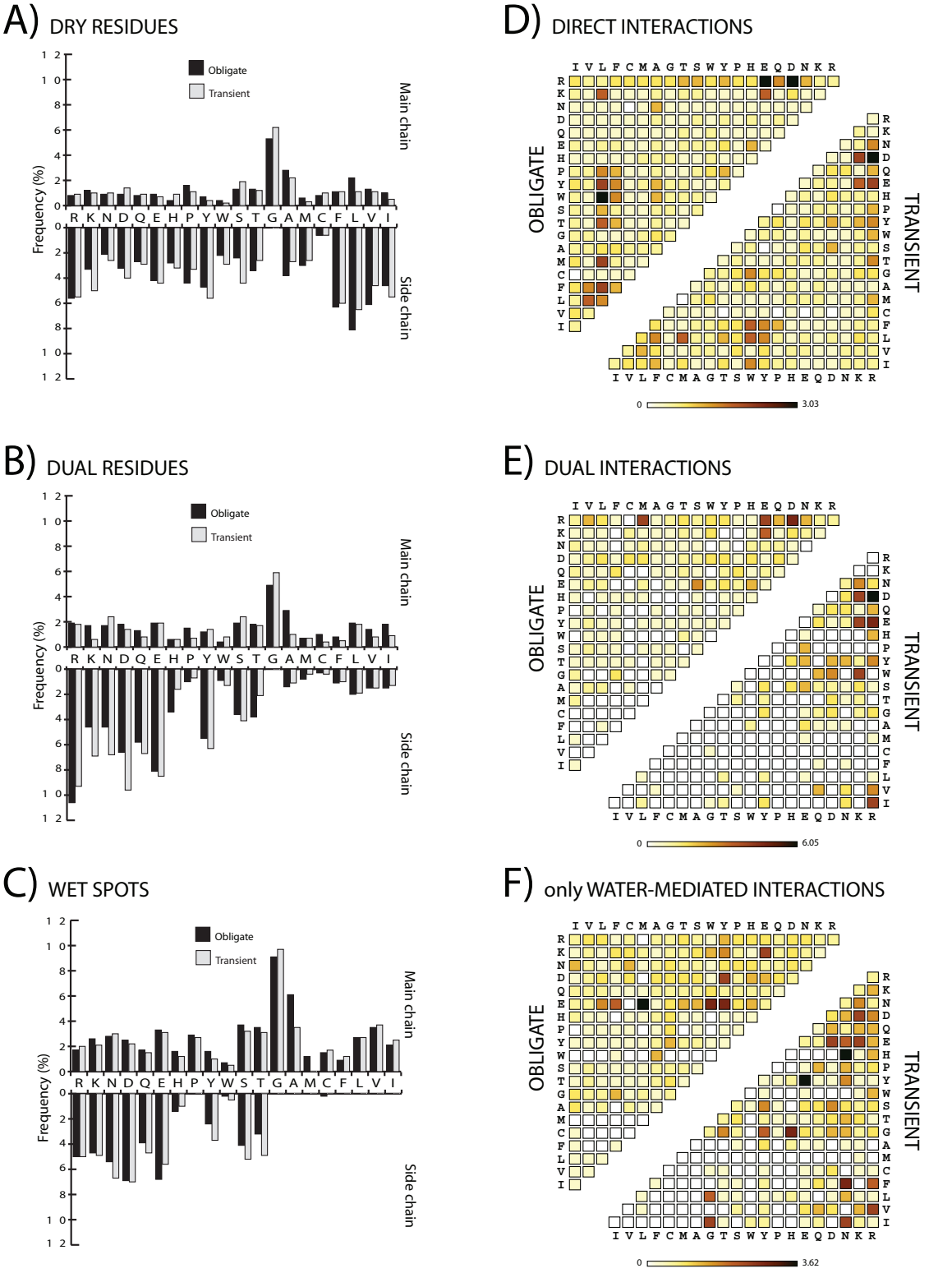


Figure 4.8: Amino acid composition and pairing preferences. Amino acid composition of (A) dry, (B) dual, and (C) wet spots. The color gradient used in the matrices represents the normalized pairing frequencies of two residues that occur in our dataset (see Materials and Methods). Pairing preferences for (D) direct, (E) dual, and (F) only water-mediated interactions. Residues are ordered according to the Kyte-Doolittle hydropathy index.

more, the inclusion of solvent allows the participation of 14.5% of the interfacial residues that only interact mediated by a water molecule called wet spots. The observed abundance of water-bridged residues in protein interfaces reinforces the relevance of including solvent in protein interaction studies. A thorough characterization of the interfacial information of our dataset underlines differences between water-bridged residues and wet spots. We observe that the preferred residues and interaction pairs vary whether the residue is already interacting directly (dual) or not (wet spot). Dual residues interact more frequently by the side-chain of long polar residues, whereas wet spots interact essentially by their main-chain and short polar side-chains. We find wet spots in different interfacial solvation patterns, but preferably located in nonregular secondary structures. Wet spots predominantly choose their main-chains to establish interfacial interactions and have Gly as the preferred residue. These characteristics are similar to those seen in residues binding buried water molecules in the core or cavities of proteins, which appear to perform the task of satisfying the unmet hydrogen bonding [130]. The abundance of wet spots together with their peculiar characteristics makes their contribution to interfacial description quite significant.

Moreover, we also analyze the solvent contribution in obligate and transient complexes based on fold considerations. Although the amount of solvent is comparable for both, we corroborate the existence of a high heterogeneity within each complex type as it has been described for homo- and hetero-complexes [68]. Protein complexes formed in a coupled folding-binding process (obligate) are expected to exhibit similarities to protein cores having a dry nature. However, although we observe that in average obligate interfaces contain less density of residues per 1000 $\mathring{A}^2$, the percentage of hydrated residues is slightly higher than that in transient interfaces. We also see that, as expected, protein complexes formed by two monomeric folded proteins contain a higher hydrophilic nature than obligate. The role of solvent in interfaces of transient complexes is to hydrate mainly the charged side-chains as they have more capabilities to form water hydrogen bonds, whereas in obligate complexes solvent tend to establish a broad range of main-chain interactions to complement the hydrophobic interactions forming the interface.

In summary, we provide an enhanced understanding of the role of interfacial solvent, and in particular, the contribution of wet spots, to the description of protein interfaces.

Wet spots may play a crucial role in binding affinity and specificity, although they are often discarded in computational protein interaction studies and rational drug design.

We believe that the information obtained in our studies will be useful to enhance descriptions of protein interfaces, to help in interface comparisons and to better understand molecular recognition. The consideration of solvent-mediated interactions in residue propensities and pairing preference matrices may be of great interest to improve the performance of empirical methods based on statistical protein structure information such as docking and potential energy functions for protein folding and ligand design.

4.2 A molecular dynamics approach to study the importance of solvent in protein interactions

by Sergey Samsonov, Joan Teyra, M. Teresa Pisabarro
in Proteins 2008, 73(2):515-525

4.2.1 Abstract

Water constitutes the cellular environment for biomolecules to interact. Solvent is important for protein folding and stability, and it is also known to actively participate in many catalytic processes in the cell. However, solvent is often ignored in molecular recognition and not taken into account in protein-protein interaction studies and rational design. Previously we developed SCOWLP, a database and its web application (http://www.scowlp.org), to perform studies on the contribution of solvent to protein interface definition in all protein complexes of the PDB. We introduced the concept of wet spots, interfacial residues interacting only through one water molecule, which were shown to considerably enrich protein interface descriptions. Analysis of interfacial solvent in a non-redundant dataset of protein complexes suggested the importance of including interfacial water molecules in protein interaction studies. In this work we use a molecular dynamics approach to gain deeper insights into solvent contribution to protein interfaces. We characterize the dynamic and energetic properties of water-mediated protein interactions by comparing different interfacial interaction types (direct, dual and wet spot) at residue and solvent level. For this purpose, we perform an analysis of 17 representative complexes from two protein families of different interface nature. Energetically wet spots are quantitatively comparable to other residues in interfaces, and their mobility is shown to be lower than protein surface residues. The residence time of water molecules in wet spots sites is higher than of those on the surface of the protein. In terms of free energy, though wet-spots-forming water molecules are very heterogeneous, their contribution to the free energy of complex formation is considerable. We find that water molecules can play an important role in interaction conservation in protein interfaces by allowing sequence variability in the corresponding binding partner, and we discuss the important implications of our observations related to the use of the corre-

lated mutations concept in protein interactions studies. The results obtained in this work help to deepen our understanding of the physico-chemical nature underlying protein-protein interactions and strengthen the idea of using the wet spots concept to qualitatively improve the accuracy of folding, docking and rational design algorithms.

4.2.2 Introduction

Water plays an extremely important role in all biological processes because of its unique physical and chemical properties. It represents not only an environment for interacting components of biochemical reactions but it is also an active participant. Without a critical level of hydration proteins are not functional, and water molecules presence is crucial in catalytic sites of many enzymes [132]. It has been shown that water molecules can be structurally conserved in protein complexes, and that their residence time and diffusion characteristics are distinct from bulk and surface solvent [133–135]. Thermodynamically, water molecules can contribute favorably to protein complex formation [136, 137]. Computationally, the inclusion of water in the Hamiltonian of protein systems has improved folding predictions compared to *in vacuo* folding models [138]. It is widely accepted that exclusion of water molecules from the proteins contact area in the process of complex formation is associated with a free energy decrease. The entropic component increase due to the transfer of water molecules into bulk solvent is considered to have the decisive energetic impact. This entropy gain is usually thought to exceed the corresponding enthalpy loss [139]. Despite all, solvent is often ignored in the analysis of protein-protein interactions.

The protein interface concept is very important in the description of protein-protein interactions. Protein interfaces could be defined differently depending on the criteria introduced for the cut-off for interacting atoms of the complex counterparts [140]. In our previous work we have developed SCOWLP, which, taking into account interfacial solvent, classifies all interfacial protein residues of the PDB into three classes based on their interacting properties: dry (direct interaction), dual (direct and water-mediated interactions), and wet spots (residues interacting only through one water molecule) [70]. Also in our preceding studies, statistical analysis of a nonredundant protein structure dataset showed that 40.1% of the interfacial residues participate in water-mediated interactions, and that 14.5%

of the total residues in interfaces are wet spots. Moreover, wet spots have been shown to display similar characteristics to residues contacting water molecules in cores or cavities of proteins [113]. This suggests that water-mediated interactions should not be disregarded in a detailed definition of protein interfaces. Our and other studies have revealed that water-mediated interactions are highly heterogeneous [42, 68, 113], making protein-protein interactions studies challenging. In fact, certain aspects of water-mediated interactions still remain unclear.

This study aims to gain insights into solvent contribution to protein interactions. Our focus is the contribution of wet spots to protein interfaces in comparison to other interfacial residues. For this purpose, we use a molecular dynamics (MD) approach to characterize dynamic and energetic properties of wet spots and the water molecules forming them. We pay special attention to the role of water molecules in interaction conservation in protein interfaces.

For our studies we use representatives of two protein families of different physico-chemical interface nature and interface size in complex with other proteins and peptides: Src-homology 3 domains (SH3) and immunoglobulin domains (Ig). SH3 are small recognition domains comprising 5 antiparallel β-strands and widely presented in signaling pathways [141]. Immunoglobulin domains are bigger and comprise 7 to 9 antiparallel β-strands [142]. Ig interfaces are more hydrophilic than SH3 domain interfaces.

The results of this study show that water-mediated interactions are similar in both protein families, and that the dynamic and energetic characteristics of wet spots and dual residues are comparable to dry interfacial residues. Interfacial water molecules display properties such as residence time and mobility more similar to water molecules in cores and cavities of proteins than to bulk or protein surface solvent. Our findings emphasize the important role of water contributing to interaction conservation of protein interfaces and strongly imply the significance of including interfacial solvent in detailed protein interface descriptions.

Table 4.3: Complexes dataset

PDBid	Family	Resol($\AA$)	PP/Pp	Short description
1UJO*	SH3	1.70	Pp	Signal transduction adaptor molecule-2(STAM-2). SH3 domain with peptide derived from deubiquitination enzyme (UBPY)
1BBZ	SH3	1.65	Pp	Abl Tyrosine Kinase SH3 domain with p40 synthetic peptide
1FYN	SH3	2.30	Pp	Fyn Tyrosine Kinase SH3 domain with 3BP-2 synthetic peptide
1OEB	SH3	1.76	Pp	Grb2-like adaptor protein Mona/Gads. SH3 domain with the peptide derived from T-cell receptor signal transducer SLP-76
1B07	SH3	2.50	Pp	Proto-oncogene Crk2 (Serine-Threonine kinase) SH3 domain with peptoid inhibitor
1UTI	SH3	1.50	Pp	Grb2-like adaptor protein Mona/Gads with the peptide derived from Hematopoietic Progenitor Kinase 1 (Hpk1)
1ABO	SH3	2.00	Pp	Abl Tyrosine Kinase SH3 domain with 3BP-1 synthetic peptide
1OPK	SH3	1.80	PP	Mouse Abl Tyrosine Kinase (SH3-PI)
2SRC	SH3	1.50	PP	Human Src Tyrosine Kinase (SH3-PI)
1AVZ*	SH3	3.00	PP	Fyn Tyrosine Kinase SH3 domain with HIV1-Nef protein
1QCF	SH3	2.00	PP	Human Hck Tyrosine Kinase (SH3-PI)
1SM3	Ig	1.95	Pp	Tumor specific Fab fragment with its peptide epitope
1QKZ*	Ig	1.95	Pp	Fab fragment with bacterial antigen
1EJO	Ig	2.30	Pp	Monoclonal 4C4 Fab fragment with G-H loop from virus FMDV
1G7I	Ig	1.80	PP	Monoclonal Fab fragment with Hen Egg White Lysozyme
1JPS	Ig	1.85	PP	Fab D3h44 fragment with tissue factor
1LK3	Ig	1.91	PP	Fab 9D7 fragment with engineered IL-10

PP, protein-protein complex; Pp, protein-peptide complex. (*) Soft harmonic force restraints (2 kcal/mol) were applied on C of the ligand to keep it in the binding site during the simulation.

4.2.3 Methodology

Protein complexes dataset

The following criteria for protein complex selection was applied: resolution ≤ 2.5 $\AA$ and existence of wet spots (as they are annotated in SCOWLP [70], Table 4.3). The complex 1avz (FYN SH3 domain with HIV1-Nef) was also taken in our dataset because of its biological relevance. Structural alignments of the proteins were done with the MAMMOTH algorithm [80].

Molecular dynamics simulations

MD simulations were carried out with the AMBER 8.0 package. All hydrogen atoms were added using the Xleap tool. Standard ff03 force field parameters were used. Parameters for phosphorylated residues (complexes 1opk, 2src, and 1qcf) were taken from the phosphorylated amino acids library set [143], and parameters for the N-substituted glycine residue (complex 1b07) were derived in the Antechamber module of AMBER 8.0. Each complex was solvated in a truncated octahedron periodic box filled with TIP3P water molecules and neutralized by counterions. MD simulations were preceded by two energy-minimization steps: 500 cycles of steepest descent and 1000 cycles of conjugate gradient with harmonic force restraints on protein atoms, then 1000 cycles of steepest descent and 1500 cycles of conjugate gradient without constraints. This was followed by heating of the system from 0 to 300 K for 10 ps, and a 30 ps MD equilibration run at 300 K and 10^6 Pa in isothermal isobaric ensemble (NPT). Following the equilibration procedure, 10 ns of productive MD runs were carried out in periodic boundary conditions in NPT ensemble with Langevin temperature coupling with collision frequency parameter $\gamma = 1$ ps^{-1} and Berendsen pressure coupling with a time constant of 1.0 ps. The SHAKE algorithm was used to constrain all bonds that contain hydrogen atoms. A 2 fs time integration step was used. An 8 $\mathring{A}$ cutoff was applied to treat nonbonded interactions, and the particle mesh ewald (PME) method was introduced for long-range electrostatic interactions treatment. MD trajectories were recorded each 2 ps. For the analysis of the trajectories PTRAJ module was used.

Trajectory processing

We defined interfacial interactions based on physico-chemical and distance criteria between atoms. For hydrogen bonds, we considered a donor-acceptor distance of 3.2 $\mathring{A}$, for salt bridges 4 $\mathring{A}$, and for van der Waals interactions the van der Waals radii distance. Three classes of residues were introduced: dry (direct interaction), dual (direct and water-mediated interactions) and wet spots (residues interacting only through one water molecule) [70]. Each frame of the trajectory was processed so that the relative time fractions (TFs) of total, dry, dual and wet spot interactions (TF_T, TF_D, TF_d, and TF_{ws}) during the simulation were corresponded to each residue. The total interaction was defined as a sum of all three defined

interaction types. A residue was considered interacting if the total time of interaction was at least 5% of the simulation time. A residue was considered to be a wet spot if it interacted only through a single water molecule more than 10% of the simulation time. Such cut-offs were chosen arbitrarily in order to consider the wide range of the interactions for analysis and to restrict the definition of wet spots in MD to a certain intuitively significant value.

Effective interface area calculations

The area of interface is usually defined as the difference between solvent accessible areas for the unbound molecule and for the same molecule in complex. We introduced effective interface areas related to water-less (ΔASA$_{wl}$) and water-mediated (ΔASA$_{w}$) interactions in order to estimate the impact of water-mediated interactions on the interface definition. The introduction of effective interface areas allows to consider that during the simulation the same interfacial residue could belong to dry (D), dual (d), and wet spots (ws) residue class for certain respective times:

$$\Delta ASA_{wl} = \sum_i \Delta ASA_i \left(TF_{D,i} + \frac{1}{2} TF_{d,i} \right) \tag{4.2}$$

$$\Delta ASA_{w} = \sum_i \Delta ASA_i \left(TF_{w,s} + \frac{1}{2} TF_{d,i} \right) \tag{4.3}$$

where TF$_{D/d/ws}$, i are the relative time fractions of residue i; ΔASA$_i$ is the accessible surface area of the ith residue calculated in the NACCESS program with a standard water probe radius of 1.4 $\AA$ [25].

Fluctuation analysis

The average fluctuation (F) for each interfacial residue was obtained with the PTRAJ module of AMBER 8.0 as a mass-weighted sum of fluctuations for atoms belonging to this residue. To implicitly decompose the impact of each type of interaction (total, dry, dual, and wet spots) on the fluctuation as an analytically unknown function F(TF$_T$, TF$_D$, TF$_d$, and TF$_{ws}$), the following method was used. The function values were averaged regarding to all other TFs except for the one of interest in order to obtain dependence on this certain TF: $\langle$F(TF$_{ij}$, TF$_{kj}$ $\geq$ a)$\rangle_i$, where i contains all interaction types (T = total, D = dry, d

= dual, ws = wet spot) except k ($i \neq k$), k is the TF of interest, j is the summing index for interfacial residues and $a \in$ [0,100] expressed in %. Dependences on these 4 TFs were compared qualitatively.

MM-GBSA free energy decomposition per residue

Energetic post-processing of the trajectories was done in a continuous solvent model as implemented in the AMBER 8.0 MM-GBSA (Molecular Mechanics-Generalized Born Surface Area) module. MM-GBSA is a method for free energy calculation utilizing implicit solvent model and is based on a Generalized Born approximation to the exact (linearized) Poisson-Boltzmann equation for electrostatics. The snapshots for the calculations were chosen as described by Lafont et al. [144]. To achieve better conformational space sampling, first, all frames of the trajectory were sorted by *in vacuo* calculated electrostatic energy values and the range of these energies was divided into 10 equal intervals. For each interval the number of corresponding conformations was calculated and served as a weight function for the interval. Then, for a conformation, most closely corresponding to the interval mean value of electrostatic energy, full MM-GBSA energy calculations were carried out. The final result was calculated as a weighted sum of values for each interval. The energy components per residue were compared by TFs (i.e., TF_T, TF_D, TF_d, TF_{ws}) in a similar way as it is described in the Fluctuation analysis section.

Residence time analysis of water molecules

The distance from the interacting heavy atoms of each wet spot counterpart to water molecules in each frame was calculated using the PTRAJ module of AMBER 8.0. If the distance did not exceed 3.6 $\mathring{A}$, the wet spot site was considered to be occupied. A surface water site was defined by the volume that was closer than 3.6 $\mathring{A}$ to one of the protein polar groups and was not located in an interface. Solvent was considered as bulk at a distance ≥ 5 $\mathring{A}$ from the protein surface. Surface and bulk sites are defined so that their total occupancy is 100%. This makes them *a priori* more occupied in comparison to interfacial sites, where total occupancy is 64% on average. Therefore, when differences between surface, bulk and interfacial sites exist, conclusions can be made even stronger. The frequency

of consecutively occupied frames for the site was presented as residence time distribution density. Maximum number of consecutively occupied frames for the site was corresponded to maximum residence time (T_{max}) of a water molecule in the site, and total occupancy was defined as the sum of all time intervals when the site was occupied.

Free energy perturbation calculations

For free energy calculations of water molecules in wet spot sites, the double decoupling method of free energy perturbation as described in the work of Hamelberg and McCammon was used [145]. This method is based on two steps of perturbation. First, electrostatic interactions are gradually turned off; then van der Waals radii of chargeless atoms are decreased to 0. The free energy difference between two states was calculated using the thermodynamic integration approach at discrete points of the coupling parameter λ, which was variated from 0 to 1 and then back from 1 to 0 with a 0.01 step along the path. Simulation for each λ value was equilibrated for 10 ps followed by a productive MD sampling for 10 ps. In the case of two water molecules in the same spot, the less mobile one was first removed. If the removal energy of the first water was negative (favorable), and the two waters were establishing hydrogen bond interactions in the site, both waters were removed from the spot at a time. AMBER prevents other water molecules from occupying the site of the perturbed one by considering the volume corresponding to the site of the perturbed water molecule as occupied by default.

Statistical analysis

Statistical analysis of data was carried out with the R-package [146].

4.2.4 Results and discussion

We have performed MD studies to deepen our understanding of the properties of protein interfacial residues (direct, dual, and wet spots) and interfacial solvent. In this work we have studied protein interactions in terms of mobility, free energy and interaction conservation. For this purpose we have analyzed water-mediated interactions in a representative set formed by 11 complexes of SH3 domains (seven with peptides and four with proteins) and six

Table 4.4: Summary of interface properties

Domains	**SH3**	**Ig**
Number of interacting residues per domain	14 ± 2	24 ± 5
Interface area ($\AA^2$)	733 ± 195	1291 ± 471
Interacting residues/1000 ($\AA^2$)	19	19
Total observed wet spots in MD (SCOWLP)	49 (15)	61 (19)
Wet spots in MD (SCOWLP)/complex	4.5 (1.4)	10 (3.2)
Wet spots in MD (SCOWLP)/1000 ($\AA^2$)	6.1 (1.9)	7.7 (2.5)
Wet spots in MD (SCOWLP)/interface residue	0.32 (0.10)	0.42 (0.13)
Effective areas ratio $\Delta ASA_w/\Delta ASA_d$, (%)	28 ± 7	39 ± 13

complexes of Ig domains forming variable antibody fragments (three with peptides and three with proteins) (Table 4.3). A total of 292 interfacial residues including 110 wet spots were analyzed.

Interaction patterns in MD simulations

MD runs of 10 ns were performed for each complex. To define interfacial residues the trajectories were processed and the corresponding TFs were calculated. A summary of properties averaged over all structures of a family for the interface description of the two representative protein families used in this study is presented in Table 4.4. Immunoglobulin interfaces are almost twice bigger than SH3 domain interfaces, both in number of residues and in area size. However, the density of interactions (number of interfacial residues per area unit) is the same for both families. In both families the number of wet spots observed in MD simulations is about three times higher than in the corresponding PDB initial structures. This could be explained, first of all, because of the different nature of the source of information (obtained from PDB files, static; defined in MD, dynamic), and partly by resolution and quality of data contained in PDB files. Classification of interfacial residues based on X-ray data falls into three classes (direct, dual, and wet spots), while in MD this discrete division is not possible due to the fact that the same residue may present different interacting modes during the simulation. A continuous model of residue interaction pattern requires more

complicated, often implicit, not direct mathematical approaches for analysis. Table 4.4 illustrates roughly same ratios for the parameters for the two representative domain families though Ig interfaces have higher relative number of wet spots than SH3 domains interfaces. The last displayed parameter in Table 4.4 corresponds to the ratio of effective areas of dry and water-mediated interfaces (see Methods section). These ratios reflect how larger the area of the interfaces would be if we would include wet spots in the interface definition. We obtained roughly 30 and 40% of interface size increase for SH3 and Ig, respectively. Considering the importance of the interfacial area as an empirical parameter in algorithms implemented for energy calculations, these numbers suggest that exclusion of the water molecules from protein interface analysis may lead to significantly biased or incomplete results. In terms of total interactions per residue, despite the differences in size and chemical nature, both SH3 and Ig interfaces have comparable averaged contribution of interfacial residues to each class (Fig 4.9). The *t*-test shows only a significant difference for the wet spots impact (at the level of *P*-value = 0.05). Although the total occurrence of wet spots interactions is about three times lower than of direct interactions, overall water-mediated interactions correspond to almost the same TF as direct interactions. The percentage of interactions in our complexes agrees with the 14% presence of wet spots in total interfacial residues obtained for a nonredundant dataset of protein complexes [113]. That means that, in terms of wet spots contribution, both SH3 and Ig families are close to average protein families. Direct and water-mediated interactions reveal differences in distributions of TFs (Fig 4.10). Direct interactions are almost uniformly distributed on all time fraction intervals, while dual and wet spots interactions are distributed mostly on intervals up to 30% of relative interaction time. At the same time, the distribution of total interactions shows that most of the residues are interacting during more than half of the simulation. Comparison of the distributions suggests that there are few interfacial residues forming wet spots interactions for a long time during simulations. However, the contribution of wet spots to total interaction is substantial (Fig 4.9). The analysis of the distributions shows that it is not correct to consider that an interfacial residue unambiguously belongs to only one class of interfacial residues. We monitored wet spots in the MD simulations of the SH3 and Ig domains, and classified them by interaction type (main-chain/side-chain) as well as by amino

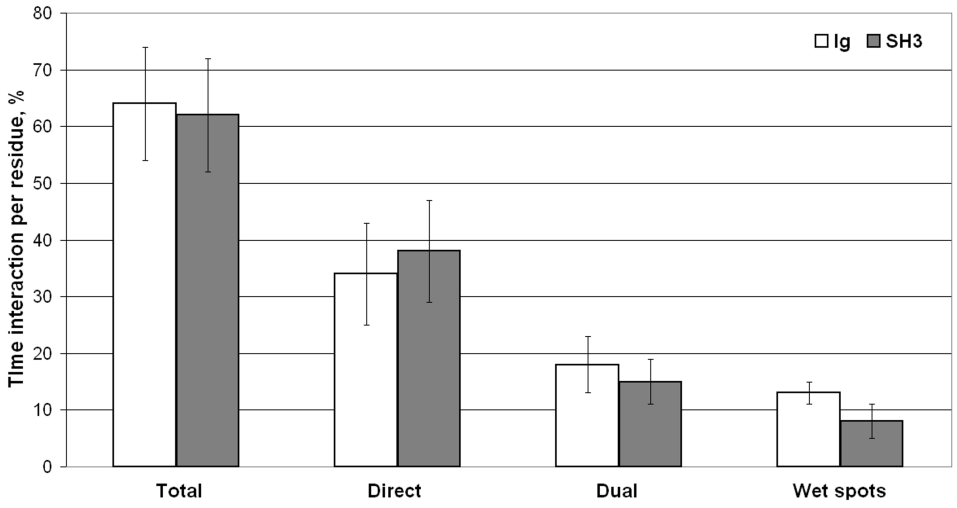


Figure 4.9: Average time fractions of interaction in the Ig and SH3 complexes.

acid composition (Fig 4.11). For both families side-chain interactions slightly prevailed (55% of all wet spots), which agrees with previous results obtained for transient complexes based on crystallographic data [113]. Wet spots show in general a strong preference for polar and negatively charged residues, mostly interacting through their side-chains. At the same time, water mediation increases the participation of hydrophobic amino acid residues through their main-chains in the formation of interfaces with a certainly less hydrophobic character.

Interaction conservation through water

Water plays a role of molecular glue in protein interfaces [59]. Water may mediate conserved interactions in protein families and thus participate in specificity. Certain water-mediated interactions can be present in most of the interfaces of a protein family (e.g., Asn52 in SH3 domains; Fig 4.12). In addition, water may also keep the interactions conserved despite the introduction of semi- or nonconservative mutations in a specific position in a protein family (e.g., sites I, IV, and V; Fig 4.12). Correlated mutations in binding partners are expected to appear as a result of coevolution and aim to keep a specific interaction conservative [147, 148]. However, due to the participation of water in protein interfaces conservation of

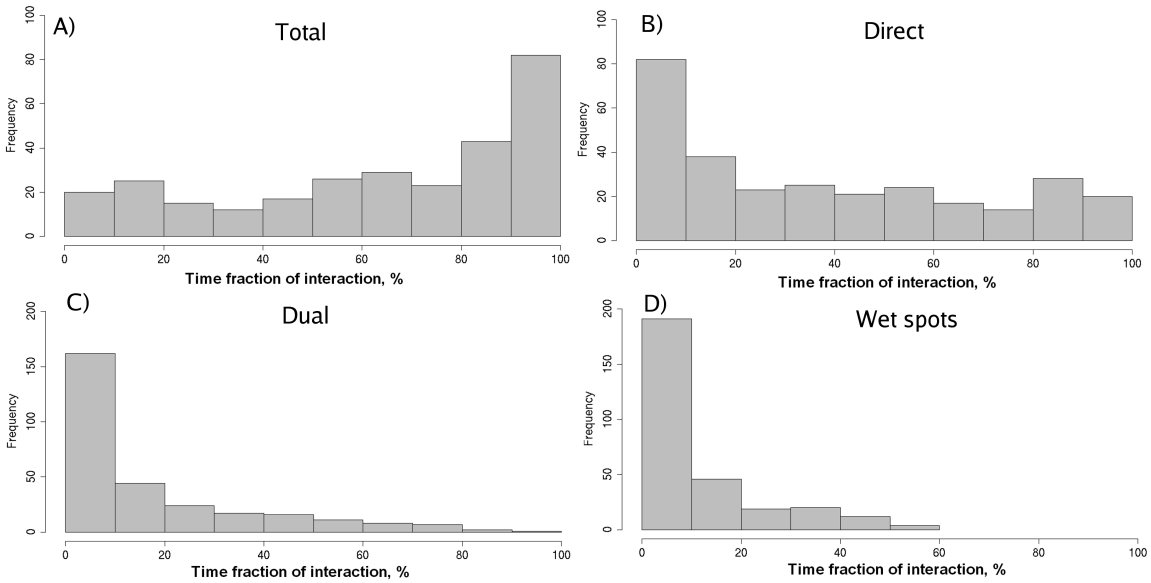


Figure 4.10: Distribution of time fractions of interaction for all simulated complexes.

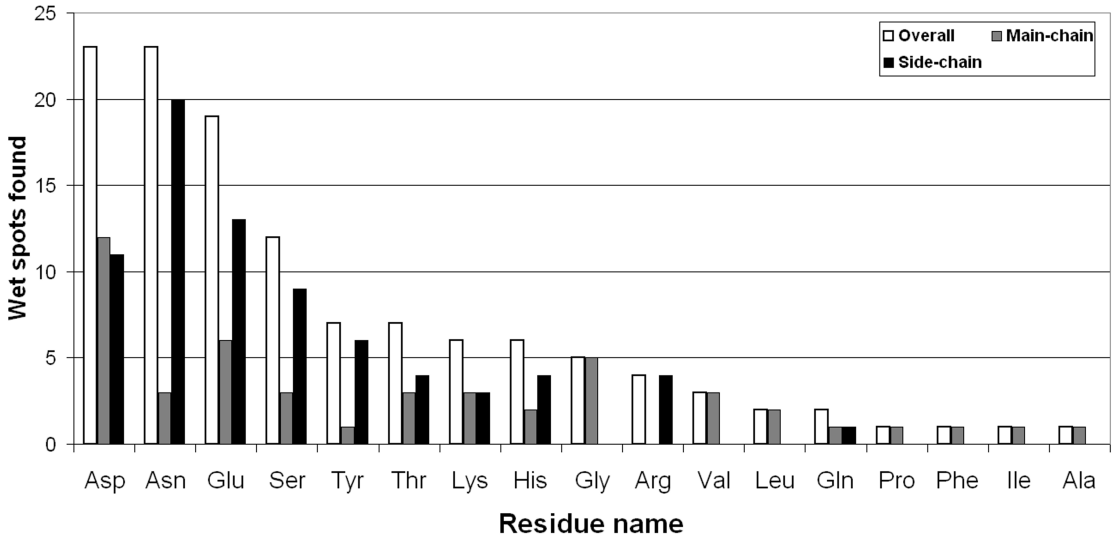


Figure 4.11: Participation of different residues in wet spots.

an interfacial interaction may occur despite non-correlated mutations (Table 4.5). Examples of interaction conservation through water found in SH3 domains are graphically shown in Figure 4.13. Figure 4.13.A illustrates site I with no correlation between the mutations in protein and ligand. The conserved interaction Glu(SH3)-Arg(ligand) is replaced in one of the complexes (1bbz) by the interaction Val(SH3)-Tyr(ligand). The interaction formed by the side-chain of different ligand residues with the protein is conserved due to the establishment of water-mediated main-chain interactions. Figure 4.13.B illustrates site III with direct interacting residues being replaced by wet spots. The conserved interaction between side-chains of ligand and side-chains of the protein is maintained in one of the complexes with a non-conservative mutation (1fyn) thanks to the establishment of a water-mediated main-chain interaction. The concept of correlated mutations in protein-protein interaction studies was introduced in the 90s [147, 148] and has been used since then to optimize protein design, predictions of protein interfaces and docking algorithms. Several matrices of residue-pairwise interacting probabilities have been built using different mathematical approaches and empirical parameters [149–151]. However, none of them considers solvent as mediator of interactions. Our results indicate that disregarding interfacial solvent may cause inaccuracies in the application of correlated mutations based approaches in the complete analysis of protein interfaces and the prediction of protein interactions.

Fluctuation analysis

In previous work we showed that thermal B-factors of wet spots are comparable to those of other interfacial residues [113]. Prior to checking if the mobility properties of different interfacial residue classes could be distinguished, we compared the mobility of surface residues and interfacial residues in terms of average fluctuations. Our results show that fluctuations of interfacial residues (side-chain and also full residue) are significantly lower than those of surface residues (at the level of t-test P-value $= 0.05$), while there is no significant difference for backbone fluctuations. Implicit decomposition of the average fluctuation function calculated for all interfacial residues in the studied complexes shows that, in general, residue fluctuations decrease with the increase of residue interaction time (Fig 4.14). Dual residues fluctuate more than dry and less than wet spots. At the same time the residues in the

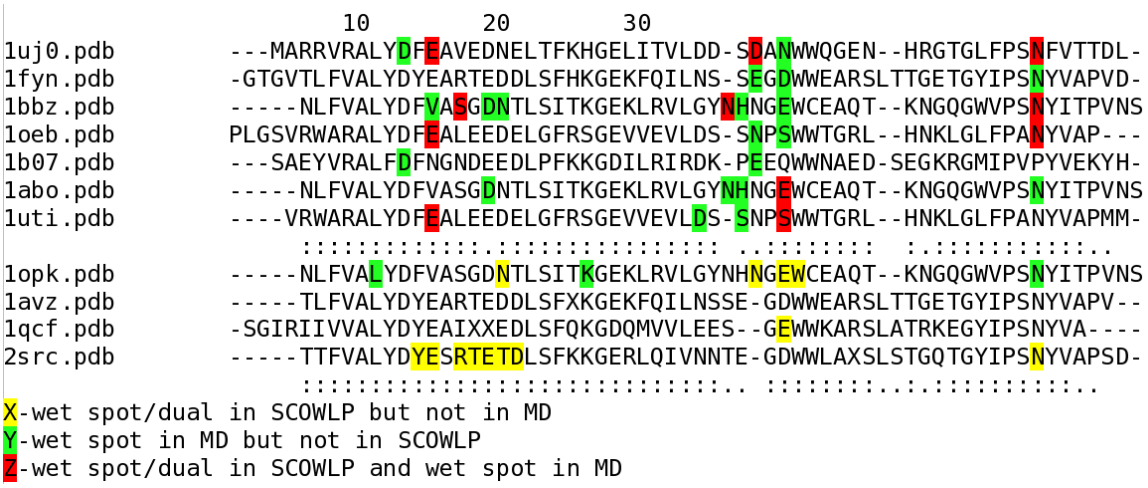


Figure 4.12: Structure-based sequence alignment of SH3 domains. Residues are colored by their participation in wet spots. The position of Asn52 (numbering by 1uj0) is labeled with an asterisk. Interaction sites from Table 4.5 are labeled with roman numbers at the top of the alignment.

Table 4.5: Examples of Interaction Conservation in SH3 Domain Interfaces.

Site	PDB ID	SH3	Ligand	Site	PDB ID	SH3	Ligand
I	1uj0	E12m	R64s	IV	1uj0	D34s	N66m
	1bbz	V10m	Y63s		1oeb	N37s	T67m
	1oeb	E15m	R65s		1uti*	N33s	E68m
	1uti	E11m	R66s	Va	1uj0	N36s	N66s
II	1bbz	S12s	Y63s		1oeb	S39s	T67m
	1qcf	I16m	E175s		1uti	S35s	E68s
	1fyn*	R16s	Y66s	Vb	1fyn	N38s	Y66m
	1uj0*	V14s	R64s		1abo	E35s	M62m
	1oeb*	L17s	R65s	Vc	1bbz	E35s	P65m
	1b07*	N14s	R68s		1avz	D237s	L106m
	1uti*	I13s	R66s		1qcf	R35s	W174m
III	2src	D16s	I172m		1b07*	R36s	P67m
	1uj0*	E18s	R64s				
	1fyn*	D20s	Y66s				
	1bbz*	T16s	Y63s				
	1uti*	E17s	R66s				

s, side-chain; m, main-chain interactions; *, direct interaction; Va, Vb, Vc correspond to different interactions in the same site

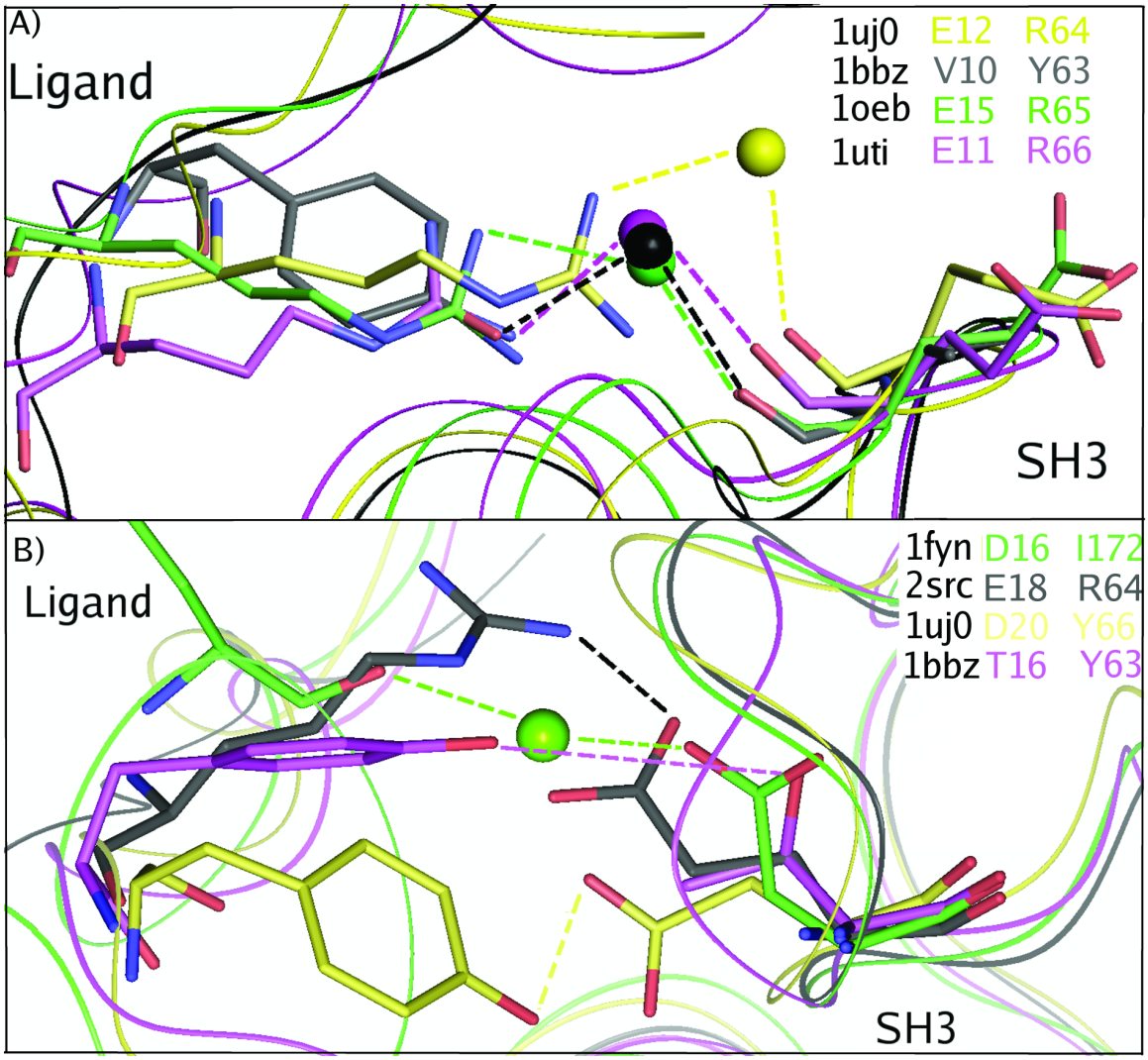


Figure 4.13: Examples of interfacial interaction conservation through water in SH3 interfaces. (A) Site I of Table 4.5 with no correlation between the mutations in protein and ligand. (B) Site III of Table 4.5 with direct interacting residues being replaced by wet spots. Proteins and ligands are represented by ribbons and labeled. Interacting residues are shown in sticks, and water molecules as spheres. The color code of the sequences in the upper right panels corresponds to the colors of the residues in the proteins and ligands as well as in the water molecules. Hydrogen bonds are represented with dash lines.

protein interior were shown to be significantly (at the level of *t*-test *P*-value = 0.05) less mobile than those in the interface or surface. For example, for the 1uj0 complex, the average fluctuations for protein interior, interfacial and surface residues were 0.50 $\mathring{A}$ ± 0.06 $\mathring{A}$, 0.74 $\mathring{A}$ ± 0.30 $\mathring{A}$ and 1.00 $\mathring{A}$ ± 0.48 $\mathring{A}$, respectively. The fluctuation analysis of 111 surface and 151 interfacial residues also revealed significant difference (at the level of *t*-test *P*-value = 0.05) between these two groups of protein residues (1.20 ± 0.61, 0.83 ± 0.59, respectively). Our data agree with the thermal factor analysis performed on a large dataset of protein complexes, which found that the closer the residues are to the core of interfaces, the higher their stability [152]. Since the fluctuations of wet spots and surface residues at temperatures higher than 180 K could be roughly explained in terms of surrounding water molecules mobility [153], it suggests that water flow around wet spots is slower in general than water molecules motion in the surface hydration shell. A similar trend was obtained for the interfacial residues participating in water-mediated interactions in another study where the speed of surrounding water flow and the mobility of the residues were analyzed in MD simulations [140]. Therefore, dynamical properties of interfacial residues and solvent mediating interaction are tightly interconnected, and they could be mechanically described as a coupling of harmonic oscillator to solvent modes via small springs (hydrogen bonds) [154]. This means that dynamic analysis of the interfacial residues might be biased without the consideration of surrounding water molecules.

Free energy decomposition per residue in interfaces

The MM-GBSA method applied for free energy decomposition calculations per residue allows to obtain the following independent components describing the energetics of a protein complex in implicit solvent: electrostatic component *in vacuo*, van der Waals interactions component *in vacuo*, Generalized Born reaction field energy [155] and hydrophobic component of solvation. The differences in energy values for interfacial residue classes were compared with the characteristic thermal motion energy value at 300 K (RT $\sim$ 0.6 kcal/mol). The differences in hydrophobic component of solvation were lower than this value, so we considered that this component does not differ significantly among the interfacial residue classes. Generalized Born reaction field energy roughly compensates for the electrostatic

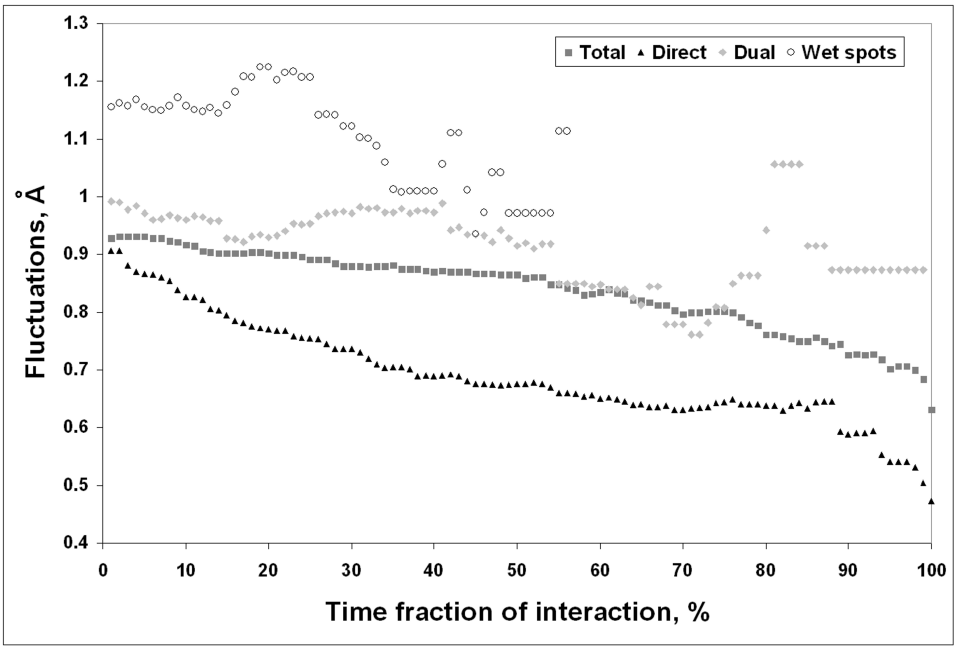


Figure 4.14: Fluctuations of interfacial residues decomposed by interaction type.

component *in vacuo*, so we discuss only the results obtained for van der Waals and electrostatic components *in vacuo*. The general trend for both components and all types of interactions is that the energy values decrease with the increase of residue interaction time, suggesting that both energy components stabilize complexes independently of the residue class. Energy decomposition (data not shown) illustrates that the electrostatic impact of water-mediated interactions is at least of the same order as of direct interactions, considering that the dielectric constant of water in the protein interface within the analyzed distance scale is approximately one order lower than the dielectric constant in bulk (and several times higher than in dry interfaces) [156]. The Van der Waals energy component is the lowest for dual residues and the highest for wet spots. Such a benefit for dual residues is explained by more tight contacts of the atoms additionally summed up with water atoms contacts. In wet spots there are only contacts with water atoms, which are not so tightly packed. Despite the quantitative differences observed for the SH3 and Ig interfaces, the important finding out of the energy decomposition is that all three interfacial residue classes are energetically

comparable even in implicit solvent, meaning that wet spots interactions are energetically of the same order as direct interactions. This conclusion could be generalized for all protein interfaces since the analyzed interface families substantially differ in physico-chemical properties. The obtained small differences between the families in the free energy components are due to intrinsic properties of the dataset, and the larger size and more hydrophilic nature of Ig interfaces. Noninterfacial residues were analyzed in the same way but their free energy contribution values were at least one order lower than of the interfacial ones. This suggests that inclusion of water-mediated interactions in protein interface definition is energetically well grounded.

Residence time of water molecules in wet spot sites

The analysis of the residence time distribution density of wet spot sites obtained from MD simulations suggests that the best theoretical model to describe the distribution density function should be defined as $\rho(\mathrm{t}) = \mathrm{Ct}^{-k}$, where C is a constant that can be obtained by normalization for each site individually, and the constant $\mathrm{k} > 0$ is the only distribution parameter. This k constant and the maximum residence time (T_{max}) in sites were taken as the parameters to compare different water sites. For wet spot sites the linear regression adjusted correlation coefficient r is equal to 0.97 ± 0.04 with *P*-value ranging from 8×10^{-13} to 3×10^{-2}. r and *P*-values for most of the surface and bulk solvent sites were not defined because the observed residence times were in the most cases less than 20 ps, meaning that there were just two points in the distribution (each point was obtained summing up the number of events on a 10 ps interval to avoid big fluctuations in the density function). As it is shown in Table 4.6, both k and T_{max} significantly differ (at the level of *t*-test *P*-value = 0.05) for different sites, indicating that water molecules in wet spot sites are less mobile than in bulk solvent or in surface hydration sites. At the same time, in each wet spot site many occupancy events occur that have as short residence as in bulk or surface sites. That agrees with the model proposed by Makarov et al., where the correlation function for residence time in hydration sites is decomposed into the sum of fast and slow diffusion exponent components. These components characterize bulk water motions and specific for hydration site events, respectively [157]. Other theoretical and experimental studies obtained similar

Table 4.6: Residence Time Parameters of Different Water Sites.

Site type	**Sample size**	**T_{max} (ps)**	**k**
Wet spot	110	137 ± 12	3.0 ± 1.0
Surface	30	18 ± 6	7.1 ± 1.1
Bulk	10	14 ± 2	8.7 ± 0.9

Sample size is the number of analyzed sites of each water site type. T_{max} is maximal residence time. k is residence time distribution parameter.

residence time values for different water sites, which vary from 1^{-10} ps for bulk solvent to 10^1-10^3 ps for protein hydration sites, cavities and cores [132]. T_{max} and k are well correlated (adjusted correlation coefficient r = 0.81 for $\ln(T_{max}) \sim$ k linear regression), meaning that maximum residence time of water molecules does not correspond to an opportunistic event of site occupation but is expected from the residence time distribution. There was no correlation between total occupancy of the sites and T_{max} ($r < 0.3$) because these parameters are independent and describe different kinetic characteristics of the site. While T_{max} is defined only by the energy barrier required for the molecule to leave the site, total occupancy is also dependent on the energy barrier of water transfer from bulk solvent to the site. The residence time analysis suggests that the potential barriers for wet spots sites are significantly higher than those for surface sites. However, it does not mean that the potential energy level of water molecules in wet spot sites are necessarily lower than in bulk solvent.

Free energy of water molecules in wet spot sites

To determine if water molecules contribute energetically favorably to complex formation we calculated their free energy using the free energy perturbation double decoupling method [145]. As a first step, free energy of removing a water molecule from bulk solvent was calculated. Electrostatic and van der Waals components were equal to 8.2 and -2.2 kcal/mol, respectively, which agrees well with the results obtained from similar calculations correlated with experimental data [145, 158] (Table 4.7). The second step consisted of the transfer of a water molecule from the wet spot site to vacuum. The difference of these two energy

Table 4.7: Free Energy Perturbation of Water Molecules (Ws) in 1UJ0 complex using the Double-Decoupling Method.

E (kcal/mol) site	Site type	E_{Elect}	E_{VDW}	-RT* $\ln(S_w S_p / S_{w*p})$	RT* $\ln(C_0 V_1)$	E_{rot}	E_{total}	ΔG^0
E12-R64	Wet spot	12.9	-1.5	0.4	-4.4	0	7.4	-1.4
D34-N66	Wet spot	8.3	0.1	0.4	-4.1	0	4.7	1.3
D34-N66, 2 Ws	Wet spot	22.9	-3.7	0.8	-8.2	0	12.6	-6.6
N52-M61,N63	Wet spot	8.9	0.1	0.4	-4.2	0	5.2	0.8
N52-M61,N63 2 Ws	Wet spot	18.1	0.1	0.8	-7.2	0	11.2	0.1
L58-R6	Surface	9.8	0.2	0.4	-3.8	0	6.4	-0.4
D31-S33	Surface	7.6	-0.6	0.4	-3.7	0	3.7	2.3
Lysozyme	Cavity	13.5	0.0	0.4	-3.9	0	10.0	-4.0
Bulk-vacuo transfer	Bulk	8.2	-2.2	-	-	-	6.0	-
Bulk-vacuo transfer[145]	Bulk	8.2	-2.2	-	-	-	6.0	-
Bulk-vacuo transfer[158]	Bulk	8.3	-2.4	-	-	-	5.9	-

E_{VDW}, van der Waals energy; -RT* $\ln(S_a S_b / S_{a*b})$, the free energy component related to the symmetry of water molecule (S_a), protein (S_b) and the complex of water molecule with protein (S_{a*b}); RT*$\ln(C_0 V_1)$, the free energy component associated with translational constraints in the V_1 volume; E_{rot}, the free energy component associated with rotational constraints; E_{total}, total free energy of a water molecule transfer from the site to vacuo; ΔG_0, free energy of transfer of a water molecule from bulk to the site [145].

components makes up the total energy of a water molecule transfer from bulk solvent to the wet spot site. The obtained results for several water sites of the SH3 domain complex 1uj0 show that the sites are very heterogeneous. In particular, the free energy of water molecule transfer from bulk to the site formed by the carboxyl oxygen of Glu12 in the SH3 domain and the side-chain of Arg64 in the ligand is -1.4 kcal/mol, meaning a favorable impact of a water molecule on the complex formation. The calculations carried out for another site formed by the side-chain of Asp34 in the SH3 domain and the side-chain of Asn66 in the ligand revealed a positive change of free energy (1.3 kcal/mol). However, as it was observed in the trajectory, another water molecule was present in this site and establishing a hydrogen bond with the first water molecule forming the wet spot. Consideration of both water molecules in the free energy calculations revealed an energy gain of -6.6 kcal/mol (Table 4.7). In this

case removal of two water molecules from a wet spot site leads to an increase of the free energy value, while a removal of each water molecule independently leads to a free energy decrease. Another example of such an effect was found in the site formed by the side-chain of Asn52 in the SH3 domain and the main-chain of Met61 in the ligand. Here, although the energy became more favorable by taking into account two water molecules transfer, water contribution was still not favorable. For the comparison of the free energies of wet spot sites we took several surface sites and a site in the cavity of lysozyme, which is not exposed to bulk solvent (1hel, 1.7 $\AA$). The X-ray structure of lysozyme presents a very stable water molecule [159], which is present in the site during the whole simulation with a residence time of several nanoseconds. The energetic impact of the cavity water to the stability of the lysozyme was quite significant (-4 kcal/mol). In surface sites no big negative values for free energy were found. In similar calculations performed with the double decoupling method for free energy calculation with AMBER, the obtained values for the free energy of water in hydration sites changed from slightly positive up to -5 kcal/mol [145, 158]. The examples of favorable energetic impact of water molecules on complex formation was also found in a study of various protein complexes by Monte Carlo calculations using different force fields [160]. The most important conclusion that can be driven from this free energy analysis is that water molecules in wet spot sites can not be characterized uniformly in energetic terms since in some cases they manifest properties similar to cavity waters and in other do not even contribute favorably to the complex free energy (just occupying an empty space between the residues). Nevertheless, it is realistic to claim that the introduction of water into protein interface description would crucially change the energy function of the system. Interestingly, a recent attempt of solvated protein docking has shown promising results [161].

4.2.5 Conclusion

We present a detailed MD study on 17 protein complexes representatives of two families of different interface nature. Our aim has been to gain insights into the contribution of interfacial solvent in protein-protein interactions. We show that water molecules in protein interfaces contribute to the conservation of protein interactions by allowing more sequence

variability in the interacting partners, which has important implications for the use of the correlated mutations concept in protein interactions studies.

Interfacial residues interacting through water are more mobile than those interacting directly but less than protein surface residues. Despite their broad heterogeneity, all interfacial residues are quantitatively comparable in terms of their contribution to the energy of complex formation, independently of their type of interaction. In the case of interfacial solvent, water molecules forming wet spots have significantly longer residence time than those on the protein surface, meaning that in terms of mobility interfacial protein residues and interfacial solvent are alike. Although interfacial water molecules are very diverse energetically, their contribution to the free energy of complex formation should be not be ignored.

Our data confirm that water plays an important active role in protein interfaces, suggesting that consideration of solvent in the development of energetic functions describing protein interactions is essential. Moreover, the introduction of water-mediated interactions into protein interface definitions should substantially increase the accuracy of protein interaction predictions based on protein contacts. We believe that the results obtained in this work could be useful for deeper understanding of the physico-chemical properties underlying protein-protein interactions in order to improve the accuracy of protein folding, docking and rational design methods.

Chapter 5

CONCLUSIONS AND FUTURE DIRECTIONS

This last chapter outlines the main achievements of my thesis work and relates them to the most recent research performed in the area. Future directions and possible implications of my results are also discussed.

5.1 SCOWLP database

During my thesis I developed the SCOWLP database and its web interface following a two-steps process: i) the automatic identification and atomic description of all protein interfaces contained in the PDB repository [70] and ii) the structural classification of protein binding regions by families (FBRs) [97].

There are some other FBR databases such as PRISM [55], PIBASE [162], SCOPPI [71], 3DID [163], although SCOWLP differs from them in: i) the accurate definition of atom-atom interactions at physicochemical level, ii) the inclusion of water-mediated interactions, iii) the inclusion of protein-peptide interactions, and iv) the clustering technique applied. The importance of including these fundamental components have been widely explained and discussed in Chapter 2.

The number of available tools and web servers for the analysis of protein-protein interactions and interfaces since I started my thesis [164] has been increasing, reflecting their need for research. SCOWLP web application permits the analysis of individual protein interfaces, and the database allows the possibility of high-throughput comparative studies of protein interactions and binding regions.

Specific structural databases for protein-nucleic acid (NA) [165] or for protein-saccharide (SAC) [166] complexes already exist, however, there is currently no database that combines ligands of different nature at once into a single database that structurally classifies protein interactions. One of my current ongoing work is the inclusion of protein-NA and protein-

SAC complexes into the SCOWLP database. This kind of classification will potentially be able to, i) differentiate protein regions recognizing other proteins from those interacting with NA or SAC, and ii) identify protein-families that can recognize proteins and other ligands (NA, SAC) through the same region.

Moreover, it is known that the role of water molecules mediating protein-NA [167] and protein-SAC [168] are as critical for recognition as for protein-protein interactions. The fact that SCOWLP is one of the few databases that takes into account water-mediated interactions, makes it suitable to describe and classify these type of complexes.

SCOWLP website (www.scowlp.org) has been already visited more than 6,000 times during the last 3 years, showing the interest of the scientific community for this kind of information. I am currently updating SCOWLP to improve its speed and user-friendliness, and to incorporate the interacting information obtained from protein-NA and protein-SAC complexes. With all these improvements SCOWLP may become a routinary tool for structural biologists and bioinformaticians.

The information contained in the SCOWLP database can be downloaded and used for independent research projects. The utility of this kind of structural information has been already highlighted in several studies:

- Validation of large-scale of protein-protein interaction information, determined by two-hybrid or pull-down analysis [169].

- Modeling the structures of multi-domain proteins and protein complexes using a hybrid method combining comparative modeling based on a template complex and protein docking [170].

- Interactions prediction of protein structures when two structures contain surface regions that resemble the complementary partners of a known interface [101, 104].

- Statistical analysis of special features of protein interactions such as the research work presented in Chapter 4 [113, 114] about the role of water molecules in protein interfaces.

- Extraction of statistical potentials to improve computational predictive techniques, such as the study we performed about the inclusion of solvent information in protein contacts prediction [171]. Other techniques such as docking or fold recognition could also benefit from this information.

5.2 Inference of binding regions within protein families

Proteins exhibit an enormous diversity of structures, and many proteins classified as a different fold present common three-dimensional architectures [84]. It has also been observed that proteins with different folds and functions often interact with other proteins through interfaces containing similar local structural features or interacting motifs [99].

During my PhD I performed the first systematic analysis of protein binding region conservation between protein families presenting structural resemblances across all known folds, in order to obtain binding inferences beyond family level. The results obtained by this approach highlighted that binding region conservation occurs systematically, specially between proteins considered to have different folds in the current protein topology classifications. Within the results, I could identify interface similarities between protein complexes that present structurally similar receptors, although classified in different folds. These cases support the feasibility of the methodology to propose putative binding regions to proteins.

My approach provides alternative binding regions for protein families, and can be used as a predictive tool for deriving putative binding inferences. However, the results obtained still require an accurate analysis at atomic level of the binding determinants to assess ligand recognition similarities. For this reason, I am currently working in a collaborative project to design a software for the comparison of the physico-chemical properties of atoms in the space, that could complement the results obtained in the work presented in my thesis.

5.3 Water in protein interfaces

Water networks in protein interfaces can complement direct interactions contributing significantly to molecular recognition, function, and stability of protein association. The description of protein interfaces may be modified by the inclusion of water-mediated interactions in the definition of protein interfaces. The inclusion of residues that cannot interact directly

with the ligand but through water molecules, so called wet spot residues, may also change interface's size and shape. During my PhD, I decided to study the role of water molecules in protein interfaces, paying special attention to wet spot residues.

The research project I present in Chapter 4 [113] about the study of water molecules in protein interfaces in a curated dataset of obligate and transient protein complexes represents one of the first high-throughput studies in this area. Indeed, it nicely complements the comparative analysis of the role of water molecules in the interfaces of biological and crystal packing complexes performed by Rodier et al. [68]. The results of my work showed that 40.1% of the interfacial residues were interacting through water and that wet spots represented a 14.5% of the total, emphasizing the importance of the inclusion of solvent in protein interaction studies, and the contribution of wet spots to interfacial description. I characterized for the first time wet spot residues in terms of secondary structure, temperature factors, residue composition, and pairing preferences. Interestingly, I also observed that obligate and transient interfaces present a comparable amount of solvent, which contrasts the old thoughts saying that obligate protein complexes are expected to exhibit similarities to protein cores having a dry and hydrophobic interfaces.

Moreover, the contact matrices generated in my work were further used by the group to study the influence of solvent in the prediction of protein contacts, implementing 'wet' together with classical 'dry' matrices in a correlated mutations approach [171].

I also contributed to studies performed in the group on the dynamic characteristics of water molecules in protein interfaces for two families of different interface nature [114]. We observed that interfacial residues interacting through water are more mobile than those interacting directly but less than protein surface residues. Also, that water molecules mediating residue-residue interactions have significantly longer residence time than those on the protein surface.

My results agree with other studies, suggesting the consideration of water molecules should improve the accuracy of many computational techniques, such as protein folding or rational design methods, as it has recently been shown for protein docking [161].

5.4 Other collaborative research projects

During my PhD I had the chance to work in some other collaborative projects. This allowed me to get in contact with a broad range of computational and experimental techniques. I highlight the research articles that have been already published although the results have not been included as part of my thesis:

5.4.1 Analysis of the impact of solvent on contacts prediction in proteins

by Sergey Samsonov, Joan Teyra, Gerd Anders and M Teresa Pisabarro
in BMC Structural Biology, 2009, 9:22

The correlated mutations concept is based on the assumption that interacting protein residues co-evolve, so that a mutation in one of the interacting counterparts is compensated by a mutation in the other. Approaches based on this concept have been widely used for protein contacts prediction since the 90s. Previously, we have shown that water-mediated interactions play an important role in protein interfaces. We have observed that current "dry" correlated mutations approaches might not properly predict certain interactions in protein interfaces due to the fact that they are water-mediated.

The goal of this study has been to analyze the impact of including solvent into the concept of correlated mutations. For this purpose we use linear combinations of the predictions obtained by the application of two different similarity matrices: a standard "dry" similarity matrix (DRY) and a "wet" similarity matrix (WET) derived from all water-mediated protein interfacial interactions in the PDB. We analyze two datasets containing 50 domains and 10 domain pairs from PFAM and compare the results obtained by using a combination of both matrices. We find that for both intra- and inter-domain contacts predictions the introduction of a combination of a "wet" and a "dry" similarity matrix improves the predictions in comparison to the "dry" one alone.

Our analysis, despite the complexity of its possible general applicability, opens up that the consideration of water may have an impact on the improvement of the contact predictions obtained by correlated mutations approaches.

- I contributed to the creation of the 'wet' contact matrices based on the data contained in SCOWLP database, and to the discussion of the research work.

5.4.2 A genome-scale DNA repair RNAi screen identifies a novel gene associated with hereditary spastic paraplegia

by Slabicki M, Theis M, Krastev DB, Teyra J, Paszkowski-Rogacz M, Junqueira M, Heninger AK, Poser I, Mundwiller E, Truchetto J, Prieur F, Confavreux C, Brice A, Shevchenko A, Pisabarro MT, Stevanin G and Buchholz FA

in Plos Biology, 2010, accepted

DNA repair is essential to maintain genome integrity and genes with roles in DNA repair are frequently mutated in a variety of human diseases. Repair via homologous recombination typically restores the original DNA sequence without introducing mutations and a number of genes that are required for homologous recombination DNA double-strand break repair (HR-DSBR) have been identified. However, a systematic analysis of this important DNA repair pathway in mammalian cells has not been reported.

Here, we describe a genome-scale endoribonuclease-prepared short interfering RNA (esiRNA) screen for genes involved in DNA double strand break repair. We report 61 genes that influenced the frequency of HR-DSBR and characterize in detail one of the genes that decreased the frequency of HR-DSBR. We show that the gene KIAA0415 encodes a putative helicase that interacts with SPG11 and SPG15, two proteins mutated in hereditary spastic paraplegia (HSP). We identify mutations in HSP patients, discovering KIAA0415/SPG48 as a novel HSP-associated gene and show that a KIAA0415/SPG48 mutant cell line is more sensitive to DNA damaging drugs.

We present the first genome-scale survey of HR-DSBR in mammalian cells providing a dataset that should accelerate the discovery of novel genes with roles in DNA repair and associated medical conditions. The discovery that proteins forming a novel protein complex are required for efficient HR-DSBR and are mutated in patients suffering from HSP suggests a link between HSP and DNA repair.

- I contributed to the functional annotation of the KIAA0415 using structure-based computational methods.

BIBLIOGRAPHY

[1] The universal protein resource (UniProt). *Nucleic Acids Res*, 36(Database issue), January 2008.

[2] NCBI. http://www.ncbi.nlm.nih.gov.

[3] J. Schultz, F. Milpetz, P. Bork, and C. P. Ponting. SMART, a simple modular architecture research tool: Identification of signaling domains. *Proceedings of the National Academy of Sciences*, 95(11):5857–5864, May 1998.

[4] E. L. Sonnhammer, S. R. Eddy, E. Birney, A. Bateman, and R. Durbin. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res*, 26(1):320–322, January 1998.

[5] C. J. Sigrist, L. Cerutti, N. Hulo, A. Gattiker, L. Falquet, M. Pagni, A. Bairoch, and P. Bucher. PROSITE: A documented database using patterns and profiles as motif descriptors. *Brief Bioinform*, 3(3):265–274, January 2002.

[6] A. Marchler-Bauer, J. B. Anderson, C. DeWeese-Scott, N. D. Fedorova, L. Y. Geer, S. He, D. I. Hurwitz, J. D. Jackson, A. R. Jacobs, C. J. Lanczycki, C. A. Liebert, C. Liu, T. Madej, G. H. Marchler, R. Mazumder, A. N. Nikolskaya, A. R. Panchenko, B. S. Rao, B. A. Shoemaker, V. Simonyan, J. S. Song, P. A. Thiessen, S. Vasudevan, Y. Wang, R. A. Yamashita, J. J. Yin, and S. H. Bryant. CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res*, 31(1):383–387, January 2003.

[7] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. CATH–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, August 1997.

[8] L. Lo Conte, B. Ailey, T. J. Hubbard, S. E. Brenner, A. G. Murzin, and C. Chothia. SCOP: a structural classification of proteins database. *Nucleic Acids Res*, 28(1):257–9, 2000.

[9] L. Holm and C. Sander. Dali/FSSP classification of three-dimensional protein folds. *Nucleic acids research*, 25(1):231–234, January 1997.

[10] PDB. http://www.rcsb.org/pdb.

[11] J. A. Mccammon and S. C. Harvey. *Dynamics of Proteins and Nucleic Acids*. Cambridge University Press, April 1988.

[12] T. Loup Verle. Computer "experiments" on classical fluids. ii. equilibrium correlation functions. *Physical Review Online Archive (Prola)*, 165(1):201–214, Jan 1968.

[13] P. P. Ewald. Die berechnung optischer und elektrostatischer gitterpotentiale. *Annalen der Physik*, 369(3):253–287, 1921.

[14] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. Dinola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8):3684–3690, 1984.

[15] T. A. Andrea, W. C. Swope, and H. C. Andersen. The role of long ranged forces in determining the structure and properties of liquid water. *J. Chem. Phys.*, 79:4576–4584, November 1983.

[16] H. C. Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *The Journal of Chemical Physics*, 72(4):2384–2393, 1980.

[17] R. W. Pastor, B. R. Brooks, and A. Szabo. An analysis of the accuracy of langevin and molecular dynamics algorithms. *Molecular Physics: An International Journal at the Interface Between Chemistry and Physics*, 65(6):1409–1419, 1988.

[18] R. J. Loncharich, B. R. Brooks, and R. W. Pastor. Langevin dynamics of peptides: The frictional dependence of isomerization rates of n-acetylalanyl-n-prime-methylamide. *Biopolymers*, 32(5):523–535, May 1992.

[19] J. A. Izaguirre, D. P. Catarello, J. M. Wozniak, and R. D. Skeel. Langevin stabilization of molecular dynamics. *The Journal of Chemical Physics*, 114(5):2090–2098, 2001.

[20] J. Ryckaert, G. Ciccotti, and H. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327–341, March 1977.

[21] S. Miyamoto and P. A. Kollman. Settle: An analytical version of the shake and rattle algorithm for rigid water models. *Journal of Computational Chemistry*, 13(8):952–962, 1992.

[22] P. Drabik, A. Liwo, C. Czaplewski, and J. Ciarkowski. The investigation of the effects of counterions in protein dynamics simulations. *Protein Eng.*, 14(10):747–752, October 2001.

[23] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case, and T. E. Cheatham. Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Accounts of Chemical Research*, 33(12):889–897, December 2000.

[24] A. Shrake and J. Rupley. Environment and exposure to solvent of protein atoms. lysozyme and insulin. *Journal of Molecular Biology*, 79(2):351–364, September 1973.

[25] S.J. Hubbard. NACCESS Computer Program. University College London; 1993.

[26] F. M. Richards. Areas, volumes, packing, and protein structure. *Annual Review of Biophysics and Bioengineering*, 6(1):151–176, 1977.

[27] M. L. Connolly. Analytical molecular surface calculation. *Journal of Applied Crystallography*, 16(5):548–558, Oct 1983.

[28] T. J. Richmond. Solvent accessible surface area and excluded volume in proteins. analytical equations for overlapping spheres and implications for the hydrophobic effect. *Journal of molecular biology*, 178(1):63–89, September 1984.

[29] M. Connolly. The molecular surface package. *Journal of Molecular Graphics*, 11(2):139–141, June 1993.

[30] T. Simonson. Electrostatics and dynamics of proteins. *Reports of Progress in Physics*, 66:737–787, May 2003.

[31] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *Journal of the American Chemical Society*, 112(16):6127–6129, August 1990.

[32] A. Onufriev, D. A. Case, and D. Bashford. Effective born radii in the generalized born approximation: the importance of being perfect. *J Comput Chem*, 23(14):1297–1304, November 2002.

[33] G. Wei and J. E. Shea. Effects of solvent on the structure of the Alzheimer amyloid-beta(25-35) peptide. *Biophys J*, 91(5):1638–47, 2006.

[34] E. M. Phizicky and S. Fields. Protein-protein interactions: methods for detection and analysis. *Microbiol Rev*, 59(1):94–123, 1995.

[35] G. D. Bader, D. Betel, and C. W. Hogue. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*, 31(1):248–50, 2003.

[36] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, and G. Cesareni. MINT: a Molecular INTeraction database. *FEBS Lett*, 513(1):135–40, 2002.

[37] I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S. M. Kim, and D. Eisenberg. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*, 30(1):303–5, 2002.

[38] P. Argos. An investigation of protein subunit and domain interfaces. *Protein Eng*, 2(2):101–13, 1988.

[39] J. Janin, S. Miller, and C. Chothia. Surface, subunit interfaces and interior of oligomeric proteins. *J Mol Biol*, 204(1):155–64, 1988.

[40] C. J. Tsai, S. L. Lin, H. J. Wolfson, and R. Nussinov. A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique. *J Mol Biol*, 260(4):604–20, 1996.

[41] S. Jones and J. M. Thornton. Principles of protein-protein interactions. *Proc Natl Acad Sci U S A*, 93(1):13–20, 1996.

[42] L. Lo Conte, C. Chothia, and J. Janin. The atomic structure of protein-protein recognition sites. *J Mol Biol*, 285(5):2177–98, 1999.

[43] J. Park, M. Lappe, and S. A. Teichmann. Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *J Mol Biol*, 307(3):929–38, 2001.

[44] P. Aloy, H. Ceulemans, A. Stark, and R. B. Russell. The relationship between sequence and interaction divergence in proteins. *J Mol Biol*, 332(5):989–98, 2003.

[45] O. Keskin, C. J. Tsai, H. Wolfson, and R. Nussinov. A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications. *Protein Sci*, 13(4):1043–55, 2004.

[46] W. S. Valdar and J. M. Thornton. Conservation helps to identify biologically relevant crystal contacts. *J Mol Biol*, 313(2):399–416, 2001.

[47] Y. Ofran and B. Rost. Analysing six types of protein-protein interfaces. *J Mol Biol*, 325(2):377–87, 2003.

[48] D. R. Caffrey, S. Somaroo, J. D. Hughes, J. Mintseris, and E. S. Huang. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci*, 13(1):190–202, 2004.

[49] S. Jones, A. Marin, and J. M. Thornton. Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng*, 13(2):77–82, 2000.

[50] A. Stein, R. B. Russell, and P. Aloy. 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res*, 33(Database issue):D413–7, 2005.

[51] F. P. Davis and A. Sali. PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics*, 21(9):1901–7, 2005.

[52] R. D. Finn, M. Marshall, and A. Bateman. iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, 21(3):410–2, 2005.

[53] S. Gong, G. Yoon, I. Jang, D. Bolser, P. Dafas, M. Schroeder, H. Choi, Y. Cho, K. Han, S. Lee, H. Choi, M. Lappe, L. Holm, S. Kim, D. Oh, and J. Bhak. PSIbase: a database of Protein Structural Interactome map (PSIMAP). *Bioinformatics*, 21(10):2541–3, 2005.

[54] S. Gong, C. Park, H. Choi, J. Ko, I. Jang, J. Lee, D. M. Bolser, D. Oh, D. S. Kim, and J. Bhak. A protein domain interaction interface database: InterPare. *BMC Bioinformatics*, 6:207, 2005.

[55] U. Ogmen, O. Keskin, A. S. Aytuna, R. Nussinov, and A. Gursoy. PRISM: protein interactions by structural matching. *Nucleic Acids Res*, 33(Web Server issue):W331–6, 2005.

[56] J. Zeng. Mini-review: computational structure-based design of inhibitors that target protein surfaces. *Comb Chem High Throughput Screen*, 3(5):355–62, 2000.

[57] T. Pawson. Specificity in signal transduction: from phosphotyrosine-SH2 domain interactions to complex cellular systems. *Cell*, 116(2):191–203, 2004.

[58] L. Castagnoli, A. Costantini, C. Dall'Armi, S. Gonfloni, L. Montecchi-Palazzi, S. Panni, S. Paoluzi, E. Santonico, and G. Cesareni. Selectivity and promiscuity in the interaction network mediated by protein recognition modules. *FEBS Lett*, 567(1):74–9, 2004.

[59] Y. Levy and J. N. Onuchic. Water and proteins: a love-hate relationship. *Proc Natl Acad Sci U S A*, 101(10):3325–6, 2004.

[60] A. Palencia, E. S. Cobos, P. L. Mateo, J. C. Martinez, and I. Luque. Thermodynamic dissection of the binding energetics of proline-rich peptides to the Abl-SH3 domain: implications for rational ligand design. *J Mol Biol*, 336(2):527–37, 2004.

[61] J. Janin. Wet and dry interfaces: the role of solvent in protein-protein and protein-DNA recognition. *Structure*, 7(12):R277–9, 1999.

[62] M. Levitt and B. H. Park. Water: now you see it, now you don't. *Structure*, 1(4):223–6, 1993.

[63] G. A. Papoian, J. Ulander, and P. G. Wolynes. Role of water mediated interactions in protein-protein recognition landscapes. *J Am Chem Soc*, 125(30):9170–8, 2003.

[64] P. M. Petrone and A. E. Garcia. MHC-peptide binding is assisted by bound water molecules. *J Mol Biol*, 338(2):419–35, 2004.

[65] Jmol. http://jmol.sourceforge.net.

[66] SCOP. http://scop.mrc-lmb.cam.ac.uk.

[67] P. Dafas, D. Bolser, J. Gomoluch, J. Park, and M. Schroeder. Using convex hulls to extract interaction interfaces from known structures. *Bioinformatics*, 20(10):1486–90, 2004.

[68] F. Rodier, R. P. Bahadur, P. Chakrabarti, and J. Janin. Hydration of protein-protein interfaces. *Proteins*, 60(1):36–45, 2005.

[69] K. Gunasekaran, B. Ma, and R. Nussinov. Is allostery an intrinsic property of all dynamic proteins? *Proteins*, 57(3):433–43, 2004.

[70] J. Teyra, A. Doms, M. Schroeder, and M. T. Pisabarro. SCOWLP: a web-based database for detailed characterization and visualization of protein interfaces. *BMC Bioinformatics*, 7(1):104, 2006.

[71] C. Winter, A. Henschel, W. K. Kim, and M. Schroeder. SCOPPI: a structural classification of protein-protein interfaces. *Nucleic Acids Res*, 34(Database issue):D310–4, 2006.

[72] K. Henrick and J. M. Thornton. PQS: a protein quaternary structure file server. *Trends Biochem Sci*, 23(9):358–61, 1998.

[73] H. Zhu, F. S. Domingues, I. Sommer, and T. Lengauer. NOXclass: prediction of protein-protein interaction types. *BMC Bioinformatics*, 7:27, 2006.

[74] B.S. Everitt, S. Landau, and M. Leese. *Cluster Analysis*. Arnold, fourth edition, 2001.

[75] A. D. Gordon. *Classification*. Crc Press Llc, second edition, 1999.

[76] C. J. Tsai, S. L. Lin, H. J. Wolfson, and R. Nussinov. A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique. *J Mol Biol*, 260(4):604–20, 1996.

[77] W. K. Kim and J. C. Ison. Survey of the geometric association of domain-domain interfaces. *Proteins*, 61(4):1075–88, 2005.

[78] W. K. Kim, A. Henschel, C. Winter, and M. Schroeder. The many faces of protein-protein interactions: A compendium of interface geometry. *PLoS Comput Biol*, 2(9):e124, 2006.

[79] J. Teyra, M. Paszkowski-Rogacz, G. Anders, and M. T. Pisabarro. SCOWLP classification: Structural comparison and analysis of protein binding regions. *BMC Bioinformatics*, [In review process], 2007.

[80] D. Lupyan, A. Leo-Macias, and A. R. Ortiz. A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics*, 21(15):3255–63, 2005.

[81] A. R. Ortiz, C. E. Strauss, and O. Olmea. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci*, 11(11):2606–21, 2002.

[82] M. J. Sippl, S. J. Suhrer, M. Gruber, and M. Wiederstein. A discrete view on fold space. *Bioinformatics (Oxford, England)*, 24(6):870–871, March 2008.

[83] A. Pascual-García, D. Abia, A. R. Ortiz, and U. Bastolla. Cross-over between discrete and continuous protein structure space: Insights into automatic classification and networks of protein structures. *PLoS Comput Biol*, 5(3):e1000331+, March 2009.

[84] D. Petrey and B. Honig. Is protein classification necessary? toward alternative approaches to function annotation. *Current Opinion in Structural Biology*, 19(3):363–368, June 2009.

[85] W. Taylor. Evolutionary transitions in protein fold space. *Current Opinion in Structural Biology*, 17(3):354–361, June 2007.

[86] J. Skolnick, A. K. Arakaki, S. Y. Lee, and M. Brylinski. The continuity of protein structure space is an intrinsic property of proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 106(37):15690–15695, September 2009.

[87] D. Petrey, M. Fischer, and B. Honig. Structural relationships among proteins with different global topologies and their implications for function annotation strategies. *Proceedings of the National Academy of Sciences*, 106(41):17377–17382, October 2009.

[88] R. Kolodny, P. Koehl, and M. Levitt. Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *Journal of molecular biology*, 346(4):1173–1188, March 2005.

[89] M. J. Sippl and M. Wiederstein. A note on difficult structure alignment problems. *Bioinformatics*, 24(3):426–427, February 2008.

[90] Y. Xin and B. Christopher. Non-sequential structure-based alignments reveal topology-independent core packing arrangements in proteins. *Bioinformatics*, 21(7):1010–1019, April 2005.

[91] A. Abyzov and V. A. Ilyin. A comprehensive analysis of non-sequential alignments between all protein structures. *BMC Structural Biology*, 7:78+, November 2007.

[92] A. Guerler and E. W. Knapp. Novel protein folds and their nonsequential structural analogs. *Protein Sci*, pages ps.035469.108+, June 2008.

[93] J. Dundas, T. A. Binkowski, B. Dasgupta, and J. Liang. Topology independent protein structural alignment. *BMC Bioinformatics*, 8:388+, October 2007.

[94] R. B. Russell, P. D. Sasieni, and M. J. Sternberg. Supersites within superfolds. binding site similarity in the absence of homology. *Journal of molecular biology*, 282(4):903–918, October 1998.

[95] P. Aloy and R. B. Russell. Ten thousand interactions for the molecular biologist. *Nature Biotechnology*, 22(10):1317–1321, October 2004.

[96] P. Aloy and R. B. Russell. Interrogating protein interaction networks through structural biology. *Proceedings of the National Academy of Sciences of the United States of America*, 99(9):5896–5901, April 2002.

[97] J. Teyra, M. Paszkowski-Rogacz, G. Anders, and M. T. Pisabarro. Scowlp classification: structural comparison and analysis of protein binding regions. *BMC bioinformatics*, 9:9+, January 2008.

[98] A. Henschel, W. K. Kim, and M. Schroeder. Equivalent binding sites reveal convergently evolved interaction motifs. *Bioinformatics*, 22(5):550–555, March 2006.

[99] O. Keskin and R. Nussinov. Favorable scaffolds: proteins with different sequence, structure and function may associate in similar ways. *Protein Eng Des Sel*, 18(1):11–24, 2005.

[100] A. Shulman-Peleg, R. Nussinov, and H. J. Wolfson. SiteEngines: recognition and comparison of binding sites and protein-protein interfaces. *Nucleic Acids Res*, 33(Web Server issue):W337–41, 2005.

[101] S. A. Aytuna, A. Gursoy, and O. Keskin. Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics (Oxford, England)*, 21(12):2850–2855, June 2005.

[102] H. Zhu, I. Sommer, T. Lengauer, and F. S. Domingues. Alignment of non-covalent interactions at protein-protein interfaces. *PLoS ONE*, 3(4), 2008.

[103] A. Shulman-Peleg, R. Nussinov, and H. J. Wolfson. Recognition of functional sites in protein structures. *Journal of molecular biology*, 339(3):607–633, June 2004.

[104] S. Günther, P. May, A. Hoppe, C. Frömmel, and R. Preissner. Docking without docking: Isearch-prediction of interactions using known interfaces. *Proteins*, September 2007.

[105] O. Carugo and P. Argos. Protein-protein crystal-packing contacts. *Protein Sci*, 6(10):2261–2263, October 1997.

[106] E. Krissinel and K. Henrick. Secondary-structure matching (ssm), a new tool for fast protein structure alignment in three dimensions. *Acta crystallographica. Section D, Biological crystallography*, 60(Pt 12 Pt 1):2256–2268, December 2004.

[107] S. Miller, A. M. Lesk, J. Janin, and C. Chothia. The accessible surface area and stability of oligomeric proteins. *Nature*, 6133(328):834–6, 1987.

[108] B. V. North, D. Curtis, and P. C. Sham. A note on the calculation of empirical p values from monte carlo procedures. *Am J Hum Genet*, 71(2):439–441, August 2002.

[109] R. Filipek, M. Rzychon, A. Oleksy, M. Gruca, A. Dubin, J. Potempa, and M. Bochtler. The staphostatin-staphopain complex: a forward binding inhibitor in complex with its target cysteine protease. *J Biol Chem*, 278(42):40959–66, 2003.

[110] Patrick Aloy and Robert B. Russell. Structural systems biology: modelling protein interactions. *Nature Reviews Molecular Cell Biology*, 7(3):188–197, March 2006.

[111] J. F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak,

R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, September 2005.

[112] Y. S. Choi, J. S. Yang, Y. Choi, S. H. Ryu, and S. Kim. Evolutionary conservation in multiple faces of protein interaction. *Proteins: Structure, Function, and Bioinformatics*, 77(1):14–25, 2009.

[113] J. Teyra and M. T. Pisabarro. Characterization of interfacial solvent in protein complexes and contribution of wet spots to the interface description. *Proteins*, 67(4):1087–1095, 2007.

[114] S. Samsonov, J. Teyra, and M. T. Pisabarro. A molecular dynamics approach to study the importance of solvent in protein interactions. *Proteins: Structure, Function, and Bioinformatics*, 73(2):515–525, November 2008.

[115] C. Mattos. Protein-water interactions in a dynamic world. *Trends Biochem Sci*, 27(4):203–8, 2002.

[116] M. Shatsky, A. Shulman-Peleg, R. Nussinov, and H. J. Wolfson. The multiple common point set problem and its application to molecule binding pattern detection. *J Comput Biol*, 13(2):407–28, 2006.

[117] R. Nair, J. Liu, T. Soong, T. Acton, J. Everett, A. Kouranov, A. Fiser, A. Godzik, L. Jaroszewski, C. Orengo, G. Montelione, and B. Rost. Structural genomics is the largest contributor of novel structural leverage. *Journal of Structural and Functional Genomics*, 2009.

[118] G. Anders, J. Teyra, and M. T. Pisabarro. APFAN- Automated Protein Fold Annotation. *Bioinformatics*, [Submitted in December], 2007.

[119] U. Langhorst, J. Backmann, R. Loris, and J. Steyaert. Analysis of a water mediated protein-protein interactions within RNase T1. *Biochemistry*, 39(22):6586–93, 2000.

[120] J. E. Ladbury. Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. *Chem Biol*, 3(12):973–80, 1996.

[121] M. Petukhov, D. Cregut, C. M. Soares, and L. Serrano. Local water bridges and protein conformational stability. *Protein Sci*, 8(10):1982–9, 1999.

[122] I. M. Nooren and J. M. Thornton. Diversity of protein-protein interactions. *Embo J*, 22(14):3486–92, 2003.

[123] J. Mintseris and Z. Weng. Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc Natl Acad Sci U S A*, 102(31):10930–5, 2005.

[124] S. Ansari and V. Helms. Statistical analysis of predominantly transient protein-protein interfaces. *Proteins*, 61(2):344–55, 2005.

[125] F. Glaser, D. M. Steinberg, I. A. Vakser, and N. Ben-Tal. Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins*, 43(2):89–102, 2001.

[126] J. Mintseris and Z. Weng. Atomic contact vectors in protein-protein recognition. *Proteins*, 53(3):629–39, 2003.

[127] I. M. Nooren and J. M. Thornton. Structural characterisation and functional significance of transient protein-protein interactions. *J Mol Biol*, 325(5):991–1018, 2003.

[128] D. Xu, C. J. Tsai, and R. Nussinov. Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Eng*, 10(9):999–1012, 1997.

[129] S. Parthasarathy and M. R. Murthy. Analysis of temperature factor distribution in high-resolution protein structures. *Protein Sci*, 6(12):2561–7, 1997.

[130] S. Park and J. G. Saven. Statistical and molecular dynamics studies of buried waters in globular proteins. *Proteins*, 60(3):450–63, 2005.

[131] J. Bonet, G. Caltabiano, A. K. Khan, M. A. Johnston, C. Corbi, A. Gomez, X. Rovira, J. Teyra, and J. Villa-Freixa. The role of residue stability in transient protein-protein

interactions involved in enzymatic phosphate hydrolysis. A computational study. *Proteins*, 63(1):65–77, 2006.

[132] T. M. Raschke. Water structure and interactions with protein surfaces. *Curr Opin Struct Biol*, 16(2):152–9, 2006.

[133] M. M. Rhodes, K. Reblova, J. Sponer, and N. G. Walter. Trapped water molecules are essential to structural dynamics and function of a ribozyme. *Proc Natl Acad Sci U S A*, 103(36):13380–5, 2006.

[134] P. M. Petrone and A. E. Garcia. MHC-peptide binding is assisted by bound water molecules. *J Mol Biol*, 338(2):419–35, 2004.

[135] D. Hamelberg, T. Shen, and J. A. McCammon. Insight into the role of hydration on protein dynamics. *J Chem Phys*, 125(9):094905, 2006.

[136] A. Amadasi, F. Spyrakis, P. Cozzini, D. J. Abraham, G. E. Kellogg, and A. Mozzarelli. Mapping the energetics of water-protein and water-ligand interactions with the natural HINT forcefield: predictive tools for characterizing the roles of water in biomolecules. *J Mol Biol*, 358(1):289–309, 2006.

[137] Z. Li and T. Lazaridis. Thermodynamics of buried water clusters at a protein-ligand binding interface. *J Phys Chem B*, 110(3):1464–75, 2006.

[138] G. A. Papoian, J. Ulander, M. P. Eastwood, Z. Luthey-Schulten, and P. G. Wolynes. Water in protein structure prediction. *Proc Natl Acad Sci U S A*, 101(10):3352–7, 2004.

[139] Y. Levy and J. N. Onuchic. Water mediation in protein folding and molecular recognition. *Annu Rev Biophys Biomol Struct*, 35:389–415, 2006.

[140] I. Mihalek, I. Res, and O. Lichtarge. On itinerant water molecules and detectability of protein-protein interfaces through comparative analysis of homologues. *J Mol Biol*, 369(2):584–95, 2007.

[141] M. T. Pisabarro, L. Serrano, and M. Wilmanns. Crystal structure of the abl-SH3 domain complexed with a designed high-affinity peptide ligand: implications for SH3-ligand interactions. *J Mol Biol*, 281(3):513–21, 1998.

[142] P. Bork, L. Holm, and C. Sander. The immunoglobulin fold. Structural classification, sequence patterns and common core. *J Mol Biol*, 242(4):309–20, 1994.

[143] Jr. Craft, J. W. and G. B. Legge. An AMBER/DYANA/MOLMOL phosphorylated amino acid library set and incorporation into NMR structure calculations. *J Biomol NMR*, 33(1):15–24, 2005.

[144] V. Lafont, M. Schaefer, R. H. Stote, D. Altschuh, and A. Dejaegere. Protein-protein recognition and interaction hot spots in an antigen-antibody complex: free energy decomposition identifies efficient amino acids. *Proteins*, 67(2):418–34, 2007.

[145] D. Hamelberg and J. A. McCammon. Standard free energy of releasing a localized water molecule from the binding pockets of proteins: double-decoupling method. *J Am Chem Soc*, 126(24):7683–9, 2004.

[146] R package Development Core Team. R: a language and environment for statistical computing. Vienna, Austria, 2006.

[147] U. Gobel, C. Sander, R. Schneider, and A. Valencia. Correlated mutations and residue contacts in proteins. *Proteins*, 18(4):309–17, 1994.

[148] D. J. Thomas, G. Casari, and C. Sander. The prediction of protein contacts from multiple sequence alignments. *Protein Eng*, 9(11):941–8, 1996.

[149] S. B. Nagl. Can correlated mutations in protein domain families be used for protein design? *Brief Bioinform*, 2(3):279–88, 2001.

[150] P. J. Kundrotas and E. G. Alexov. Predicting residue contacts using pragmatic correlated mutations method: reducing the false positives. *BMC Bioinformatics*, 7:503, 2006.

[151] I. Halperin, H. Wolfson, and R. Nussinov. Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families. *Proteins*, 63(4):832–45, 2006.

[152] G. R. Smith, M. J. Sternberg, and P. A. Bates. The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking. *J Mol Biol*, 347(5):1077–101, 2005.

[153] D. Vitkup, D. Ringe, G. A. Petsko, and M. Karplus. Solvent mobility and the protein 'glass' transition. *Nat Struct Biol*, 7(1):34–8, 2000.

[154] V. Helms. Protein dynamics tightly connected to the dynamics of surrounding and internal water molecules. *Chemphyschem*, 8(1):23–33, 2007.

[155] P. Koehl. Electrostatics calculations: latest methodological advances. *Curr Opin Struct Biol*, 16(2):142–51, 2006.

[156] K. D. Collins, G. W. Neilson, and J. E. Enderby. Ions in water: characterizing the forces that control chemical processes and biological structure. *Biophys Chem*, 128(2-3):95–104, 2007.

[157] V. A. Makarov, B. K. Andrews, P. E. Smith, and B. M. Pettitt. Residence times of water molecules in the hydration sites of myoglobin. *Biophys J*, 79(6):2966–74, 2000.

[158] Y. Lu, C. Y. Yang, and S. Wang. Binding free energy contributions of interfacial waters in HIV-1 protease/inhibitor complexes. *J Am Chem Soc*, 128(36):11830–9, 2006.

[159] T. Imai, R. Hiraoka, A. Kovalenko, and F. Hirata. Water molecules in a protein cavity detected by a statistical-mechanical theory. *J Am Chem Soc*, 127(44):15334–5, 2005.

[160] C. Barillari, J. Taylor, R. Viner, and J. W. Essex. Classification of water molecules in protein binding sites. *J Am Chem Soc*, 129(9):2577–87, 2007.

[161] A. D. van Dijk and A. M. Bonvin. Solvated docking: introducing water into the modelling of biomolecular complexes. *Bioinformatics*, 22(19):2340–7, 2006.

[162] D. Korkin, F. P. Davis, and A. Sali. Localization of protein-binding sites within families of proteins. *Protein Sci*, 14(9):2350–2360, September 2005.

[163] A. Stein, A. Panjkovich, and P. Aloy. 3did update: domain-domain and peptide-mediated interactions of known 3d structure. *Nucleic acids research*, 37(Database issue):gkn690+, January 2009.

[164] N. Tuncbag, G. Kar, O. Keskin, A. Gursoy, and R. Nussinov. A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. *Brief Bioinform*, 10(3):217–232, May 2009.

[165] Semin Lee and Tom L. Blundell. Bipa: a database for protein-nucleic acid interaction in 3d structures. *Bioinformatics*, 25(12):1559–1560, June 2009.

[166] Rene Ranzinger, Stephan Herget, Thomas Wetter, and Claus W. von der Lieth. Glycomedb - integration of open-access carbohydrate structure databases. *BMC Bioinformatics*, 9(1), 2008.

[167] B. Jayaram and T. Jain. The role of water in protein-dna recognition. *Annu Rev Biophys Biomol Struct*, 33:343–361, 2004.

[168] S. M. Tschampel and R. J. Woods. Quantifying the role of water in protein-carbohydrate interactions. *J Phys Chem A.*, 43:9175–81, 2003.

[169] Christina Kiel, Pedro Beltrao, and Luis Serrano. Analyzing protein interaction networks using structural information. *Annual review of biochemistry*, 77(1):415–441, 2008.

[170] D. Korkin, F. P. Davis, F. Alber, T. Luong, M. Y. Shen, V. Lucic, M. B. Kennedy, and A. Sali. Structural modeling of protein interactions by analogy: application to PSD-95. *PLoS Comput Biol*, 2(11):e153, 2006.

[171] S. A. Samsonov, J. Teyra, G. Anders, and M. T. Pisabarro. Analysis of the impact of solvent on contacts prediction in proteins. *BMC structural biology*, 9:22+, April 2009.