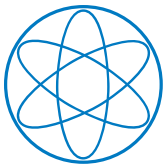


Technische Universität München
Physik Department
Lehrstuhl für Molekulardynamik T38

Solvation of protein binding sites and optimisation of protein-protein complex prediction

Dipl.-Inform.
Sebastian Eckehart Schneider

Vollständiger Abdruck der von der Fakultät für Physik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. Thorsten Hugel
Prüfer der Dissertation:	1. Univ.-Prof. Dr. Martin Zacharias
	2. Univ.-Prof. Dr. Iris Antes

Die Dissertation wurde am 16.05.2012 bei der Technischen Universität München eingereicht und durch die Fakultät für Physik am 05.07.2012 angenommen.

Abstract

Knowledge of the binding sites of biomolecules and understanding the complex formation of proteins are essential for the understanding of the functionality of biological processes. Bioinformatic methods can be used if experimental methods are unable to resolve the structure of individual protein complexes.

The aqueous solvent in the cell surrounds the proteins and other biomolecules and can influence the stability and functionality of those molecules. This thesis is concerned with the question, whether solvent molecules and the protein's surface interact differently depending on the function of the surface. The study analysis the behavior of solvent molecules at protein-protein and protein-ligand binding sites.

Experimental methods are often applied to solve the three dimensional structure of protein-protein complexes at high resolution. If this is not possible for individual protein complexes, computational docking can be used to generate possible complexes by means of a force field based approach. In the second part of this thesis, methods for the optimization of complex prediction are investigated. Predictions from bioinformatic methods ans information from experiments are used to optimize the docking procedure. This method is then applied to predict the three dimensional structure of a protein complex which could not be resolved so far.

Solvation of proteins, mixed solvents, protein binding site prediction, protein-small ligand binding site prediction, desolvation penalty, unbound protein protein docking, C1q, complement system, complex prediction

Zusammenfassung

Die Kenntnis der Bindestellen von Biomolekülen sowie des Bindungsprozesses zweier oder mehrerer Proteine ist von entscheidender Bedeutung für das Verstehen biologischer Prozesse. Rechnerbasierte Methoden können eingesetzt werden, wenn experimentelle Methoden für die Aufklärung einzelner Proteinkomplexe nicht anwendbar sind.

Die wässrige Lösung innerhalb der Zelle, welche die Proteine während ihres Lebenszyklus umgibt, kann die Funktion und die Form ebendieser beeinflussen. Daher beschäftigt sich der erste Teil dieser Arbeit mit der Frage, ob die Interaktion zwischen Molekülen der Flüssigkeit und der Proteinoberfläche in Abhängigkeit von der Funktion der Oberfläche variiert. Im Fokus der Untersuchungen stehen dabei Bindestellen zwischen zwei Proteinen sowie zwischen Proteinen und kleinen organischen Bindepartnern.

Mit Hilfe experimenteller Methoden ist es oftmals möglich dreidimensionale Strukturen von Proteinkomplexen in hoher Auflösung zu bestimmen. Sind diese für spezielle Proteinkomplexe nicht anwendbar, können computergestützte Methoden zur Bestimmung des Komplexes herangezogen werden. Diese generieren, zum Beispiel mit Hilfe von Kraftfeld basierten Interaktionsberechnungen, mögliche Komplexe. Im zweiten Teil dieser Arbeit werden Methoden zur Optimierung der Komplexbildungsvorhersage untersucht. Dazu werden computergestützte Bindestellenvorhersagen sowie Informationen aus experimentellen Studien herangezogen. Abschließend wird die Leistungsfähigkeit dieser Prozedur an einem Beispiel mit unbekannter dreidimensionaler Struktur demonstriert.

Solvatisierung von Proteinen, Bindeseitenvorhersage, Protein-Liganden Bindung, Desolvatisierungskosten, Optimierung des Protein-Protein Dockings, Komplexvorhersage

Ist das die Simulation oder gilt das schon?
Gustav – Soldat/in oder Veteran

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Interactions between biomolecules are essential for the functioning of biological systems. The life cycle of a cell is determined by cellular processes, based on interactions between proteins, organic ligands, DNA and RNA. They take place in the aqueous solution within the cell, under the direct or indirect influence of the aqueous solvent, which all these interactions have in common.

The binding of water and other solvents to the surface of proteins has been under investigation by experimental groups for decades, and has revealed information on dry and wet protein binding interfaces. Often single solvent molecules tightly bound to the surface, or known to be involved in biological processes, were observed. In recent years, more and more computational approaches arose, investigating the binding and the dynamics of diverse solvents. These methods are becoming of increasing interest in the context of binding site as well as binding affinity prediction, especially for ligand design in pharmaceutical research.

Besides the prediction of binding sites, the prediction of the three-dimensional structure of a complex between two or more proteins is of essential interest. X-Ray crystallography and other experimental methods reveal the structure of protein complexes, but are limited to the size of the system, its dynamics and structural features. In the past decades, many different methods have been developed to computationally predict the structure of associated proteins, as well as the driving forces of this complex formation. The finding of such complexes is often supported by knowledge on the binding process derived from experimental studies or computational prediction methods.

In this thesis, both of the above introduced fields of research have been investigated. The behavior of solvents in known protein binding sites was analyzed and the use for binding site prediction explored. Furthermore, the best practice for the use of external data for complex prediction was developed within a competitive complex prediction approach, and the limit of achievable, using state of the art bioinformatics methods, shown. This procedure has been applied to a protein complex of the complement system with an unknown three dimensional structure, relevant for medical research in the field of immunology, to predict the former unknown three dimensional structure.

Chapter 2

Fundamentals

2.1 Protein interactions

Many processes in biological systems are based on the interaction between proteins and other biomolecules. Therefore, knowledge on the kind of interaction is crucial for understanding the function and dynamics of biomolecular processes. Proteins can interact with most of the components in the cell, as there are other proteins, peptides, small organic ligands, DNA and RNA, lipids and others. Some form high affinity complexes which last for nearly the whole lifetime of the proteins (Gavin et al., 2002) and fulfill their function only in those special and often complex arrangements (for example transporter proteins in membranes). Other complexes form just for a short period to activate processes in the cell (e.g. in the signal pathway) or bind until other interactions occur (e.g. enzyme inhibitors). A full understanding of cellular functions requires structural knowledge of all these interactions. It will not be possible, in the foreseeable future, to determine the structure of all detected protein-protein interactions (PPI) experimentally at high resolution. Structural modeling and structure prediction is of increasing importance to obtain at least realistic models of complexes (Halperin et al., 2002, Bonvin, 2006, Andrusier et al., 2008, Vajda and Kozakov, 2009, Zacharias, 2010a). If the structure of the isolated protein partners or of closely related proteins is available it is possible to use a variety of computational docking methods to generate putative complex structures.

The driving force for the protein binding process corresponds to the associated change in free energy which depends on the structural and physicochem-

ical properties of the protein partners. The "lock and key" concept of binding proposed by E. Fischer (Fischer, 1894) emphasizes the importance of optimal sterical complementarity at binding interfaces as a decisive factor to achieve high affinity and specificity.

However, proteins and other interacting biomolecules are not rigid but can undergo a variety of motions, even at physiological temperatures. The induced fit concept has evolved based on the observation that binding can result in significant conformational changes of partner molecules (Koshland, 1958). Within this concept protein partners induce conformational changes during the binding process that are required for specific complex formation. It should be emphasized, that in principle all possible molecular recognition processes require a certain degree of conformational adaptation. In recent years, extensions of the induced-fit concept emerged, based on ideas from statistical physics. A preexisting ensemble of several inter convertible conformational states being in equilibrium has been postulated (Tsai et al., 1999). Structures are among these states close to the bound and unbound forms. Binding of a partner molecule to the bound form shifts the equilibrium towards the bound form. Since every conformation is, in principle, accessible albeit with a potentially low statistical weight already in the unbound state the original induced fit concept is a special case of ensemble selection where only the presence of a ligand gives rise to an appreciable concentration of the bound partner structure.

Proteins can often form a complex with a variety of different binding partners, either at different or shared binding sites (Gavin et al., 2002, Rual et al., 2005). Interactions between those partners are of diverse strength and can base on different types of interactions as electrostatic interactions (Schreiber and Fersht, 1996, Sheinerman and Honig, 2002), hydrogen bonding and salt bridges across the binding site (Xu et al., 1997), the hydrophobic effect (Tsai et al., 1997), water mediated hydrogen bonding (Rodier et al., 2005) or cofactor induced binding (Ryu et al., 1999, Giangrande et al., 2000). Hints on the interactions that play a role in an individual binding process can be derived from the protein structure of the isolated proteins, as well as assumptions on the contribution of particular residues to binding, undergoing some kind of interaction (e.g. electrostatic interactions, hydrophobicity). Other possible binding factors (hydrogen bonding networks, cofactor induced binding) need additional information that is often not available or has to be generated with costly methods.

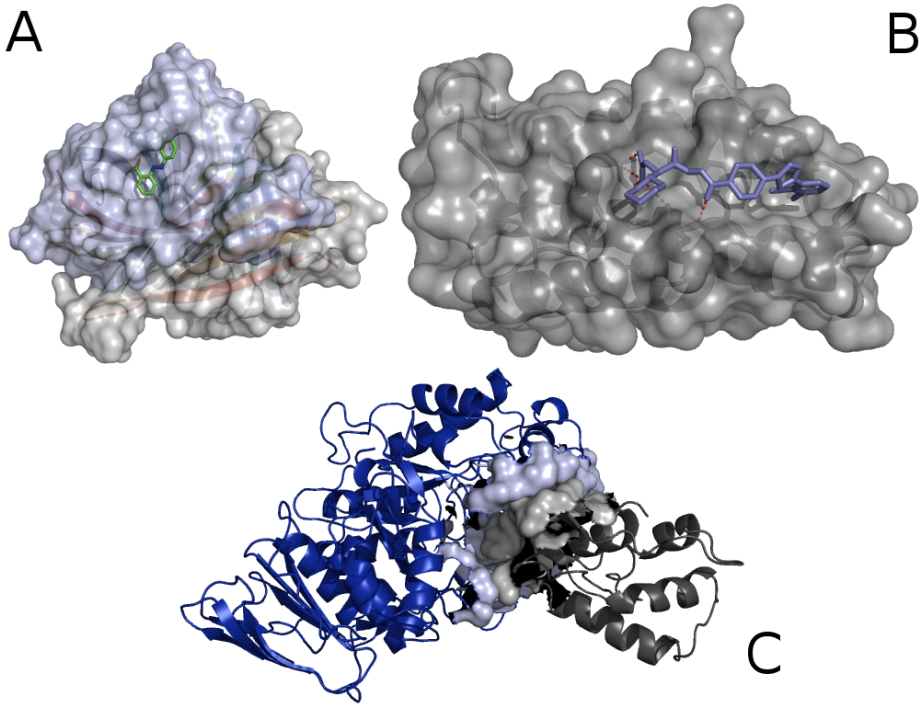


Figure 2.1 – Examples of protein-protein and protein-ligand associations. A) Small ligand bound within an enclosed cavity with mostly hydrophobic residues in the surrounding (pdb1srf). B) Ligand bound to the surface of a protein, inhibiting the protein-protein association of Interleukin 2 (IL 2) and IL-R2α (pdb1pw6). Possible polar contacts between protein and ligand are shown as red dots. C) Example of a protein-protein binding with the binding site of alpha-Amylase and it's inhibitor shown in surface representation (pdb1tmq).

A binding site can be determined as those residues buried under complex formation and which become inaccessible from solvent. To ease this criterion often atoms within a distance of 5 Å to the protein partner in a known complex are counted as binding site atoms. Since not all residues within a binding site have to contribute to the binding in the same manner (Reichmann et al., 2005), regions that add a high amount to the binding free energy are called a protein binding sites "hot spot" (Clackson and Wells, 1995, Bogan and Thorn, 1998) and are obligatory for many complex formation processes.

It is often found that in case of protein-protein binding the binding region of a protein can be split into a core region and a rim region. The core is completely buried upon complex formation and often of hydrophobic nature, whereas residues in the rim region can be, at least partially, solvent exposed (Clackson and Wells, 1995, Janin and Seraphin, 2003, Bahadur et al., 2004). The size of the binding site also contributes to the strength of the coupling between proteins as well as the packing of residues (Conte et al., 1999). Additionally, the composition of those regions or hot spots often differs from

other regions on the protein's surface (Jones and Thornton, 1996, Tsai et al., 1997, Bahadur and Zacharias, 2008), as also the geometric arrangement of the residues. Binding sites for small ligands, like organic molecules, often tend to be concave and form in many cases a cavity with distinct physiochemical properties like a high amount of hydrophobic residues (Mattos and Ringe, 1996, Vajda and Guarnieri, 2006) whereas protein-protein binding sites often have a more flattened contact area (with exceptions, e.g. enzyme-inhibitor complexes; some examples of protein-protein and protein-ligand binding are shown in Figure 2.1).

Many statistical analysis have been performed from known crystal structures to characterize protein-protein binding sites (Jones and Thornton, 1996, Tsai et al., 1997, Glaser et al., 2001, Janin and Seraphin, 2003, Reichmann et al., 2005) as well as protein-small ligand binding sites (Mattos and Ringe, 1996, Liang et al., 1998, Vajda and Guarnieri, 2006). In some cases a significant difference between binding site and non binding site could be found, while in other cases the binding site followed the average distribution of physiochemical properties and residue concentrations (Lichtarge et al., 1996, Conte et al., 1999, Chakrabarti and Janin, 2002).

Experimental methods have been developed to identify protein binding sites, such as mutagenesis experiments, e.g. alanine scanning. In these experiments guessed or predicted residues are substituted with an alanine residue and the change in binding free energy is measured (Cunningham and Wells, 1989, Wells, 1990, 1991). High changes in binding free energy above a given threshold, indicate the relevance of the residue in the binding process (Thorn and Bogan, 2001, Keskin et al., 2005). Hot spots of binding can be identified with this method at the cost of high lab effort. Computational methods try to follow these methods (Kortemme and Baker, 2002, Benedix et al., 2009) by identification of such hot spots due to alanine scanning. Hot spot and binding site prediction has also been the effort of many computational approaches (reviewed in Zhou and Qin (2007), de Vries and Bonvin (2008), Leis et al. (2010), Wass et al. (2011)) which base on different interaction motifs or statistical evidence.

2.2 Protein interaction prediction

If there is no information on the binding sites of a protein given by experimental methods, in silicio approaches are used to predict atoms or residues that are involved into binding processes. Those predictors aim to define atoms, residues or approximated regions on the protein's surface involved into biological processes on the base of either the properties of the protein itself or meta informations e.g. from data mining processes. The former try to identify characteristics of active sites: One starting point is to scan and select physicochemical properties of the protein that mediate protein-protein or protein-small ligand interactions while alternative methods take the geometric characteristics into account and ascertain cavities and clefts where small ligands or peptides can bind. The latter use information on evolutionary conservation or structure alignment methods to identify known interactions in similar proteins that can be assigned to the investigated protein. As a benefit, if the number of similar proteins is reasonably high, it is also possible to get information for structures where only the sequence is known and, with the help of homology modeling, propose three dimensional structures including their interaction hot spots.

2.2.1 Protein-protein interaction prediction

One of the most challenging parts of protein-protein interaction prediction is the combination of properties with independent predictive power in such way that the wide range of possible protein-protein complex formations is captured. Since no basic rule defines the association of proteins and many different types of aggregations with different dominating interactions exist, predictors tend to be more predictive for certain types of association than for others. Protein-protein complexes can be divided roughly into two groups: permanently (homodimers/heterodimers or multimers) or transiently bound. The former type of complexes can be determined with experimental methods – at least up to a certain size – while, for the latter, the duration and conformation of the complex formation determines if it can be captured with experimental techniques – this class includes interesting cases like enzyme-inhibitor or antigen-antibody complexes.

The properties of binding sites can be divided into three groups: 1. Properties of residues; 2. Evolutionary conservation; 3. Data obtained from atomic coordinates; the latter property includes, for example, secondary structure or

solvent accessibility of residues or protein regions (according to de Vries and Bonvin (2008)). Zhou and Qin (2007) and de Vries and Bonvin (2008) analyzed existing predictors which are available as a web-server and evaluated the performance of these servers, using 25 structures from the CAPRI targets and several other datasets. They showed the diversified methods, each targeting one or more of the different properties of binding sites, and their key benefits on very different test sets, including permanent and transient bound complexes.

An increasing amount of information on protein-protein interaction has been gathered during the last decade using experimental methods. Improved techniques in structure identification like X-Ray crystallography or NMR Spectroscopy enrich databases such as the protein database (Berman et al., 2000) with new complexes which are the basis for statistical analysis. This makes it possible to include information from those analysis of known protein binding sites into PPI prediction, beside the individual physicochemical property examination. Out of these, evolutionary methods analyzing the conservation of certain key residues show a high amount of significance.

Evolutionary methods

Residues that fulfill an essential role in the life cycle of a protein are often evolutionary conserved for a family of proteins within one species or through comparable classes of proteins in different species. Those residues play a leading role in protein folding or in the interaction with other molecules (Lichtarge et al., 1996, Lichtarge and Sowa, 2002). Highly conserved residues, at the protein surface not involved in folding, are often found to be part of fundamental protein functions in the cell.

Information from data bases are used to generate classes and subclasses of protein families. Sequence alignment is a frequently used method to identify conserved residues within a family of structures (Lichtarge et al., 1996, Armon et al., 2001, Mihalek et al., 2004). Although residue conservation shows high benefit in identifying protein-protein interaction sites, it is not sufficient to determine the correct binding site for all complexes (Caffrey et al., 2004). This is also, because in many similar structures, single residue mutations can occur, reducing the efficiency of sequence conservation methods and entail more sophisticated methods to cluster families of relevant proteins (La and

Kihara, 2012). Therefore, many PPI predictors based on sequence analysis, combine this information with other data e.g. derived from physicochemical analysis. Some methods focus on sequence conservation and physicochemical property analysis like WHISCY (surface sequence conservation and physicochemical features, de Vries et al. (2006)), consPPISP (position specific sequence conservation and solvent accessibility, Chen and Zhou (2005)), others identify evolutionary coherence using phylogenetic trees or evolutionary trace methods. This methods roughly assemble proteins in families and divide them into subgroups (Lichtarge and Sowa, 2002, Mihalek et al., 2004, La and Kihara, 2012) and might end up with one protein per subgroup and can identify functionally conserved residues by merging subgroups (e.g. ET-viewer (Morgan et al., 2006) or JET (Engelen et al., 2009)).

All these methods have the advantage that only the sequence is necessary to generate the background for the prediction. Backbone or side chain movement does not affect such kinds of analysis, so that bound as well as unbound structures can be used. In contrast the local arrangement of residues often matter for physicochemical analysis (Reuveni et al., 2008, Ruvinsky et al., 2011).

Methods based on physicochemical properties

Amino acid composition and their chemical properties, the arrangement of amino acids as well as the overall polarity or hydrophobicity of the binding site became the target of exhaustive investigations (Bordner and Abagyan, 2005, de Vries and Bonvin, 2008, Ezkurdia et al., 2009). Statistical analysis are based mostly on the comparison of properties of amino acids in the binding site and the rest of the surface of known structures in the protein database (Berman et al., 2000) or other databases. In such analysis, especially statistics on residues not in the known binding site, might be diluted by the lack of knowledge of all interaction sites of a protein. Additionally, the form of the protein taken for analysis might influence the results, e.g. comparison of unbound and bound structures for known protein binding sites (Reuveni et al., 2008). A change in side-chain conformation can influence calculations of solvent accessibility of single residues as well as side-chain rotameric conformational space or side-chain packing and are therefore concerned in recent approaches (Lexa and Carlson, 2010, Ruvinsky et al., 2011, Kozakov et al., 2011). Also none of the investigated properties showed themselves to be ade-

quate to identify all protein binding sites on its own. As a consequence, most predictors combine the detection of several of these physicochemical features as well as additional sequence conservation methods.

Nevertheless, some distinguishing features could be found. Differences in the solvent accessibility of residues had been observed and included into prediction methods, showed larger accessibility in the binding site than on average (Jones and Thornton, 1997a, Porollo and Meller, 2007). This might also be an implication of secondary structure arrangements found in binding sites which showed a preference for beta strand formations over helices and in other cases an enlargement of unstructured loops (Neuvirth et al., 2004), which are often found in inhibitor complexes. Such kind of analysis have been used in ProMate (Neuvirth et al., 2004), giving valid results for many targets and has therefore found it's way into recent consensus prediction methods. Also the ability to adapt different rotameric states is limited compared to residues not in the binding site (Liang et al., 2006). The composition of the residues can also differ so that single residues or combination of residues are more often found within certain binding sites than on the rest of the surface (Lichtarge et al., 1996).

Of all physicochemical features the desolvation property of a protein area has shown to be a promising characteristic and is used directly or indirectly in a variety of methods. The desolvation penalty denotes the energy necessary to remove water from a protein surface and can be calculated as the loss of solvent accessibility upon binding.

Optimal docking areas (ODA) are used to predict binding sites by calculating the solvent accessible surface area (SASA) of patches of residues (Fernández-Recio et al., 2005). A second possibility is to calculate the change in electrostatic energy upon binding using small probes (Fiorucci and Zacharias, 2010a). This method also detects changes in the electric field upon loss of solvent accessibility on the protein's surface by solving the Poisson-Boltzmann equation (Baker et al., 2001) but which comes with the demand of more computational effort than the calculation of SASA. Electrostatic recognition also plays a role in complex formation and formation stability for many complexes (Schreiber and Fersht, 1996, Camacho et al., 1999, Sheinerman et al., 2000). Since also the overall hydrophobicity differs from binding site to non binding site, for single targets exhaustive analysis were made to identify binding sites of pharmaceutical interest (Chennamsetty et al., 2011).

Consensus prediction

Although many of the above mentioned predictors already combine several different approaches, the attempt to use multiple predictors to generate consensus predictions has shown noticeable improvement. Cport (de Vries and Bonvin, 2011) for example uses with WHISCY (de Vries et al., 2006), PIER (Kufareva et al., 2007), ProMate (Neuvirth et al., 2004), cons-PPISP (Chen and Zhou, 2005), SPPIDER (Porollo and Meller, 2007) and PINUP (Liang et al., 2006) six stand alone predictors to generate a combined prediction. Three modes in Cport allow a precise prediction up to a strong overprediction, including none up to all of the predictions of the single predictors. Meta-PPISP (Qin and Zhou, 2007) includes the three predictors cons-PPISP, ProMate and PINUP and generates a new scoring using a linear regression method. A third consensus predictor MetaPPI (Huang and Schroeder, 2008) generates surface patches with a score depending on the occurrence in the results of the single predictors as binding site. PPISP, SPIDDER, PINUP, PPI-PRED (Bradford and Westhead, 2005) and ProMate were used to generate the initial predictions.

All these meta prediction tools were designed to fulfill a specific task. Cport was developed for data driven docking with HADDOCK (Dominguez et al., 2003) providing a list of residues for restrained docking. MetaPPI was developed to filter results after docking with the BDOCK docking program (Huang and Schroeder, 2008) and therefore provides continuous patches. Meta-PPISP was designed to combine the different approaches of the used predictors to consider all possible types of physicochemical features as well as data from evolutionary conservation but was not designed towards a specific complex prediction approach. Since none of these predictors is able to identify all possible binding sites, various predictors as well as meta-predictors have been developed recently and will be in the foreseeable future.

2.2.2 Protein-small ligand interaction prediction

Small organic ligands or small peptides are often bound in concave regions on the protein's surface. Those shallow areas or cavities are often found to be of observable physicochemical composition with a large hydrophobic areas and polar residues with distinct functionality (Miller and Dill, 1997, Liang et al., 1998). Hydrophobicity, desolvation, electrostatics and sequence conservation have been found significant to detect small ligand binding sites (Burgoyne and

Jackson, 2006). The often predetermined location of the ligand in a specific arrangement of residues enables to focus on geometric properties of the surface, such as deep cavities, as an additional prediction criterion. Most of the small ligands are found in cavities (Laskowski et al., 1996, Fuller et al., 2009), which makes geometry-based cavity detection a valuable technique.

Beside the good performance of geometry-based cavity detection (also called pocket detection), these methods often do not allow prediction of both, the correct position of the ligand within the pocket and the binding sites' hot spots. Whether a pocket fulfilling geometric criteria for binding small molecules is actually able to bind small molecules is part of further investigations. Subsequently the question evolves of what kind of small molecules can bind in these detected pockets and is this cavity druggable, id est of interest for medical applications. The first problem is treated by different methods to exclude parts of the pockets from the predicted binding area within a larger pocket: DogSite extracts subpockets (Volkamer et al., 2010), ROLL defines the deepness of cavities (Yu et al., 2010) and PocketPicker uses shape descriptions to define the buriedness of the cavity (Weisel et al., 2007, 2009), to name but a view examples of recent techniques. The binding sites' hot spots detection is e.g. done by placing probes on the surface and calculating favorable interactions (Laurie and Jackson, 2005), or by adding calculations of solvent accessible surface areas (SASA) and conservation analyzes to the pocket detection (Huang and Schroeder, 2006).

More sophisticated approaches not only try to identify the binding site but also attempt to measure binding affinities or suggest chemical components as favorable binders. This is done for example, using molecular dynamics simulations or fragment docking to calculate binding energies within pockets found to attract small organic ligands (Seco et al., 2009, Kozakov et al., 2011).

Geometry-based approaches - cavity detection

Cavity detection algorithms base on the assumption that small ligands bind in one of the largest clefts on the protein's surface. Therefore, all algorithms of this class try to identify suitable cavities on the protein's surface. After the cavities on the protein's surface are detected, algorithms use either geometric criteria to identify the most favorable pocket (free volume, deepness, etc.) or energetic evaluations.

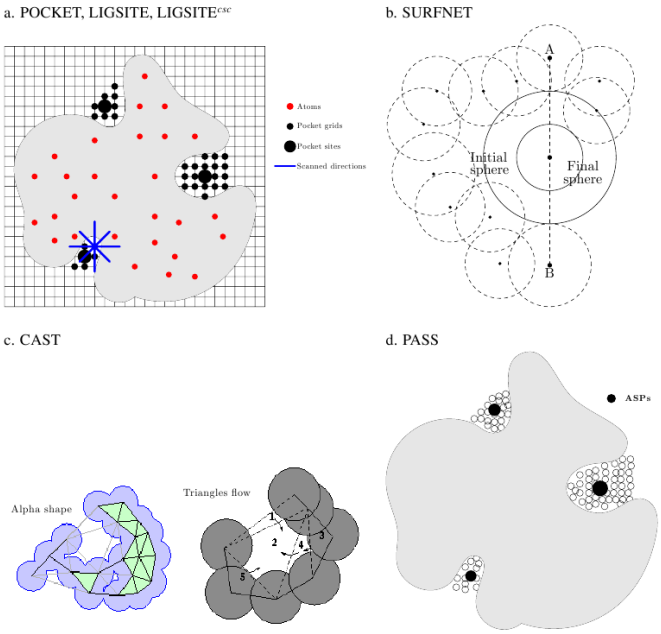


Figure 2.2 – Figure taken from Huang and Schroeder (2006). The procedure of several geometric binding site predictors are shown. A) The basic algorithm of Pocket and LIGSITE (Levitt and Banaszak, 1992, Hendlich et al., 1997) places a grid over the protein and discards grid points which are not placed on grid lines intersecting with the protein. B) The Surfnet algorithm (Laskowski, 1995) places probes within two atoms and return the largest spheres not intersecting with the protein. C) CAST (Liang et al., 1998) describes pockets via merged clusters of triangles between protein atoms. D) PASS (Brady and Stouten, 2000) places probes on the protein's surface and detects those with many protein atom contacts. This procedure is repeated until no new probe fulfills the minimal neighbour criteria.

A method often used to detect cavities, is to set up a grid covering the complete protein and beyond. The LIGSITE algorithm deletes all grid points which intersect with the protein and spreads lines from each leftover grid point along all axis. The grid points are kept that are nested on a line which intersects with both ends with the protein (Hendlich et al., 1997). Other methods place large probes on the surface not intersecting with the protein itself and keep all grid points that have not been covered by the large probes but are in contact with the protein's surface as for example ROLL does (Yu et al., 2010).

The PASS algorithm (Brady and Stouten, 2000) uses small probes to calculate atom contacts within a given radius of 8 Å. This procedure is repeated, also covering already placed probes, until no new ones can be found which satisfy the atom neighbor criterion. Probes with many atom contacts are kept and ordered concerning the number of probes defining a cavity (Brady and Stouten, 2000) – for a graphical illustration of widely used algorithms see Figure 2.2.

Cavities are often marked with single spots in the center of the prediction or deliver the grid used, often spreading through the complete cavity (Hendlich

et al., 1997, Yu et al., 2010), without further investigation of possible binding hot spots. Some approaches try to reduce the area predicted by deleting edge points of the accepted grid points (Yu et al., 2010) or by dividing pockets into subpockets if narrow regions within the predictions could be found (Volkamer et al., 2010). Cavities are than ranked according to their free volume, other shape criteria e.g. the deepness of the cavity (Weisel et al., 2007, Yu et al., 2010) or physicochemical properties of the cavity like solvent accessibility or hydrophobicity (Halgren, 2007). Cavity detection is also used as a first step towards more exhaustive methods using the detected cavities as initial point (Venkatachalam et al., 2003, Huang and Schroeder, 2006).

Surface property approaches and methods of comparison

Besides pure geometric cavity detection, energy-based methods or combined methods have been developed. An extension of the LIGSITE algorithm combines the volume based cavity detection with calculation of solvent accessibility and residue conservation (LIGSITEcsc by Huang and Schroeder (2006)). The Q-SiteFinder algorithm calculates van der Waals interactions between a methyl-group placed on the surface of the protein and the protein atoms in contact with the probe (Laurie and Jackson, 2005). Also hydrophobic regions have been investigated and successfully used to predict ligand binding sites with the Fuzzy-Oil-Drop method (Brylinski et al., 2007).

A more exhaustive approach uses molecular dynamics simulations to investigate the binding of alcohol within an aqueous solution on the surface (Seco et al., 2009). The small alcohol was used to identify the binding sites for eight targets and to estimate the maximum binding free energy ΔG and the dissociation K_d for each binding site, showing good agreement with experiments (Seco et al., 2009). A grid was used to estimate the amount of alcohol bound to the surface compared the occurrence in the bulk. The calculation of the binding free energies also allowed a more sophisticated ranking of the binding sites and revealed informations for druggable sites (Seco et al., 2009).

Fragment based methods have also been used to identify putative binding sites, binding geometries and chemical components most likely binding in the identified pockets. Often grid based approaches are used to place chemical components from a large library on the protein's surface and calculate the energy change upon binding (methods reviewed in Hajduk and Greer (2007)). Chemical fragments bound to the surface with high affinity are clustered afterwards and large clusters or combinations of small and independent clusters are used to identify binding sites. Also, the composition of different chemical fragments within one cluster is used to predict favorable attributes for possible ligands. With a smaller amount of fragments but with the inclusion of evolutionary information Kozakov et al. (2011) were able to identify the binding hot spots of 15 protein-protein inhibition sites and determined their druggability. Since such kind of approaches are time consuming, only small test sets can be investigated but, as a benefit, more informations than the simple geometry of the binding site is generated and often, e.g. with molecular dynamics simulation, flexibility of the protein is included (Seco et al., 2009, Kozakov et al., 2011, Schmidtke et al., 2011a).

The higher amount of time and computational power that has to be invested, compared to strictly geometric methods, comes with more information on the binding spots. For pure detection of the cavities, geometric based method already show very good results – up to 90% of the binding sites were detected in the three top ranked predictions, depending on the test set (Huang and Schroeder, 2006, Yu et al., 2010, Volkamer et al., 2010) and on the criterion chosen for hit counts (Chen et al., 2011) – but do not reveal any information beyond the indication of the most likely cavity.

Some methods also include conservation properties in their prediction (Glaser et al., 2003, Huang and Schroeder, 2006, Kozakov et al., 2011) which have already shown good performance for protein-protein binding sites. FINDSITE (Brylinski et al., 2007) enlarges the idea of sequence comparison: Known proteins with bound ligands are aligned with the targeted protein if sequence identity and RMSD does not exceed a given threshold. Ligand positions are extracted and clustered and the geometric center of these clusters are returned as prediction (Brylinski et al., 2007). With a sufficiently large template library, sequence comparison showed a very good performance (Chen et al., 2011).

2.3 Protein-protein docking

The purpose of computational protein-protein docking methods is to predict the structure of a protein-protein complex based on the structure of the isolated protein partners. One general strategy is to use rigid bodies for the docking procedure, another one to add flexibility in terms of side chain or backbone movements. The latter follows the natural behavior of many biomolecules but adds the challenge of predicting the protein's flexibility to the docking problem.

In the most interesting cases only the unbound geometry, if at all, and no information on the change upon binding of the protein is known. The structure of the unbound protein might differ several Ångstroms from its bound form, including large side-chain movements in the putative binding site. If the structure of the isolated partner protein is not known in atomic resolution but as the sequence of amino acids, it is often possible to build structures based on sequence homology to a known structure using comparative modeling methods.

2.3.1 Rigid protein-protein docking methods

A variety of computational methods have been developed in recent years to efficiently generate a large number of putative binding geometries. The initial stage consists typically of a systematic docking search keeping partner structures rigid (Bonvin, 2006, Vajda and Kozakov, 2009). Subsequently, one or more refinement and scoring steps of a set of preselected rigid docking solutions are added to achieve closer agreement with the native geometry and to recognize near-native docking solutions preferentially as the best or among the best scoring complexes. In the initial search some unspecific sterical overlap between docking partners is typically tolerated to implicitly account for conformational adjustment of binding partners (e.g. Pons et al. (2010)).

Among the most common are geometric hashing methods to rapidly match geometric surface descriptors of proteins (Norel et al., 1994) and fast Fourier transform (FFT, see Figure 2.3) correlation techniques to efficiently locate overlaps between complementary protein surfaces (Katchalski-Katzir et al., 1992). In the latter approach the two protein partners are represented by cubic grids, the grid points are assigned discrete values for inside, outside and on the surface of the protein. A geometric complementarity score can be calculated for the two binding partners by computing the correlation of the two grids representing each protein. Instead of summing up all the pair products

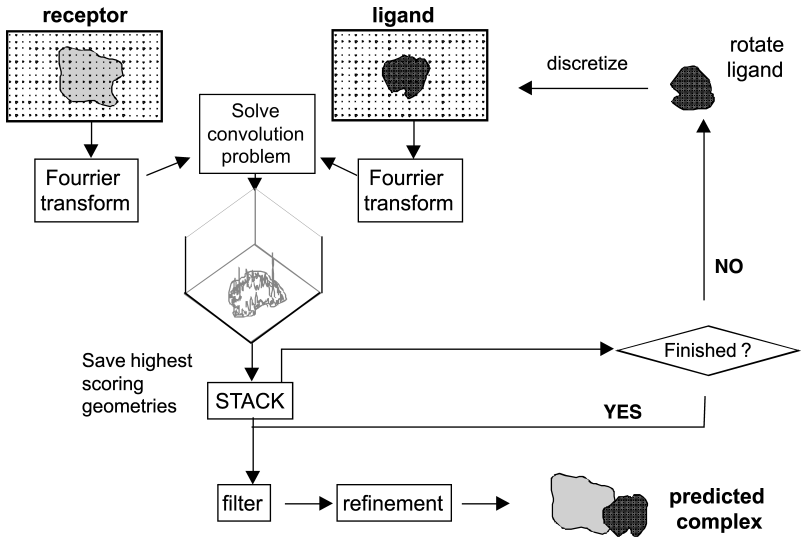


Figure 2.3 – Rigid protein-protein docking using Fast Fourier Transformation to solve a correlation problem. Receptor and ligand proteins are discretized on three-dimensional grids and are portioned into inside, surface and outside regions, respectively. Matching of surfaces is measured by the overlap of surface regions. For each ligand rotation with respect to the receptor the correlation problem is solved using Fast-Fourier-Transformation (FFT). After filtering and possible refinement steps solutions with high overlap of surface regions (high surface complementarity) are collected as putative solutions.

of the grid entries one can make use of the Fourier correlation theorem. The corresponding correlation integral can easily be computed in Fourier space. The discrete Fourier transform for the receptor grid needs to be calculated only once. Due to the special shifting properties of Fourier transforms, the different translations of the ligand grid with respect to the receptor grid can be computed by a simple multiplication in Fourier space. This process is repeated for various relative orientations of the two proteins. A disadvantage of standard Cartesian FFT-based correlation methods is the need to perform FFTs for each relative orientation of one protein molecule with respect to the partner. This can be avoided by correlating spherical polar basis functions that represent, for example, the surface shape of protein molecules. It has been successfully applied in the field of protein-protein docking (Ritchie et al., 2008). Recently, new multidimensional correlation methods have been developed that allow the correlation of multi-term potentials. Each function needs to be expressed in terms of spherical basis functions characterizing the surface properties of the protein partners (Ritchie et al., 2008, Zhang et al., 2009b).

Geometric hashing is another common approach to identify possible protein-protein arrangements. Each protein surface is discretized as a set of triangles, which are stored in a hash table. By means of a hash key similar matching triangles on the surface of protein partners can be found quickly. During dock-

ing, these triangles comprise points on a molecular surface, having a certain geometrical (concave, convex) or physicochemical (polar, hydrophobic) character. By matching triangles belonging to different molecules and being of complementary character, putative complex geometries can be generated.

A third class of methods uses either Brownian Dynamics (Gabdoulline and Wade, 2002, Schreiber et al., 2009), Monte Carlo, or multi-start docking minimization to generate large sets of putative protein-protein docking geometries (Zacharias, 2003, Fernández-Recio et al., 2003, Gray et al., 2003). These methods have in principle the capacity to introduce conformational flexibility of binding partners at the initial search step. Since these approaches are computational more expensive compared to FFT based correlation methods or geometric hashing a search is frequently limited to predefined regions of the binding partners (Bonvin, 2006). Alternatively, it is possible instead of atomistic models to employ coarse-grained (reduced) protein models to perform systematic docking searches. With such reduced protein models it is possible to optimize docking geometries starting from tens of thousands of protein start configurations (Zacharias, 2003, May and Zacharias, 2005). In order to limit the number of putative complex structures generated during an initial docking search cluster analysis is typically employed to reduce the number to a subset of representative complex geometries.

Recently, the limitations of rigid docking strategies combined with a rescoring step have been systematically investigated by Pons et al. (2010). The authors applied a combination of rigid FFT-correlation based docking and rescoring using the pyDock approach (Cheng et al., 2007). PyDock combines electrostatic Coulomb interactions with a surface-area-based solvation term (and an optional van der Waals term). The protocol showed a very good performance for most proteins that undergo minor conformational changes upon complex formation ($<$1 Å Rmsd between unbound and bound structures) but unsatisfactory results for cases with significant binding induced conformational changes or applications that involved homology modeled proteins.

A conclusion is that more specific scoring requires, at the same time, an improvement of the prediction accuracy of proposed binding modes in terms of deviation from the experimental binding interface. It also indicates the coupling between realistic scoring and accurate prediction of the complex structure.

2.3.2 Flexible protein-protein docking methods

A significant fraction of experimentally known protein-protein complexes belongs to the class that show only little conformational change upon complex formation. In such cases, it is possible to separate the initial rigid search from a subsequent flexible refinement and re-scoring step. However, for many interesting docking cases with large associated conformational changes explicit consideration of conformational flexibility during the entire docking procedure or at an early refinement step appears to be necessary.

One possibility to directly use computationally rapid rigid docking algorithms is to indirectly account for receptor flexibility by representing the receptor target as an ensemble of structures. The structural ensemble can, for example, be a set of structures obtained experimentally (e.g. from nuclear magnetic resonance (NMR) spectroscopy) or can be formed by several structural models of a protein. It is also possible to generate ensemble from MD simulations (Grünberg et al., 2004) or from distance geometry calculations (de Groot et al., 1997). Docking to an ensemble increases the computational demand and due to the large number of protein conformations may also increase the number of false positive docking solutions. In the field of small-molecule docking, a variety of ensemble based approaches have been developed in recent years (reviewed in Totrov and Abagyan 2008). Cross docking to ensembles from MD simulations have also been used to implicitly account for conformational flexibility in protein docking (Krol et al., 2007). Mustard and Ritchie 2005 generated protein structures deformed along directions compatible with a set of distance constraints reflecting large scale sterically allowed deformations. Subsequently, the structures were used in rigid body docking searches to identify putative complex structures.

Conformer selection and induced fit mechanism of protein-protein association have been compared by ensemble docking methods using the RosettaDock approach (Chaudhury and Gray, 2008). The RossettaDock approach includes the possibility of modeling both side chain as well as backbone changes for a set of starting geometries obtained from a low-resolution initial search (Wang et al., 2007). The method was able to successfully select binding-competent conformers out of the ensemble based on favorable interaction energy with the binding partner (Chaudhury and Gray, 2008). It was recently shown that the Rosetta approach can also be used to simultaneously fold and dock the

structure of symmetric homo-oligomeric complexes starting from completely extended (unfolded) structures of the partner proteins (Das et al., 2009).

For a limited number of start configurations (in the case of knowledge of the binding sites) it is possible to combine docking with molecular dynamics (MD) or Monte Carlo (MC) simulations. This allows, in principle, for full atomic flexibility or flexibility restricted to relevant parts of the proteins during docking. The HADDOCK program employs MD simulations including ambiguous restraints to drive the partner structures towards the approximately known interface (Dominguez et al., 2003). The success of HADDOCK in many Capri rounds for targets where some knowledge of the interface region was available underscores also the benefits of treating flexibility explicitly during early stages of the docking process. For protein-protein docking it is always helpful to include some knowledge on the putative interaction region. In these cases the docking problem can often be reduced to the refinement of a limited set of docked complexes close to the known binding site. Fortunately, for proteins of biological interest and with experimentally determined structure there is often also some biochemical (e.g. mutagenesis) data available on residues involved in binding to other proteins.

Protein partner structures can undergo not only local adjustments (e.g. conformational adaptation of side chains and backbone relaxation at the interface) during association but also more global conformational changes that involve for example large loop movements or domain opening-closing motions. Proteins in solution are dynamic and the question to what extent the accessible conformational space in the unbound form overlaps with the bound conformation, has been at the focus of several experimental and computational studies. Elastic Network Model (ENM) calculations are based on simple distance dependent springs between protein atoms and despite its simplicity are very successful to describe the mobility of proteins around a stable state (Bahar et al., 1997, 2007).

Systematic applications to a variety of proteins indicate that there is often significant overlap between observed conformational changes and a few soft normal modes obtained from an ENM of the unbound form (Tobi and Bahar, 2005, Keskin et al., 2008, Bakan and Bahar, 2009). ENM-based normal mode analysis has been used to identify hinge regions in proteins (Emekli et al., 2008) and can also be used to design conformational ensembles. It is also possible to use soft collective normal mode directions as additional variables

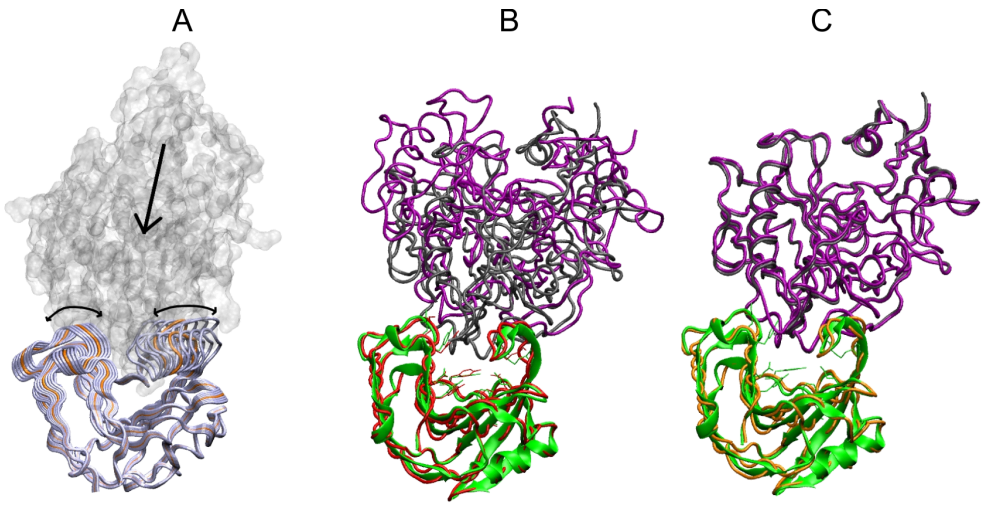


Figure 2.4 – Docking including minimization in soft flexible normal modes (A) Illustration of the flexible docking process of the taxi-inhibitor protein (pdb3hd8) to the xylanase target receptor protein (pdb1ukr) using the ATTRACT program (May and Zacharias, 2005). Putative translational motion of the inhibitor during docking approach is indicated by an arrow and the deformability of the xylanse by the superposition of several structures deformed in the softest normal mode (grey backbone tube representation). Best possible docking solutions (in pink) of the inhibitor relative to the bound (green cartoon) and unbound xylanase (red tube) are shown for rigid (B) and flexible (C) docking employing minimization along the 5 softest normal mode directions of the xylanase receptor protein. The placement of the inhibitor in the experimental

during docking by energy minimization (Zacharias and Sklenar, 1999, May and Zacharias, 2005). This allows the rapid relaxation of protein structures on a global scale involving much larger collective displacements of atoms during minimization then conventional energy minimization using Cartesian or other internal coordinates (see Figure 2.4). The application of refinement in normal mode variables has been applied successfully in a number of studies (May and Zacharias, 2005, Mashiach et al., 2010). Based on a coarse-grained protein model in the ATTRACT docking program (Zacharias, 2003) it has also been used in systematic docking searches to account approximately for global conformational changes already during the initial screen for putative binding geometries (May and Zacharias, 2008). In cases where protein partners undergo collective changes that overlap with the NM variables, the approach can result in improved geometry and ranking of near-native docking solutions and can also lead to an enrichment of solutions close to the native complex structure.

2.3.3 The ATTRACT docking program

The ATTRACT docking program (Zacharias, 2003, Fiorucci and Zacharias, 2010a) employs a coarse-grained protein representation with two pseudo atoms per residue representing the main chain (located at the backbone nitrogen and backbone oxygen atoms, respectively). Small amino acid side chains (Ala, Asp, Asn, Cys, Ile, Leu, Pro, Ser, Thr, Val) are represented by one pseudo atom (geometric mean of side chain heavy atoms). Larger and more flexible side chains are represented by two pseudo atoms to account for the shape and dual chemical character of some side chains. Effective interactions between pseudo-atoms are described by soft distance(r_{ij})-dependent Lennard-Jones(LJ)-type potentials of the following form:
Attractive pairs:

$$V = \epsilon_{AB}\left(\left(\frac{R_{AB}}{r_{ij}}\right)^8 - \left(\frac{R_{AB}}{r_{ij}}\right)^6\right) + \frac{q_i qj}{\epsilon(r_{ij})r_{ij}} \tag{2.1}$$

Repulsive pairs:

$$V = -\epsilon_{AB}\left(\left(\frac{R_{AB}}{r_{ij}}\right)^8 - \left(\frac{R_{AB}}{r_{ij}}\right)^6\right) + \frac{q_i qj}{\epsilon(r_{ij})r_{ij}}, \quad if\, r_{ij} > r_{min} \tag{2.2}$$

or

$$V = 2e_{min} + \epsilon_{AB}\left(\left(\frac{R_{AB}}{r_{ij}}\right)^8 - \left(\frac{R_{AB}}{r_{ij}}\right)^6\right) + \frac{q_i qj}{\epsilon(r_{ij})r_{ij}}, \quad if\, r_{ij} <= r_{min} \tag{2.3}$$

where R_{AB} and ϵ_{AB} are effective pairwise radii and attractive or repulsive Lennard-Jones parameters. At the distance r_{min} between two pseudo atoms the standard LJ-potential has the energy e_{min}. A Coulomb type term accounts for electrostatic interactions between real charges (Lys, Arg, Glu, Asp) damped by a distance dependent dielectric constant ($\epsilon = 15r$). This form allows for purely repulsive interacting pseudo-atom pairs. The attractive and repulsive parameters for each pseudo-atom pair were iteratively optimized by minimizing the root-mean-square-deviation of near-native docking minima and by comparing the scoring of near-native minima with many high-scoring decoy complexes (published in Fiorucci and Zacharias 2010b).

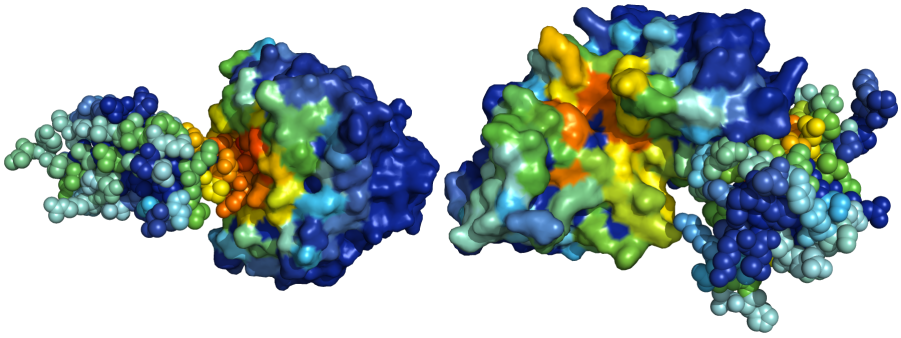


Figure 2.5 – Prediction of putative protein binding interfaces Predictions were performed with the meta-PPISP server (Qin and Zhou, 2007) on the partner proteins of an enzyme inhibitor complex (pdb2sic, left panel) and partners of a second complex (pdb1buh, right panel). In each case one partner is represented as surface or collection of spheres, respectively. Protein partners are slightly displaced from the complexed state to indicate the native binding interface. Red indicates high predicted probability for a residue to be in the binding site and dark blue represents a low probability. Left example: The results match the real binding site. Right panel: The prediction for the smaller protein overlaps with the real binding site while for the larger protein residues quite far apart from the correct binding site are marked as putative binding site residues.

2.3.4 Application of protein-protein binding site prediction in protein-protein docking

If no experimental data on binding sites is available, binding site prediction methods can provide useful data for information driven docking (for examples of predictions compare Figure 2.5). This type of information can be very helpful in order to limit the docking search or to evaluate and filter docking results. Docking approaches like HADDOCK (Dominguez et al., 2003) are based on applying restraints derived from experimentally known binding sites or predicted binding regions. Several different approaches exist to identify putative protein-protein binding sites. These methods focus on different characteristics of protein interaction sites like solvent accessibility (Chen and Zhou, 2005) or desolvation properties (Pons et al., 2010, Fiorucci and Zacharias, 2010a) and in many cases on combining different surface properties (Neuvirth et al., 2004, Liang et al., 2006).

The data generated by predictors using one or more binding site features is presented either as a list of residues (Qin and Zhou, 2007) or as a patch on the protein's surface (Jones and Thornton, 1997a,b). Patch methods generate one or more patches of circular shape which can be found close to each other or distributed on the surface, sometimes additionally center coordinates of these spots are given. In the other case residues from residue list predictors do not have to be nearby each other but are often clustered afterwards to

receive a joint prediction at one or more spots on the protein's surface. Since proteins often have more than one binding site, prediction tools can indicate a correct binding site but maybe for the wrong binding partner. The binding site predictions can be used to evaluate possible predicted docking geometries but also to generate artificial binding sites around the prediction to bias the docking run towards a desired region. Additionally, predictions can be used to discard complexes with a low overlap of predicted contacts after a systematic docking run or re-score previously sampled structures.

2.4 Molecular dynamics simulations

The function of a biomolecule depends on its three dimensional structure and its flexibility upon contact with other biomolecules: internal motions, conformational backbone as well as side-chain movements are an essential part in many biological processes. To calculate and observe the dynamics of such systems, Molecular Dynamics (MD) simulation is a widely used tool. Software packages including the Molecular Dynamics algorithms compute the movement and the interactions of and between atoms over time such that it is possible to create a trajectory of the dynamics of a system from pico- up to microsecond timescales. The application area of MD simulations ranges from refinement of experimentally solved as well as computationally designed structures over sampling of the dynamics for the investigation of the behavior of the system to observations of the equilibrium state (Karplus, 2002).

2.4.1 Molecular dynamics simulation steps and algorithms

As an approximation of the more accurate but very time demanding quantum chemical simulations, Newtons second law

$$F_i = m_i a_i \tag{2.4}$$

is used in a classical description of the system. To calculate the effective forces between the atoms classical mechanical force fields are used which include approximations of interactions based on experimental methods as well as quantum chemical calculations. Such a force field can treat interactions of atoms, not covalent bound to each other, with electrostatic Coulomb and Lennard-

Jones potentials and atoms covalent bound with terms for bond length, rotations around bonds and angles:

$$\begin{aligned} V = & \sum_{i<j} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} + \sum_{i<j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) \\ & + \sum_{bonds} \frac{1}{2} k_{ij}^{b} (r_{ij} - b_{ij}^{0})^2 \\ & + \sum_{angles} \frac{1}{2} k_{ijk}^{\theta} (\theta_{ijk} - \theta_{ijk}^{0})^2 \\ & + \sum_{dihedrals} k^{\phi} (1 + \cos n(\phi - \phi^{0})) \end{aligned} \tag{2.5}$$

V describes a classical force field term with q_i the charge of atom i, r_{ij} the distance between atom i and atom j, Lennard-Jones parameter Aij, B_{ij}, the constants $k_{ij}^{b}, k_{ijk}^{\theta}, k^{\phi}$ for bond length, angles and dihedrals and $b_{ij}^{0}, \theta_{ijk}^{0}, \phi^{0}$ equilibrium values. Those values differ from force field to force field and are often optimized for special tasks. Commonly used force fields (FF) are AMBER (Weiner et al., 1984), CHARMM (Brooks et al., 1983), OPLS/AA (Jorgensen et al., 1996) or GROMOS (van Gunsteren et al., 2002). It is possible to simulate molecules at a preset temperature. This allows to simulate biomolecules at physiological as well as high and low temperatures, depending on the question to be answered. This kind of temperature coupling is done via an external heat bath e.g. the Berendsen thermostat (Berendsen et al., 1981).

Newtons equation of motion (see equation 2.4) has to be integrated for each time step. Two widely implemented algorithms, the Leap-Frog algorithm (van Gunsteren and Berendsen, 1988) and the Verlet algorithm (Gillia and William, 1992), are used for numerical integration. The time step has to be conform to small molecular movements like hydrogen bond vibration which is less than 1 femtosecond but can be extended due to the use of constraining algorithms like SHAKE (Hünenberger et al., 2001) or LINCS (Hess et al., 1997) to up to 2 femtoseconds. Those algorithms correct after each time step the values for bond length between covalently bound atoms.

To expand MD simulations to a more realistic scenario solvents can also be included into simulations, providing a more realistic environment but come with an increase of computational costs. The basic set up of an MD simulation consists of one or more molecules to be studied, a shell of water around the

molecules surface and ions to neutralize the simulation box. Since it is not possible to realize large boxes of water around the molecules, periodic replication of the box in all directions are created so that water molecules, ions or other solvent components from one side of the simulation box can interact with those on the other side. To avoid self interaction, the box size has to be chosen large enough so that molecules cannot interact with them self. Pairwise electrostatic interactions are usually restricted to a cut-off value while the long range interactions are calculated with a grid based Ewald summation. Also Lennard-Jones interactions are reduced to cut off of about 10 Å. Electrostatic and Lennard-Jones interaction calculation is one of the most costly parts and with restrictions to cut-off distances speed up is traded for slight loss of accuracy. It is in general possible to do simulations as accurate as desired but with each enhancement in accuracy a loss of simulation size or time is associated; however, this loss of time or size can be compensated by the technical development, allowing longer simulation times, larger systems, faster calculations or more accuracy.

Classical molecular dynamics simulations are limited for example in the aspect of bond formation and breaking which can be overcome with the inclusion of quantum chemical simulations as part of a combined simulation. Also the simulation length is limited, due to the computational cost of a simulation of a biomolecule in water. Simulation length depends also on the size of the system, so that extremely large protein assemblies can hardly be simulated for a reasonable time. With specially designed hardware, millisecond simulations of smaller proteins are already reachable (Shaw et al., 2007, 2009) and offer insight into time scales including e.g. protein folding and protein associations for larger complexes.

2.4.2 Solvation of biomolecules

Biomolecules are within the cell in an aqueous environment in contact with a variety of other molecules, ions and small organic compounds. Water is the basis for most solvents used in MD simulations as it is in the natural environment of the cells. Besides water molecules, ions are added to the solvent to neutralize the overall electrostatic net charge commonly but are sometimes added to simulations to mimic natural or experimental concentrations.

Many processes in the cell depend on the special properties of water e.g. folding of a protein due to the hydrophobic effect (Tsai et al., 1997, Chandler, 2005) or the long range electrostatic interactions. To meet the basic requirements of a protein, within a simulation or further analysis, an aqueous solvent has to be considered either explicitly, each atom of the water molecule is treated explicitly in the calculations of the force field, or implicitly, thus accounting for the overall effect of the water on other biomolecules as an approximation in the force computation.

Explicit water model

Different models of water molecules have been developed which model the three atoms explicitly (SPC/SPCE (Berendsen et al., 1981, 1987) or TIP3P (Jorgensen et al., 1983)) or add additional features e.g. the center of mass (TIP4P (Jorgensen et al., 1983)) or charge distributions (TIP5P (Mahoney and Jorgensen, 2000)) as pseudo atoms. These water models mimic the behavior of water slightly differently (Zielkiewicz, 2005) and were optimized for different force fields. Besides the large costs for the calculations of explicit water in MD simulations, models with additional pseudo atoms increase the computational cost accessorily.

Solvents play a crucial role for the formation and folding of biomolecules due to direct interactions (Ebbinghaus et al., 2007, Reichmann et al., 2008) or the hydrophobic effect (Chandler, 2005). Single water molecules are also elementary for the function of several biomolecules in the cell (Sue et al., 2001, Carla and Mattos, 2002). Analysis of the water around biomolecules have been performed using computer simulations (Makarov et al., 1998a, Henchman and McCammon, 2002b, Samsonov et al., 2008) or comparison of experimentally revealed bound water (Rodier et al., 2005, Barillari et al., 2011). One of the important interactions of water with biomolecules, besides the general long range electrostatic interactions, is the formation of hydrogen bonds (Tarek and Tobias, 2002) which allow strong interactions for the biomolecule with the solvent e.g upon binding processes or to stabilize proteins and complexes (Reichmann et al., 2007, 2008). One limitation of molecular dynamics simulations is that proton transfer from water to a protein can not be represented but can be included with time consuming quantum mechanical (QM) simulations combined with classical methods (Sproviero et al., 2008).

As used for experiments (Mattos and Ringe, 1996, 2001) also organic compounds are added to simulations (Schellman, 1990, 2003, Shulgin and Ruckenstein, 2005). These computational experiments try to describe the behavior of the protein in an environment with a high amount of organic molecules or take advantage of the physicochemical features of those molecules to mimic interactions with larger partners (Seco et al., 2009).

Implicit water model

Beside the explicit calculation of each water molecule, implicit water models can be used to mimic the overall effect of water on biomolecules. Due to the lower number of atoms involved in inter molecular force calculation, the computational cost can often be reduced. This allows to increase the simulation time and enables to sample more of the conformational space. Additionally the cost for the equilibration of explicit water around the molecule – essential in explicit water simulations – can be omitted.

To calculate the energies of solvated solutes the free energy of the transfer of a molecule into the solvent is summed up with the potential energy of the molecule in vacuum and can be written as (Onufriev, 2010)

$$E_{tot} = E_{vac} + \Delta G_{solv} \tag{2.6}$$

with the decomposition of ΔG_{solv} as

$$\Delta G_{solv} = \Delta G_{el} + \Delta G_{nonpolar} \tag{2.7}$$

Thereby $\Delta G_{nonpolar}$ is often approximated by the calculation of solvent accessible surface areas (SASA) or separation in repulsive and attractive interactions (Tan and Luo, 2007) with low computational demand, while the approximation of the electrostatic interactions ΔG_{el} is the most time consuming step (Onufriev, 2010).

A frequently used approximation is the Poisson-Boltzmann (PB) model (Fixman, 1979, Tironi et al., 1995, Im et al., 1998). The molecule is represented in all atom resolution, while the solvent is treated as a continuum. The electric field between the charges of the solute and the continuum of the

solvent can be calculated with the help of the Poisson-Boltzmann equation (Warwicker and Watson, 1982, Klapper et al., 1986):

$$\nabla[\epsilon(r)\nabla\Phi(r)] = -4\pi\rho(r) - 4\pi\lambda(r)\sum_i z_i c_i exp(-z_i\Phi(r)/k_bT) \tag{2.8}$$

with ϵ the dielectric constant, Φ the electrostatic potential, ρ the charge of the molecule, λ the Stern function, z_i charge of an ion and c_i the bulk density of the ion. To solve this equation, numerical methods have been implemented to approximate the analytical solution of the Poisson-Boltzmann equation (e.g. Baker et al. (2001)).

Another approximation is the generalized Born (GB) model which represents the protein as N particles with distinct radii r_i and charge q_i settled in a medium of permittivity ϵ. It has emerged from the formula of a single ion in solvent expressed by Born:

$$\Delta G_{el} = -\frac{q_i^2}{2p_i}\big(1 - \frac{1}{\epsilon}\big) \tag{2.9}$$

with q_i the charge and p_i the van der Waals radius of the ion. For the general approach, the radii have to be approximated within each time step to account for the degree of burial of a single atom in context to the surrounding atoms. The interior of each sphere is filled with a dielectric medium and surrounded by a continuum of higher dielectric constant representing the solvent. For a multi-atom system, e.g a protein, the formula for the calculation of ΔG_{el} is often (e.g. in the simulation software suite AMBER (Weiner et al., 1984)) written as

$$\Delta G_{el} = -\frac{1}{2}\sum_{ij}\frac{q_i q_j}{f_{GB}(r_{ij}, R_i, R_j)}\big(\frac{1}{\epsilon_{in}} - \frac{exp[-\kappa f_{GB}]}{\epsilon_{out}}\big) \tag{2.10}$$

with r_{ij} the distance between the atoms, R the effective Born radii, ϵ_{in} and ϵ_{out} the dielectric constant for the interior and the exterior, respectively, and κ the Debye-Hückel screening parameter which accounts for the salt concentration in the solvent (Srinivasan et al., 1999). f_{GB} denotes a smoothing function often used (due to Still et al. (1990))

$$f_{GB} = \sqrt{r_{ij}^2 + R_i R_j exp\big(\frac{-\lambda r_{ij}^2}{R_i R_j}\big)} \tag{2.11}$$

to correct the radii for the electrostatic calculations, where λ can be varied (Onufriev, 2010). To calculate the effective Born radii and the border between the molecule and the solvent, many methods have been developed (Hawkins et al., 1995, Ghosh et al., 1998, Onufriev et al., 2000, Romanov et al., 2004), based on different approximations for volume and surface, a comparison of different methods can be found in Onufriev (2010).

Implicit solvent approaches allow to account for solvation effects with a low amount of computational effort compared to explicit solvent simulations. Depending on the chosen model (PB vs. GB) and the level and type of approximation made within these models, different levels of accuracy can be reached. This makes it possible to choose the optimal modeling settings, depending on the demands of the simulation and the available computer power.

2.5 Motivation

As already noted, most interactions between proteins and their binding partners occur in an aqueous media and therefore, water molecules and other compounds of the solvent interact with the protein. The protein's physiochemical properties influence the aqueous solvent at close distance to the protein's surface. Hydrophobic regions on the surface are expected to be differently populated by water molecules than the hydrophilic areas and highly bound water molecules are more likely to be found in polar cavities than at exposed and flexible residues.

Pure water simulations as well as water molecules observed by X-ray crystallography have often been investigated with the attempt to identify special properties of water e.g at the binding site of target proteins (Rodier et al., 2005, Amadasi et al., 2006, Barillari et al., 2007, Samsonov et al., 2008). Hydration mapping is a widely used computational approach to identify spots and regions of high and low water concentration around protein surfaces from molecular dynamics simulations (Henchman and McCammon, 2002a,b, Virtanen et al., 2010) and diverse experimental methods (Svergun et al., 1998, Ebbinghaus et al., 2007, Zhang et al., 2009a). Recently published studies report on the thermodynamic properties of water around pharmacological relevant proteins and the attempt to connect the hydration of the protein with its functional sites (Beuming et al., 2011), with limited but notable success. Seco et al. (2009) also used solvents to identify small ligand binding sites and determine

their druggability. They mimicked experimental studies with proteins saturated in mixed solvents, composed of varying volumes of alcohol molecules in the solvent, for a small number of proteins. In their report, coherence between high affinity binding spots of the solvent with the known drug target site could be shown in some cases. It was emphasized, however, that this procedure was subject to several limitations concerning the selection of hot spots, as well as the unambiguity of the binding site prediction.

Former approaches focused either only on one aspect (Beuming et al. (2011) concentrating mostly on low hydration sites of water to identify small ligand binding sites) or on a very limited set of proteins (Seco et al., 2009). In this thesis (see chapter 3), the interplay between solute and solvent was studied for a pure water solvent as well as for a mixed solvent system, using molecular dynamics simulations. In contrast to previous studies, which investigated small ligand binding sites, large protein-protein binding sites are focused in this thesis. The presence and absence of the solvent molecules within functional sites of the proteins was investigated and used to predict possible binding sites. Previous studies often focused on a certain type of protein or protein families, while in this work, a wide range of protein interfaces are investigated to identify general trends. Additionally, 15 proteins were analyzed for which is known, that the protein-protein binding can be inhibited by small ligands, with some of the inhibitors binding at the protein-protein binding site.

Distinct criteria have been observed for small ligand binding sites. As reported in chapter 2.1 small ligands are often found in one of the largest cavities and therefore, the search for these ligand binding sites can be cut down to the detection of cavities. Many geometry-based methods were developed (summarized in Leis et al. (2010)) which focus on the detection of cavities, but only a few methods came up which analyze energetic properties (Laurie and Jackson, 2005, Brylinski et al., 2007). While geometry-based methods only focus on the size or deepness of a cavity, energy-based methods try to identify the key interaction between the probes and the protein's surface. It has been found, that many ligands bind in hydrophobic cavities with a low amount of clearly defined polar contacts, which is not included in geometric cavity detection, but is covered at least partially by energetic methods.

A new method proposed in this thesis (see chapter 4) combines a fast geometry-based approach to identify small ligand binding site cavities with calculations of desolvation free energies using an implicit solvent model. This

enables the detection of the main interactions of ligands with proteins, the hydrophobic as well as the polar contacts. Recently developed cavity detection algorithms try to reduce the size of the predicted binding site by applying additional geometrical criteria (Weisel et al., 2009, Yu et al., 2010, Volkamer et al., 2010). The procedure presented here does not only include the shape of the cavity to identify binding regions but uses the detection of distinct energetic profiles to overcome the limitations of pure geometry-based pocket area reduction and sub-pocket prediction.

Adjacent to the identification of binding sites, the prediction of the complex formed between a protein and its partner are of fundamental interest for the understanding of molecular recognition. As discussed in chapter 2.3 there are many approaches in existence, which can be applied to identify the complex of two proteins. Fast approaches based on geometrical correlation follow the assumption of the lock-and-key mechanism and try to identify optimal sterical complementarity (e.g. FFT methods, for example see Katchalski-Katzir et al. (1992)). Other methods use force fields to mimic the driving forces of protein-protein interaction, providing less approximated scoring of docked complexes but come with an increase of computational costs (e.g. Zacharias (2003)).

During the docking procedure, thousands and hundreds of thousands of complexes are docked and the selection of the near native complex among all results, invokes an additional level of complexity. Therefore, knowledge on a protein's binding site or its prediction are often included to re-score the docking results afterwards (Ben-Zeev and Eisenstein, 2003, Huang and Schroeder, 2008) or are directly included as restraints during the docking procedure (Dominguez et al., 2003). Both attempts have drawbacks: the re-scoring procedure does not generate any new complexes, besides the ones predicted without any information on the binding site, and the restraining procedure does not allow any results apart from the restraint regions. To overcome these restrictions, in this thesis information from binding site predictors are used to control the scores of a force field based docking procedure (see chapter 5). This enables increasing the sampling and scoring of complexes in contact at proposed surface regions, without reducing the sampling at sites of the protein's surface far away from the predicted regions. Additionally, the effect of inclusion of experimental information into complex prediction was analyzed using optimal parameters (compare chapter 5) and for an important complex of the complement system with unknown three dimensional structure (see chapter 6).

2.5.1 Acknowledgment

Parts of this chapter have been published in Schneider and Zacharias (2011).

Chapter 3

Solvation of proteins in explicit solvent simulations

Interaction between proteins and other biomolecules is based on a variety of effects including hydrophobic collapse, electrostatic interaction, solvation effects, hydrogen and salt bridges or cofactor binding (Xu et al., 1997, Tsai et al., 1997, Sheinerman and Honig, 2002, Rodier et al., 2005, Chandler, 2005). This occurs in an aqueous solvent within a compartment of the cell or in the cytosol, and some of this interaction is directly or indirectly related to the solvent. The hydrophobic effect (Tanford, 1978, 1979) is based on the presence of water and is one of the major contributors to the binding of obligatory protein complexes. It was found, that most enzymes are only able to fulfill their functionality in the presence of water (Monsan and Combes, 1984, Smolin et al., 2005) and also that protein association is mediated and stabilized by water molecules (Bhat et al., 1994).

Upon binding of a ligand, water has to be replaced from the surface of the binding partners, contributing to the energy barrier to be overcome during such processes (Beuming et al., 2011). Then again, spots of water can reduce the influence of unfavorable residues in the binding site of a complex (Samsonov et al., 2008). Small ligands, binding within a deep cavity, would have to displace water in the interior of a cavity to the bulk (Michel et al., 2009). Within polar regions, this includes the breaking of hydrogen bonds between the protein's surface and the water molecules, within hydrophobic regions, water is released to the bulk, allowing it to form hydrogen bonds with other water molecules or hydrophilic residues.

Several different approaches to analyze the presence or absence of water or other solvents in the binding sites of proteins are available. One kind of method analyzes waters found in experimental studies, e.g. X-Ray crystallography. Different concentrations of bound water molecules were found within binding sites of small ligands (Barillari et al., 2007, 2011) as well as a high amount of water in specific protein-protein binding sites (Rodier et al., 2005, Reichmann et al., 2008). Another method uses molecular dynamics simulations, or other computational approaches, to sample the behavior of water around the protein's surface. Several approaches exist to identify hydration sites around the protein and correlate these regions to binding sites, either aiming at the identification of permanently bound waters (Merzel and Smith, 2005) or regions with low solvation (Beuming et al., 2011). Besides pure water solvents, also mixed solvents have been widely used to investigate the behavior of proteins. Solvent analysis were used to identify binding sites of proteins using experimentally determined structures (Mattos and Ringe, 2001) or computer simulations (Seco et al., 2009).

In this chapter, the solvation of proteins will be discussed using several distinct methods to analyze the interacting behavior of solute and solvent. In previous studies, only small ligand binding sites, a small amount of proteins or proteins of one family have been the target of investigations. Therefore, the study presented in this chapter focuses on a set of protein-protein complexes of different types, binding site size and binding site residue composition to investigate general trends of solvation. Also protein-protein binding inhibition sites are investigated in a second test set, including inhibitors binding at the protein-protein binding site.

Pure water simulations were performed for the former test set to study the behavior of water at the protein-protein binding site, with a main focus on low hydration areas. Also, the protein-protein binding sites as well as the inhibitor binding sites were investigated using mixed solvent simulations. Mixed solvent simulations have the advantage, that regions on the protein's surface with low or medium affinity for water binding can be populated by the other solvent ingredients. This allows accounting for the different aspects of hydration, which is the main focus of this study: the low hydration found in single protein complexes as well as wet spots filled with water in the interior of a binding site. Analysis of the solvation of the unbound partner enables investigation into whether this low or high solvation information can already

be found before complex formation and if this information can be used to directly predict binding sites.

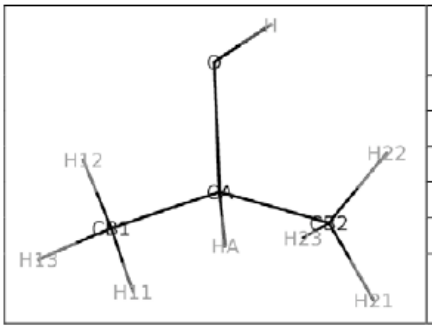
3.1 Molecular dynamics simulations of proteins: setting and application

Molecular dynamics simulations were performed to sample solvent molecules on the protein's surface. Therefore, a long simulation of nearly 25 ns per protein structure was performed. The first 15 ns of the simulation were used to equilibrate the system, followed by the sampling of solvent distribution in equilibrium state, within the last 10 ns of the trajectory. Solvent molecules will be counted afterwards using a grid based approach to identify regions of high and low solvation near the protein's surface.

3.1.1 Simulations in aqueous solvents

Water is the most dominant solvent enclosing biomolecules within a cell. Solvents are often investigated to identify either interactions of individual water molecules or the global behaviour within protein-protein and protein-small ligand binding sites (Steinbach and Brooks, 1993, Timasheff, 2002, Barillari et al., 2007, Amadasi et al., 2008, Beuming et al., 2011). Water was found to mediate the binding of biomolecules and is in many cases essential for ligand binding and drug design (Oubridge et al., 1994, Bhat et al., 1994, Ladbury, 1996). Replacing water upon ligand binding, especially from deep cavities, is connected to an energy barrier which has to be overcome (Barillari et al., 2007, Michel et al., 2009). Since small ligands can often be found in more hydrophobic cavities (Miller and Dill, 1997, Liang et al., 1998), such binding sites were recently studied with respect to low hydration sites (Beuming et al., 2011).

Also, mixed solvents have widely been used to analyze biomolecules in experiments (Timasheff and Inoue, 1968, Wüthrich et al., 1992) as well as in computer simulations (Schellman, 1990, 2003, Shulgin and Ruckenstein, 2005). Since in nature proteins do not exist in pure water but in a mixture of diverse molecules, mixed solvents allow the investigation of protein surface features (Seco et al., 2009, Dechene et al., 2009), co-solvent contribution to binding



Atom Names	Amber Atom Type	Partial charge
O	OH	-0.612
H	HO	0.373
CA	CT	0.309
HA	H1	-0.012
CB1, CB2	CT	-0.176
H11, H12, H13 H21, H22, H23	HC	0.049

Figure 3.1 – Figure taken from (Seco et al., 2009) Supporting Information. Isopropyl alcohol and its partial charges for the amber forcefield.

(Shukla et al., 2009), protein functionality (Buhrman et al., 2003) or general interaction networks (Vagenende et al., 2009).

Isopropyl alcohol (iPrOH) is a small alcohol often used to investigate protein features in experiments (Mattos and Ringe, 1996, 2001), e.g. small ligand binding sites on different protein surfaces (English et al., 2001, Mattos et al., 2006, Ho et al., 2006). Also, isopropyl alcohol was used in computer simulations to predict the binding sites of small molecules (Dennis et al., 2002, Seco et al., 2009). The work of Seco et al. (2009) revealed that for one structure iPrOH bound exactly at the protein-protein binding site of protein phosphatase (PTP-1B) and insulin receptor tyrosine kinase (IRK) (compare Seco et al. (2009)). Therefore, it can be assumed that the high affinity of the mixed solvent in cavities containing small ligands also can be found in protein-protein binding sites. On one hand, the accumulation of the iPrOH molecules at the binding site can be used to indirectly account for surface regions which are unfavorable for water molecules. On the other hand, the interaction between iPrOH and the protein can be assumed as a simplified contact to a possible partner protein.

Isopropyl alcohol is soluble in water and offers two different chemical features (Figure 3.1): the hydroxyl group (OH-group) in the second position can be involved in hydrogen bonding and the propyl group (Me-group) avoids polar regions on the protein's surface and therefore, can often be found at hydrophobic surface areas. Both sites of the isopropyl alcohol represent chemical features, often found in organic ligands as well as several amino acids.

In the simulations studied in this chapter, pure water solvents as well as mixed solvents are used to investigate the solvation of protein-protein binding sites. In contrast to most approaches so far, which focused on either only small

ligand binding sites or a distinct type of protein, here, a discriminable set of proteins is investigated to analyze the global effect of solvation within binding sites.

3.1.2 Molecular dynamics simulation protocol

Simulations were run following roughly the protocol of Seco et al. (2009). The Amber package was used to perform the simulations using pmemd (Cornell et al., 1995). The solution was composed of a 20 % v/v mixture of Isopropyl alcohol (iPrOH) molecules and TIP3P water molecules for the mixed solvent simulation and 100% TIP3P water for the water simulations. Proteins were simulated in a 10 Å octahedral box and counter ions added (either Na+ or Cl-). Minimization was performed for the protein and for the solvent before heating the system in 50 K steps from 100 K to 600 K and back to 300 K. Typically, these steps lasted 10 ps with the exception of the 600 K step (600 ps) and the last step at 300 K (800 ps). Equilibration was performed at 300 K for 1 ns followed by two productive runs with each 10 ns at 300 K. Restraints were used during heating on C_α atoms (25.0 $kcal\ mol^{-1}$Å^{-2}) and gradually released during the 800 ps step to 0.0 for the equilibration. During the productive run restraints of 0.25 $kcal\ mol^{-1}$Å^{-2} on all heavy atoms were used to keep the protein and the side-chains in shape to allow proper sampling along the entire surface.

Restraining atoms during the simulation, after heating to high temperatures and cooling down to 300 K with subsequent, fully free, 1 ns simulation of the protein, might result in an unfavored conformation. Therefore, it must be emphasized, that results might differ slightly in general, but significantly in an opened or closed cavity, depending on the structures gained from the free simulation. Allowing full flexibility during the sampling procedure would result in less meaningful grid calculations and even side-chain flexibility can reduce the magnitude of solvent sampling with a grid based approach near the residue's surface. Full flexibility is on one hand prerequisite to allow opening of pockets, which is necessary to identify small ligand binding sites, but on the other hand it reduces the sampling of solvation sites at exposed regions while it accentuates the population in cavities (for restraining protocols for hydration site mapping compare e.g. Beuming et al. (2011) and therein). It can be a general drawback of all grid based approaches that, for exposed and flexible

residues, the magnitude observed might be lower than for enclosed residues with less fluctuation, as found in cavities, maybe even if a comparable amount of iPrOH is in contact with those residues.

3.1.3 Test sets

Solvents do not behave in the same way all over the protein's surface (Makarov et al., 1998b, Rodier et al., 2005). Hydrophobic regions as well as polar contacts have a significant influence on the behavior of the solvent within the first solvation shell (Pizzitutti et al., 2007, Beuming et al., 2011). The specific features of cavities, found to bind small ligands, differ from those of exposed regions. Therefore, different test sets have been investigated. A mixed set of ten protein complexes from a published benchmark set (Mintseris et al., 2005) was used to track the behavior of water and iPrOH at protein-protein binding sites. These proteins were chosen, because of their highly diverse contact areas, difference in size, shape and residue composition and are members of different protein families (see Table 3.1). The unbound protein partners were used for the simulations, as prepared in the Benchmark 2.0 test set (Mintseris et al., 2005). This set is called protein-protein binding test set during the rest of the chapter.

Additionally, 15 proteins have been investigated with iPrOH simulations, which can form a complex with other proteins, but can be inhibited by small (organic) molecules (called protein inhibition test set). This test set was part of other studies for small ligand binding site detection and druggability studies (Kozakov et al. (2011) and within, pdb codes: 1m47, 1r2d, 1z1m, 1r6k, 1f46, 1tnf, 1f9x, 1iu2, 1i6c, 2o8t, 1cqr, 1e31, 1bi4, 1aly, 1kd7, if several models or chains had been within the files, models and chains were chosen according to Kozakov et al. (2011)). For three of the 15 molecules, no bound structure is available and they will therefore be excluded from most of the protein-inhibitor binding site statistics (1e31, 1aly, 1kd7) but included in the general analysis. For binding site statistics, only those sites were taken into account for which the partner was given by Kozakov et al. (2011). Alternative small ligand binding partners, as well as the protein partners, will be discussed separately.

Table 3.1 – Test set containing unbound structures from Benchmark 2.0

pdb code	partner	category	protein name	residues
1ACB	receptor	enzyme	Chymotrypsin	184
	ligand	inhibitor	Eglin C	62
1AY7	receptor	enzyme	Barnase	91
	ligand	inhibitor	Barstar	75
1BUH	receptor	other	CDK2 kinase	294
	ligand	other	Ckshs1	71
1AKJ	receptor	other	MHC Class 1 HLA-A2	375
	ligand	other	T-cell CD8 coreceptor	193
1D6R	receptor	enzyme	Bovine trypsin	179
	ligand	inhibitor	Bowman-Birk inhibitor	58
1KAC	receptor	other	Adenovirus fiber knob protein	158
	ligand	other	Adenovirus receptor	105
1KTZ	receptor	other	TGF-beta	103
	ligand	other	TGF-beta receptor	91
1TMQ	receptor	enzyme	alpha-amylase	470
	ligand	inhibitor	RAGI inhibitor	102
2JEL	receptor	bound antibody	Fab Jel42	435
	ligand	antigen	HPr	67
1IQD	receptor	bound antibody	Fab	334
	ligand	antigen	Factor VIII domain C2	132

3.2 Solvent profiles on the proteins' surfaces

Solvent molecules at the protein's surface behave differently from those in the bulk region (Pal and Zewail, 2004). Hydrogen bonds play a crucial role in the interaction between solvent molecules and the surface of a biomolecule and are involved in many processes (Park and Saven, 2005, Raschke, 2006). The number of hydrogen bonds a solvent molecule can establish with the surface depends on the physicochemical composition of the surface as well as the shape (Lee and Rossky, 1994, Scatena et al., 2001). Depending on the surface, hydrogen bonded water networks can be established that influence the activity and dynamics of the protein (Smolin et al., 2005).

To investigate the behavior of water and iPrOH at the proteins' surfaces, a grid based approach was used to count the occurrence of central parts of the molecules at certain positions in the simulation box. From these values, density profiles for regions around surface residues could be extracted, giving insight into most favored binding contacts of water and, in the case of iPrOH, of the Me-groups as well as the OH-group. Additionally, proximal radial distribution

functions were calculated to identify the position and amplitude of solvation shells. This allows the determination of the proximal average position of the solvent molecules on the solutes surface within the first solvation shell.

3.2.1 Measurement of the radial solvent distribution around proteins

In experimental studies, a solvation shell with higher density than the bulk concentration is found within close distance to the protein's surface (amongst others, Svergun et al. (1998)). Such solvation shells, or the probability of finding a molecule **A** at a certain distance to molecule **B**, is often measured using radial distribution functions (RDF) (Madan and Sharp, 1999, Henchman and McCammon, 2002b, Schmidtke et al., 2011b). This function can provide useful information on the number of observable shells and the intensity of the peaks, as well as the distance between those peaks, but are uncertain if the shape of **B** is less spherical (Brooks et al., 1990). This is especially the case for larger macromolecules like proteins, so that proximal radial distribution functions (pRDF) are used instead (Makarov et al., 1998b, Jha et al., 2005, Lin and Pettitt, 2011). This method identifies the density around a molecule as a function of the minimal distance of the solvent to the solute's atoms. Often, the solute's atoms are separated into atom groups (see Lin and Pettitt (2011) and within) or pruned to backbone atoms (Schmidtke et al., 2011b).

In this work, shells are defined in 0.2 Å steps from the protein's van der Waals surface and solvent molecules located within each shell are counted. To define a shell's free accessible volume, a grid is used with a 0.1 Å step size. Using

$$d(a,r) = \begin{cases} 1, & \text{if } r \leq dist(a, protein) < r + b \\ 0, & \text{else} \end{cases} \tag{3.1}$$

$$f_{shell}(r) = \frac{\sum_{i=0}^{N} d(m_i, r)}{\sum_{i=0}^{G} d(g_i, r) V_{bulk} T} \tag{3.2}$$

$f_{shell}(r)$ returns the relative probability to find a solvent molecule within a

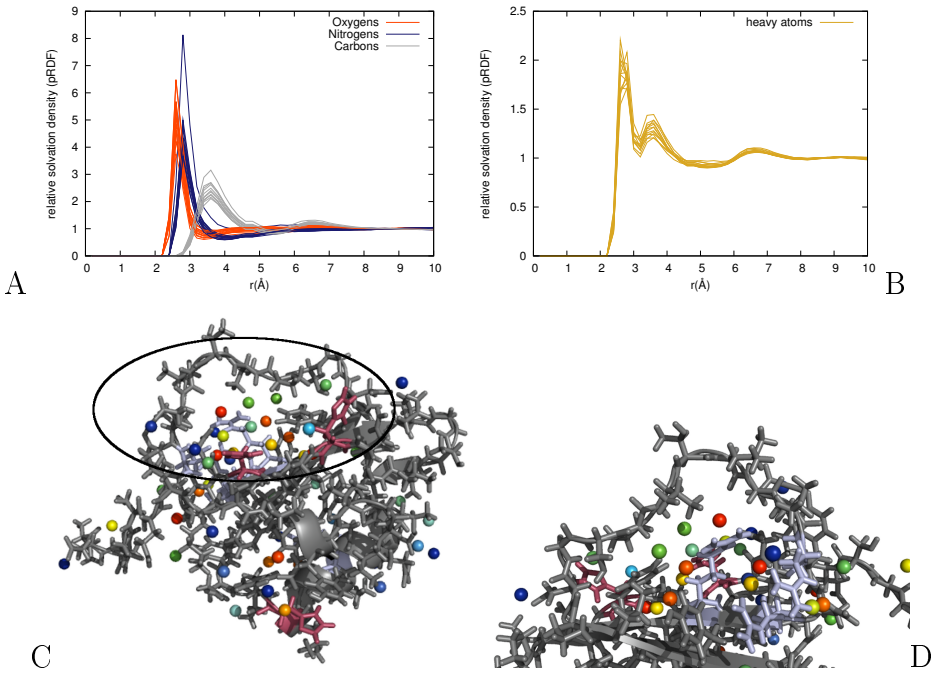


Figure 3.2 – Proximal radial distribution function of the proteins in simulations with pure water solvent. A) Single atom groups are shown separately. B) pRDF for all heavy atoms is shown. The peaks for the entire heavy atoms can be explained from the pRDFs of the atom groups: the first peak corresponds to the contacts made with oxygen and nitrogen rich amino acids and the second peak results from contacts with carbon atoms. C) and D) pdb1acb with the hydration sites shown as spheres (color coded with blue hydration of 4 times the bulk as the lowest and red as the highest). Arginine residues are shown in light blue and histidine residues in red.

given shell at distance r to the protein atoms, with b the bin size, N the total number of solvent molecules, G the number of grid points, T the total number of time steps and V_{bulk} the expected bulk value. The latter is done to normalize the value towards a relative probability of 1 in the bulk region. The distance between a solvent and the solute is defined as the minimal distance between the solvent coordinates and the protein atom coordinates in each time step. Depending on the type of atom observed (three atom types represent the different chemical groups: oxygen, nitrogen and carbon atoms (Makarov et al., 1998a)), only distances to atoms of the corresponding atom type are used to determine the relative distribution around the atom groups.

In general, pRDFs show the same position of the peak, with little divergence, for each protein, while the height of the peak can change significantly for the different atom groups (see Figure 3.2). The latter depends on the composition of the amino acids at the protein's surface as well as on the shape of the surface. Lin and Pettitt (2011) showed on nine different proteins and peptides the universality of pRDFs, surrounded by TIP3P water molecules, in a 10Å box,

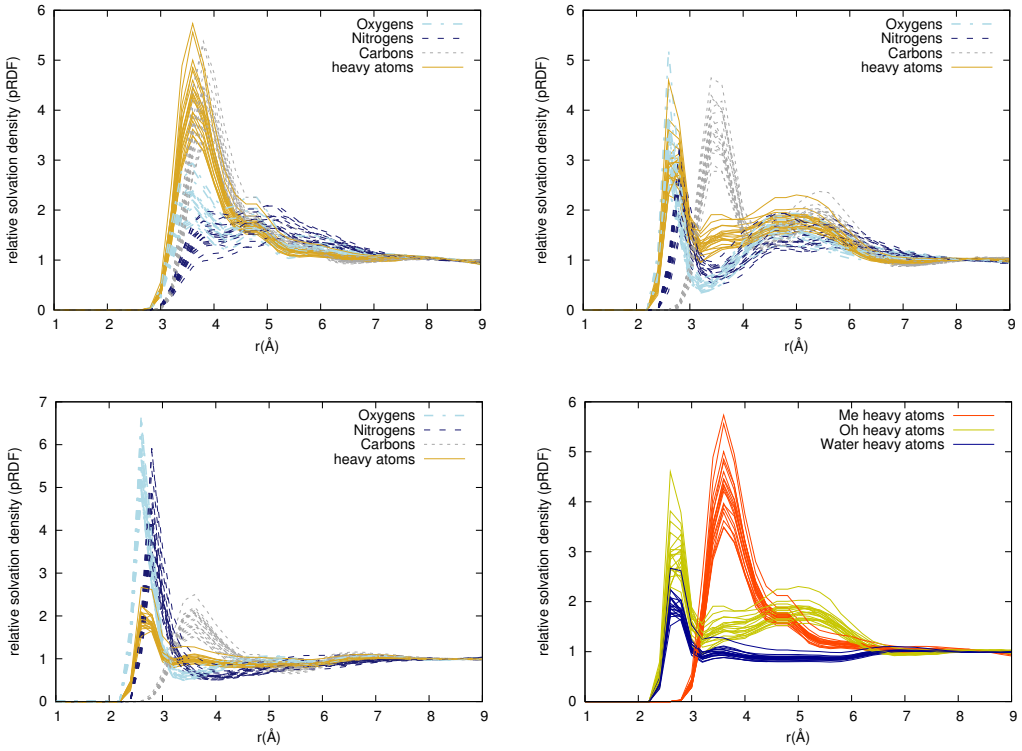


Figure 3.3 – pRDFs of the different solvents are shown. Upper left shows the density distribution of the Me-group of iPrOH at the surface, divided by atom contact groups. Upper right shows the same for the OH-group and lower left for water. Lower right pRDF shows the relative density of iPrOH and water relative to all heavy atoms.

using snapshots from 2 ns of simulation time. Figure 3.2A shows for one pRDF a significant difference in the height of the peak for nitrogen contacts. The high peak is mostly influenced by hydration sites in a cavity formed by three arginine residues (Figure 3.2D). A loop involved into enzyme inhibition (Figure 3.2C, black circle) creates a large cavity, surrounded by two histidine residues, the backbone of the loop and the arginine residues, in which proximity a high amount of hydration sites can be found.

3.2.2 Orientation and density distribution of isopropyl alcohol molecules at the surface

The orientation at the protein's surface can be manifold for mixed solvents. In general, three different classes of poses for Isopropyl alcohol on the protein's surface exist (examples can be found in Figure 3.4):

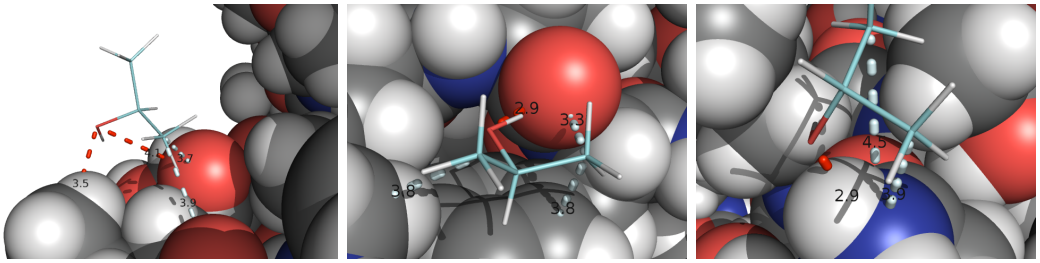


Figure 3.4 – Three examples for different binding poses of iPrOH. Protein surface is shown in spheres and iPrOH in line representation. Red spheres show oxygen, gray carbon and blue nitrogen atoms of the protein. Distances are given from the center coordinates of the atoms.

1. The OH-group can be found at the surface of the protein while both Me-groups point away from the surface. The distance between the center of the OH-group and the carbon of the Me-group is ∼2.3 Å. Dependent on the angle between the molecule and the surface, the distance of the Me-group to the (ideally planar) surface can vary. In the most extended conformation, both Me-group centers are 4.3 Å away from the protein's vdw-surface.

2. Either one Me-group or both can be found at the protein's surface with a distance of 2.5 Å between the centers of the Me-groups. In the former case, the second Me-group and the OH-group are up to 4.8 Å away from the vdw-surface, while in the latter case only the OH-group is far away from the surface and can often be found at carbon rich areas.

3. The OH as well as one or both Me-groups are at the protein's surface. This conformation is observed often, if the iPrOH molecule is stacked between residues. Often, one Me-group is not in contact with the surface, but exposed to the solvent, and can be found more than 4 Å away from the protein's vdw-surface.

For the Me-group of iPrOH, the density with contact to carbon atoms is distinct as compared to contacts with oxygens and nitrogens (compare Figure 3.3). Contacts to oxygens are represented by two small peaks, the first at the direct contact distance and the second determined by the binding of the two other groups of iPrOH to the surface. The contact to nitrogen atoms is defined by the contact of the OH-group to the nitrogen, depending on the surrounding atoms (see Figure 3.4 on the right). The OH-group preferably contacts the oxygen and nitrogen atoms on the protein's surface, which is shown in the

density profiles in Figure 3.3. A second broad peak can be identified for OH-oxygen and OH-nitrogen contacts, which can be related to iPrOH bound to the surface via the Me-groups. This is also the reason why a large peak of the OH-group can be found around carbon atoms. The varying distribution of iPrOH beyond the first peaks can also be related, up to a certain degree, to isopropyl molecules building clusters at the protein's surface, resulting in layers of iPrOH that are not directly in contact with the protein (see also Seco et al. (2009)).

The contacts for each specific atom group are represented in the diagram including all heavy atoms (see Figure 3.3). For the Me-group, the contacts with carbon rich surface areas have the largest impact, resulting in a relative density three to five times the bulk concentration. Oxygen and nitrogen contacts are essential for the relative distribution of the OH-group, while the carbon as well as the second peak of oxygen and nitrogen contacts rely on the binding of the Me-group. Water molecules in the mixed solvent simulation are slightly lower represented beyond the first peak, due to the solvation of the solute with iPrOH molecules. This results in a decrease of relative concentration lower than the bulk, within 3 to 6 Å from the protein's surface.

3.2.3 Solvent-solute contact time

The shape of the protein's surface as well as the residue composition have an influence on the time a solvent stays in contact with the solute. In deep and polar cavities, a water molecule can establish the maximum number of hydrogen bonds with the protein. Water molecules in such cavities can be found strongly bound, and therefore, replacement of such a solvent molecule requires the breaking of hydrogen bonds and replacement of several other solvent molecules and takes more time, than the replacement of water at exposed and hydrophobic surface areas. The protein-protein binding sites in the test set used in this study are often rather flat, so that water molecules in the binding site are able to exchange with water molecules at the surrounding surface area, as well as with the bulk water molecules. Only in single cases (pdb1ay7 the protein-protein binding site, pdb1acb the enzymatic region, pdb1buh the ATP binding site), solvent molecules are strongly bound within a small cavity, but in general the time in which a single molecule stays within the binding region during the simulation is rather small. If a solvent molecule is bound

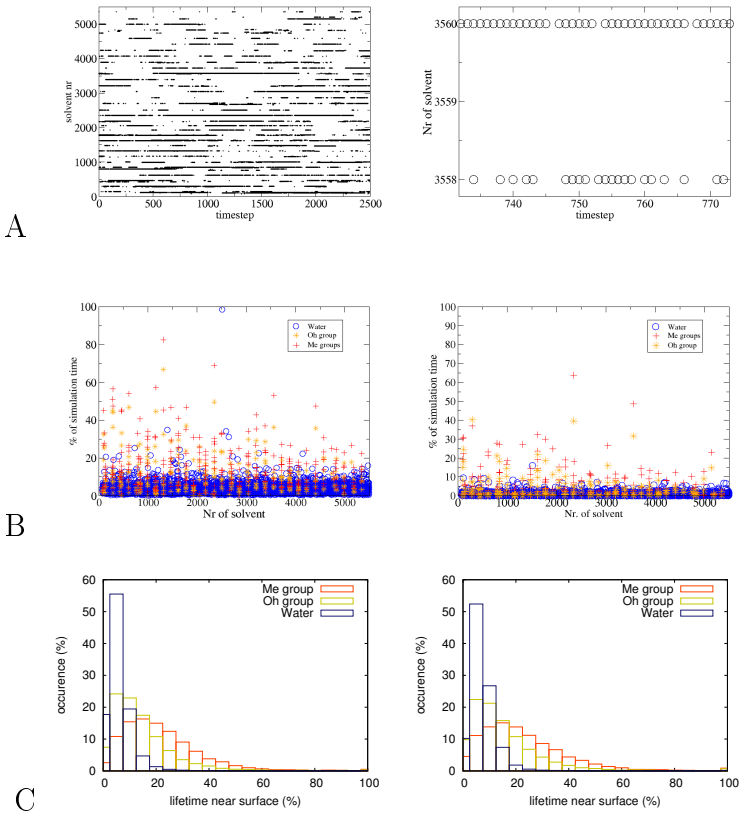


Figure 3.5 – For the ligand of protein pdb1buh: A) shows the occurrence of the Me group of iPrOH on the protein's surface at a certain time step. Distance cut-off was 5.0 Å to be counted to be at the surface. Me-iPrOH was counted at the surface if one of the two groups fulfill the criterion. A) on the right shows a close up of two Me-groups. B) Lifetime of solvent molecules near the surface during the simulation for the entire surface (left) and the binding site (right). C) Distribution of the three solvent types at the surface of the protein-protein (left) and protein inhibition (right) test set (using 5% bins).

in a small cavity, as they are found in the second test set, the time of single solvent molecules within the first solvation shell can be higher than on flat or exposed areas. However, only a few percent of the solvent molecules are in contact to the surface longer than half of the simulation time (see Figure 3.5). This shows that in general, the exchange between surface and bulk is given and only single solvent molecules in internal cavities are trapped during the entire simulation.

Figure 3.5A shows that some molecules, which are in contact with the surface for a long period of time, can lose contact with the surface, either fluctuating between surface and bulk or leaving the surface region and enter the bulk (for cut-off criteria compare legend of Figure 3.5). Fluctuation of one chemical group of iPrOH, for example the OH-group, often does not depend

on the binding of the group to the surface itself, but depends on the opposing Me-group, which give a major contribution to the binding of the molecule to the surface. On average, only 0.1% of the iPrOH molecules within a simulation have never been at the surface and 26% were never in the binding site (19% in the protein-protein binding sites and 37% in the small ligand binding sites). On one hand, the probability of not being at the binding site at least once increases within large boxes, especially for the protein inhibition test sets, including small ligand binding sites. On the other hand, within smaller boxes around more spherical proteins, all solvent molecules are in contact with the protein's surface at least once and only in four cases (out of 35, receptor of pdb1ktz and pdb1iqd as well as pdb1cqr, pdb1f9x) a solvent molecule was not in contact with the surface.

3.2.4 Grid based sampling principles and sampling issues

The identification of local solvent accumulations is done via a rigid grid approach. Therefore, a grid is constructed around the protein's surface using a grid spacing of 0.5 Å. An average structure of the protein, sampled during simulation, was used as a reference protein state for grid generation and distance calculations. Since restraints have been applied, the general geometry of the protein was preserved. Grid points at a distance of more than 4.0 Å from the protein's (vdw-)surface were deleted to reduce the amount of computational effort. This value covers the first solvation shell of water as well as iPrOH (see Chapter 3.2.1). 2500 frames were taken from the second 10 ns of the simulation, and the occurrence of the solvents on each grid point counted. The O1 of the isopropyl was counted for the OH group, either the C1 or the C2 counted for the Me-Group of the isopropyl and the central O for the position of water molecules. Those concentrations were mapped on the grid points and normalized by the expected concentration (taken from an only solvent simulation following the exact protocol). Then, as the first step, grid points were averaged over the direct neighbors to generate smoother patches and reduce grid artefacts. In the second step, beginning with the grid point with the highest concentration, all neighbors within 1.4 Å (the approximated radius of a water molecule) were deleted and the procedure repeated for the next grid points, according to their concentration, until none is left to be deleted. This is done for water, OH-group of iPrOH and Me-group of iPrOH separately, resulting

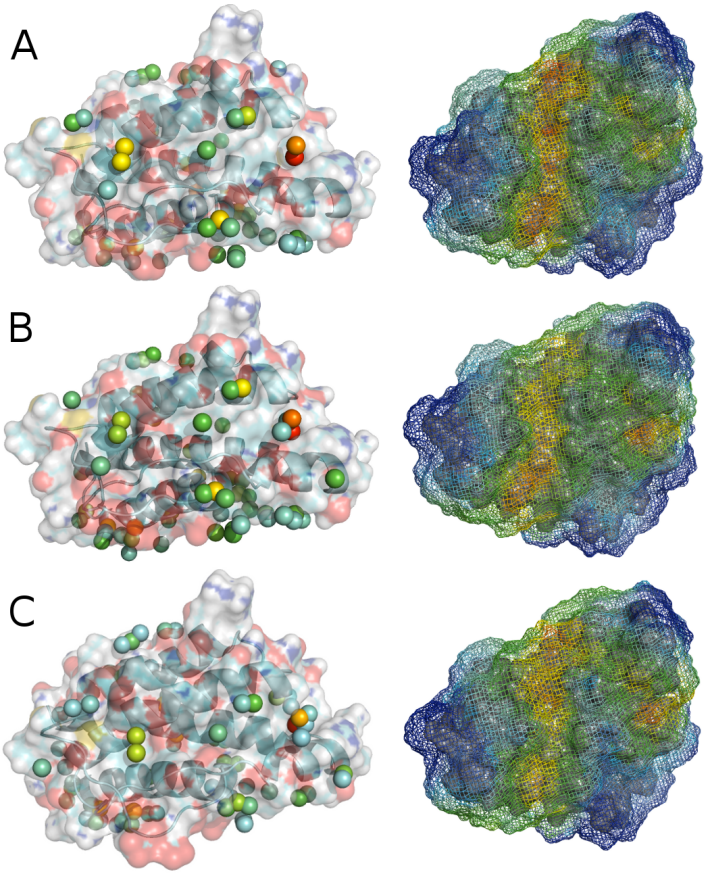


Figure 3.6 – For pdb1m47, two independent simulations have been performed. On the left, high affinity binding spots (spheres) are shown for the Me groups of iPrOH while on the right values averaged of neighbors within 6 Å are shown (mesh). A) second 10 ns of the first simulation. B) third 10 ns of the first simulation. C) second 10ns of the second simulation. Coloring is relative to the highest and lowest values found. These values might differ, especially in cavities surrounded by flexible sidechains.

in one (in case of water only simulations) to three (in case of water/iPrOH simulations) grids, each presenting the local ratio of sampled value to bulk value on each grid point.

Seco et al. (2009) reported convergence of the simulation after 1.5 ns of equilibration, for the proteins used in their study. Besides the fact that, in at least one case, differences between the first eight of the 16 ns productive run and the second eight nano seconds could be found, all changes of iPrOH population during the simulation could be related to conformational changes. The modified procedure presented in this chapter implies a 2.2 ns free equilibration, followed by additional 10 ns restrained equilibration at 300 K before sampling. The first 5 ns from the first 10 ns simulation differ from the second, while the second 5 ns of the first 10 ns simulation only modestly differ from the second 10 ns simulation.

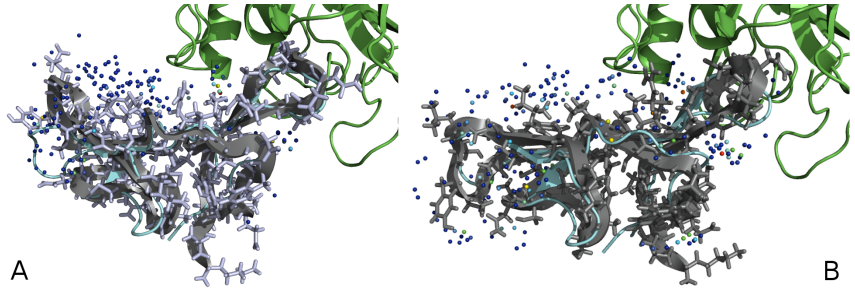


Figure 3.7 – A) and B) show different conformations of the ligand of complex pdb1d6r from independent simulations. The protein complex is shown in bound form for illustration purpose with the receptor shown in green and the ligand in mint, respectively. A representative of the average structure of the second 10 ns run (lowest RMSD between average structure and one structure of the trajectory) is shown in each case as extended cartoon with sticks. Small spheres show binding spots of Me-iPrOH with 8 fold the expected bulk value or more.

For some cases (pdb1m47, Figure 3.6 and the ligand of pdb1d6r, Figure 3.7), additional 10 ns were performed as well as a completely independent simulation, showing, that only slight differences occur and the simulations are converged within the second 10 ns. If the protein undergoes larger deformations during simulation, the profile on the surface can change, due to formation and deformation of cavities where water and iPrOH bind. In the case of the ligand of pdb1d6r, the difference in backbone RMSD between two representatives of independent runs is 2.8 Å and results in a significant difference in solvation in opened and closed cavities (compare Figure 3.7).

Larger structural changes, especially at exposed loop or terminal regions, may influence the binding of water and iPrOH. In some cases (e.g. in Figure 3.7A and B) the contact residues stay the same, while the global or local position of the side-chains changes. In many cases, if a formed cavity is narrowed or even closed upon side-chain or backbone movement, less iPrOH molecules are able to bind in such cavities, resulting in a divergent solvent distribution at this local spot in independent simulations.

3.2.5 Gibbs free energy approximation from rigid grid calculations

Values obtained by the grid calculations can be directly transfered into ΔG values, so that for each grid point

$$\Delta G = -RTln(\frac{N_i}{N_0}) \tag{3.3}$$

is calculated, with N_0 the bulk value extracted from only solvent simulations and N_i the value measured on grid point i. Seco et al. (2009) used this method to calculate binding affinities of clusters of iPrOH in small ligand binding sites and to propose the druggability of these sites. Seco et al. (2009) truncated values above 12.5 times the bulk value to 12.5 (-1.5 kcal/mol) and neglected values higher than ΔG of -0.84 kcal/mol. The truncation was made according to the maximal possible contribution found for ligand atoms (Kuntz et al., 1999). Here, values above 12.5 times the bulk are truncated, but higher values than -0.84 kcal/mol accepted. One has to emphasize, that a certain exchange of isopropyl alcohol in the binding site with the bulk must take place, to approximate ΔG values reasonably well. Figure 3.5 shows, that a high exchange is given, as reported for other test sets (Seco et al., 2009). Nevertheless, calculating ΔG values for large regions on the protein's surface is less straight forward than for small clusters within small binding sites. Therefore, ΔG values can hardly be used to estimate the binding affinity of protein-protein binding sites. However, the distribution of ΔG values can be compared between different regions on a protein's surface or between independent simulations.

Figure 3.8 shows the difference of ΔG distribution of the solvents near the surface (Figure 3.8A, 4.0 Å to the vdw surface of the protein) and within a larger box (Figure 3.8B, cut-off 10 Å). Values around zero mark spots where the concentration of the solvent was equal to the bulk concentration. If considering only the near surface area, then the occurrence of these values decreases, but does not disappear. The latter is also true for values lower than the bulk (resulting in positive ΔG), showing that local spots on the protein's surface are avoided by the solvent or are only transiently populated. Using ΔG values, it is possible to identify regions on a specific protein with high or low affinity to the solvent.

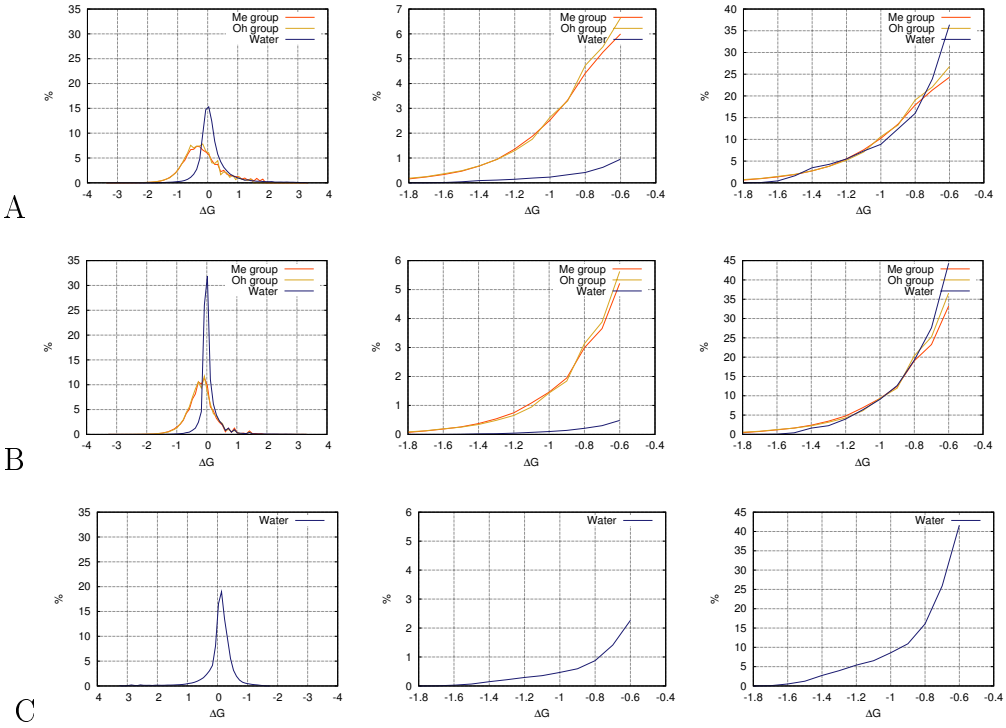


Figure 3.8 – Distribution of ΔG values of the solvent within the truncated grid (A) and the complete grid (B) for all 35 proteins. The first column shows the entire distribution of the solvent's ΔG values up to a cut-off of 10 Å. Column two shows a close up of the distribution of the attractive energies (from -1.8 kcal/mol to -0.6 kcal/mol) and column three the same but normalized towards the amount of solvent found within this regime. C) Distribution of ΔG values of water in the unary simulation for the 4 Å grid for the 20 protein-protein test set.

The relative solvation density derived from proximal radial distribution functions is also reflected in the distribution of ΔG values around the different atom groups (Figure 3.3 and 3.9). The Me-group prefers binding in carbon as well as oxygen rich areas. The affinity to nitrogen atoms is only slightly shifted towards higher affinity. Nevertheless, contacts to favored carbon atoms do not necessarily have to be high affinity contacts. Oxygen contacts are largely favored, in terms of affinity, over carbon contacts by the OH-group, while contacts to nitrogen atoms are balanced. Water avoids carbon contacts in the mixed solvent simulation as well as in the pure water simulation. The peak of the water-oxygen contacts is shifted towards lower affinity within the mixed solvent simulations. This is due to the accumulation of iPrOH at the surface, reducing the accessibility for water molecules.

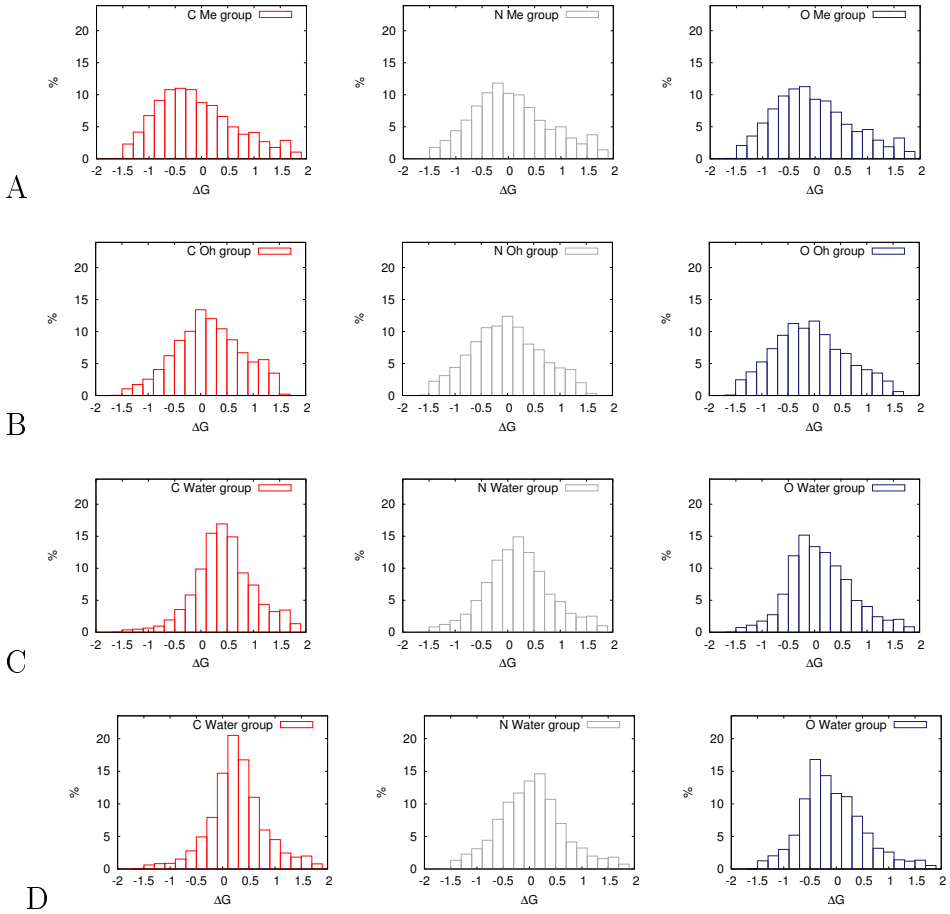


Figure 3.9 – Histogram of solvent distribution around certain atom groups. The first column shows the occurrence of solvent in close distance to carbon atoms, the second to nitrogen and the third to oxygen, respectively. A) Distribution of the Me-groups using a maximal distance of 4.2Å. B) Same for OH-group using a distance criterion of 3.2Å. C) Distribution of water oxygen from the mixed solvent around the atom groups within 3.2Å. D) Distribution of water oxygen from the water only simulation.

3.2.6 Solvation profiles for iPrOH

Proximal radial distribution functions allow the identification of the distance of the peak of solvation concentration around certain atom groups and the protein's surface as a whole, while grid calculations provide approximated information on local solvent agglomerations. To investigate the binding affinity of the solvent in respect to individual residues, the information on the distance of the peak to the protein's surface is taken as a cut-off value to calculate the amount of solvent around a single residue. This can be done for each residue of the protein to identify high and low affinity binding regions, but can also be summed up to identify favored and disfavored types of residues. This method can be extended to analyze the behavior of the solvent in a certain

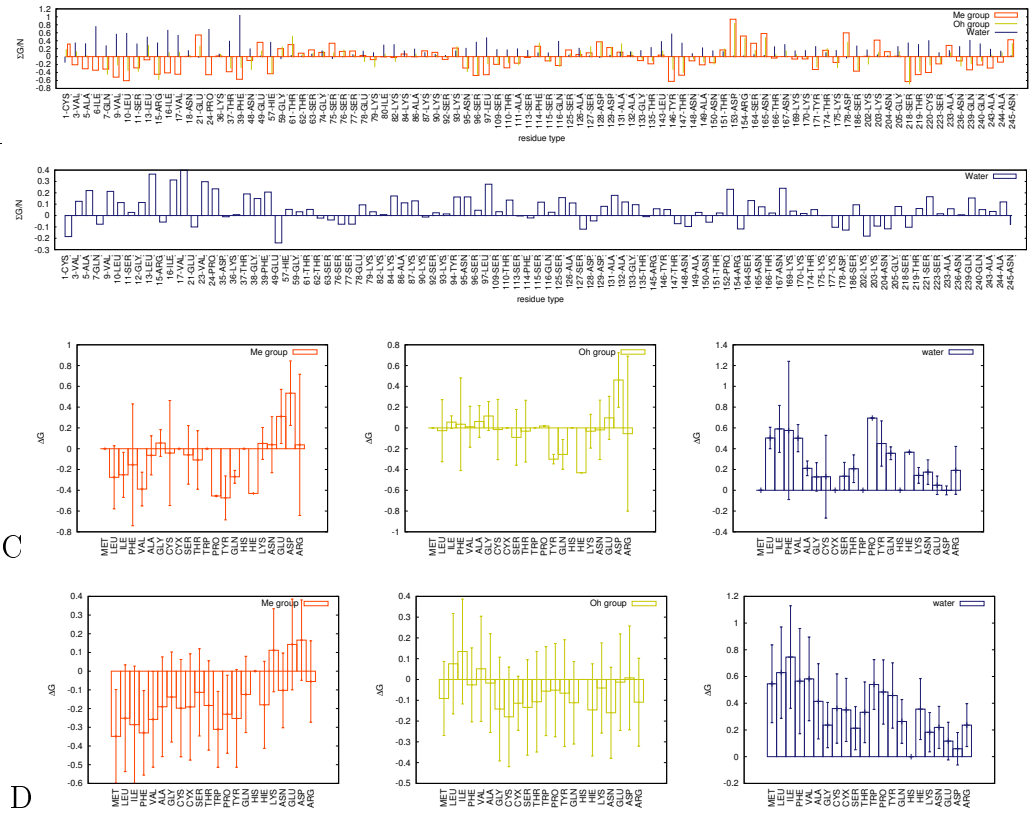


Figure 3.10 – Profiles for solvent accumulation near certain residues. Shown are results for pdb1acb receptor (A - C) and both test sets (D). Red colors Me-group, yellow OH-group and blue water. A) Profile for the mixed solvent simulation. B) Profile for the water only simulation. C) Summed up ΔG values around residue types. The average value is given as a bar with an error bar representing the standard deviation. Residues are roughly ordered by their hydrophobicity with hydrophobic residues on the left and hydrophilic on the right. D) ΔG values for solvation sites near a certain residue type, averaged over all test set proteins.

environment of distinct chemical properties. Therefore, all residues with less than 20% of their surface accessible to the solvent are discarded, while the remaining residues are counted as surface residues. Around each residue, the concentration of the solvent is counted on the grid, using the turning point after the highest peak as a distance cut-off value. For the Me-group, the cut-off is determined as 5.0 Å and for the OH-group, as well as the water, 3.6 Å. The sum of the ΔG values is afterwards normalized for each residue by the amount of grid points around the residue itself. If the values are summed up for a residue type, the resulting value is again normalized by the amount of residues of one type that occurred at the surface.

Such kind of analysis enables the investigation of the population of solvent around individual residues. As the profile of a single protein shows (compare Figure 3.10A and B), the distribution of solvent around one type of residue

varies along the surface. Since the local structure and the flanking residues influence the accumulation of solvent, it is not possible to identify a certain accumulation of solvent around residue types (Figure 3.10C). Even though the iPrOH molecules on average behave as expected, e.g. Me-group favors hydrophobic residues over hydrophilic, the standard deviation shows, that predicting the affinity of the solvent at a residue of certain type, cannot be done accurately.

As Figure 3.9 indicates, iPrOH, especially the Me-group, tends to be highly affine to carbon rich regions. Since hydrophobic residues contain a high amount of carbon atoms at the surface, Me-iPrOH can be found more often within hydrophobic than hydrophilic regions. The head group of OH-iPrOH can be found in hydrophilic regions, but is also often located near hydrophobic areas due to the high affinity to hydrophobic residues of the Me-group. To measure the affinity of solvent molecules in regions of higher hydrophobicity or hydrophilicity, for each surface residue, neighbor residues within 5 Å are taken to generate a surface patch for which the average hydrophobicity/hydrophilicity is calculated (hydrophobicity values are taken from Eisenberg et al. (1984)). Afterwards, each grid point providing ΔG values is taken into account that is within the above given cut-off distance to any of the residues within the generated patch. In such a way, the surface is scanned to identify solvent affinity in areas with distinct hydrophobicity.

Figure 3.11 shows the trend of iPrOH and water within certain regions of averaged hydrophobicity. The distribution of ΔG values of the Me-group of iPrOH follows the expected affinity towards hydrophobic regions (Figure 3.11A), showing a clear lower average ΔG for regions with high hydrophobicity and lower for hydrophilic regions. Nevertheless, the standard deviation is relatively high, so that even in very hydrophobic regions a lower amount of Me-iPrOH can be found. This uncertainty is even more reflected when focusing on OH-iPrOH: following the affinity of Me-iPrOH, OH-iPrOH can sometimes be found in hydrophobic regions, but with an even higher variance. There is still a tendency for OH-iPrOH to populate hydrophilic regions, but since the variance is high (see Figure 3.11B), a prediction of OH-iPrOH binding to regions with distinct hydrophilicity is not possible. One major reason is the avoidance of highly polar regions on the surface (Seco et al., 2009) which also might contribute to the saturation of hydrophobic regions on the protein's surface. Water, in contrast, avoids hydrophobic regions and prefers hydrophilic

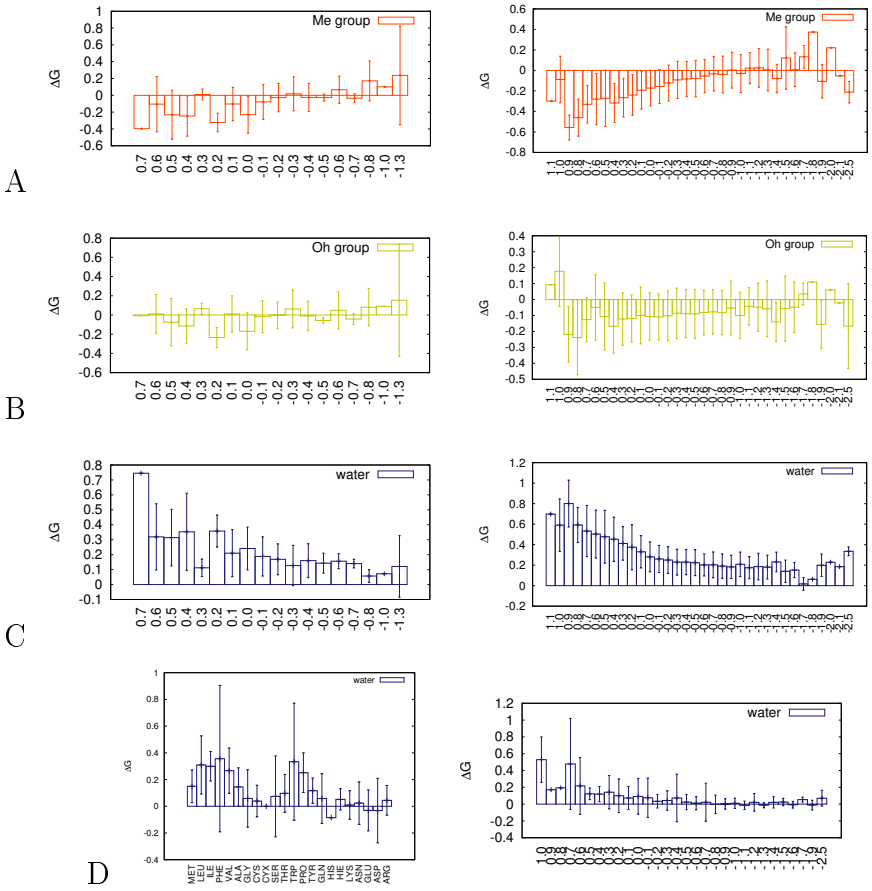


Figure 3.11 – Profiles for solvent accumulation near certain hydrophobicity patches. Shown are results for pdb1acb receptor (A, B and C left column). A, B and C right column: same but for the entire test set of 35 proteins. A) shows the affinity of the Me-group, B) the OH-group and C) the affinity of water in the mixed solvent simulations. D) Shows the accumulation of water from the only water simulation around residue types (left) and patches of certain hydrophobicity (right). Positive values on the x-axis account for hydrophobic regions and negative values for hydrophilic (Eisenberg et al., 1984).

areas. The fact that the affinity of water at hydrophilic areas is still in a more positive regime than in water only simulations, allows the assumption that iPrOH is still, at least transiently, present in most parts of these areas. Water molecules within the pure water simulations show also a trend to hydrophilic areas but also a general trend to be more affine at the surface compared to the mixed solvent simulations (Figure 3.11D). Since water only simulations have been performed only for the protein-protein test set, the sampling is slightly worse than for the mixed solvent simulations (which also included the protein inhibition test set).

3.3 Solvation of binding sites

Accumulation of iPrOH at the binding site of small ligands was investigated by carrying out experiments as well as computer simulations. Seco et al. (2009) could identify the binding sites for small ligands for a subset of their eight protein test set and estimate the binding affinity of the generated clusters representing the ligand binding site. Beuming et al. (2011) clustered cold spots and hot spots of water around proteins known to bind drug-like molecules and tried to identify the binding sites clustering the coldest spots, which are the spots mostly disfavored by water. For specific protein-protein binding sites high or low affinity spots for water could be found within experiments and simulations (Rodier et al., 2005, Reichmann et al., 2008). Using a different approach, the penalty of replacing water was estimated using continuum solvent calculations to identify regions with lower costs of water replacement in the binding sites (Fiorucci and Zacharias, 2010a).

In the following chapter, the solvation of binding sites of different types of interfaces will be explored and differences to the solvation of the entire surface illustrated. Low hydration as well as high solvation of the protein interfaces will be the focus of the analysis and the predictive power of low and high solvation patterns presented.

3.3.1 Water in protein-protein binding sites

Ten protein complexes with diverse binding site characteristics have been investigated with pure water simulations. Since hydrophobicity plays an important role for several association processes and hydrophobic residues can often be found in the center of the binding sites (Jones and Thornton, 1997a, Tsai et al., 1997, Chandler, 2005), the focus is set on low hydration sites. Also, water that is not bound with high affinity to the surface can be less costly replaced upon complex formation. Water molecules bound to the surface with high affinity, as well as the absence of these, can influence the binding process and complex stability (Nagendra et al., 1998, Tarek and Tobias, 2002, Barillari et al., 2007, Reichmann et al., 2008, Samsonov et al., 2008).

Radial distribution functions can identify the relative solvent density around a protein's surface. To estimate this relative density for the binding site of the protein, only those residues were taken into account which are found in the binding site of the known complex. These distributions can differ significantly

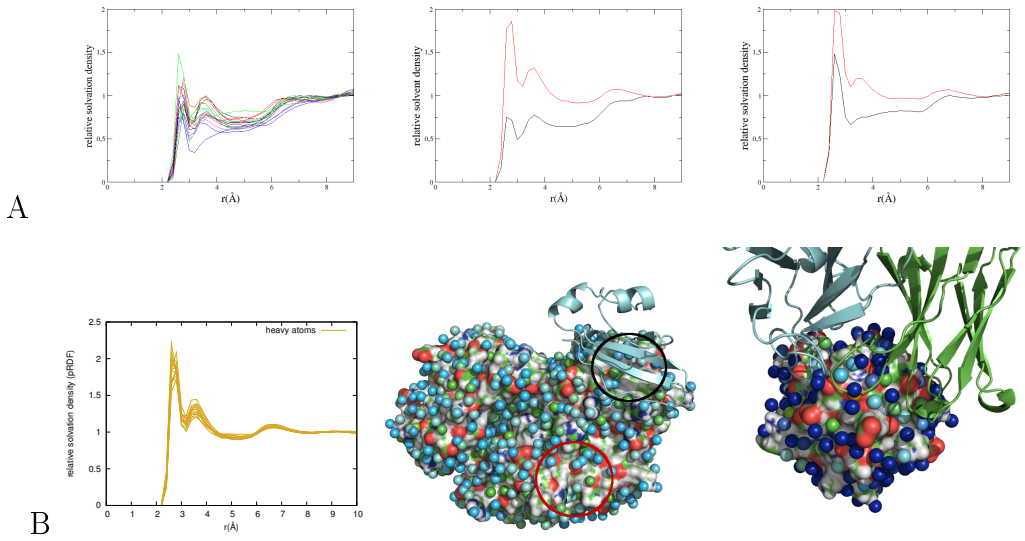


Figure 3.12 – A) pRDFs for the binding site (left) for all proteins in the test set. One has to emphasize, that the approximated bulk value used to normalize the pRDFs differs in the case of binding site pRDFs from those of the entire surface. This is due to the fact, that in the case of binding site pRDFs in a distance of 10 Å not only bulk solvent contributes to the pRDFs but also other parts of the surface. Therefore, pRDFs for the binding site and the entire surface cannot be compared directly concerning the height of the peak. Proximal relative solvation density for the entire protein (red line) and the binding site only (black line) for receptor of pdb1buh (middle) and ligand of pdb2jel (right). B) On the left, the pRDFs for the entire surface are given, showing clear differences in the relative density of the binding sites. In the middle, the receptor of pdb1buh is shown in surface representation while the ligand is shown as cartoon. The small spheres indicate hydration sites with at least 3 fold the expected bulk value. The black circle indicates the binding region, the red circle an area on the surface showing similar solvation density. On the right, same is shown for the ligand of pdb2jel.

from those including the entire protein. Figure 3.12A shows the wide variance of the local relative density of water and can be clearly distinguished from the relative density along the entire surface (Figure 3.12A, B and C). As shown in Figure 3.12B for the receptor of pdb1buh, the proximal radial distribution function for the binding site clearly differs from the overall surface. This is also reflected in the grid representation of hydration sites, showing less high affinity spots in the binding site than on many other spots on the surface (Figure 3.12B: black circle indicates binding site, red circle indicates another spot of similar water density).

Figure 3.12C shows the ligand of pdb2jel, which has the highest peak among all relative density profiles for binding sites. In this case, the distribution of water is relative even around the protein's surface, showing only small spots with less high affinity. Since this protein is rather small, no large clusters of increased or lowered affinity can be found at first sight as they have been found for the receptor of pdb1buh. Also, the binding site of pdb2jel is more

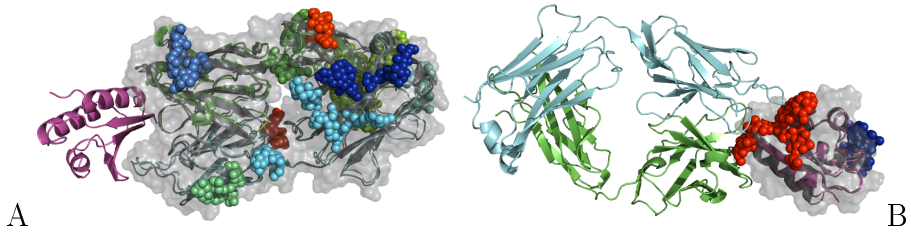


Figure 3.13 – A) The receptor of pdb2jel is shown in mixed surface/cartoon representation and the binding ligand shown as cartoon. Small spheres indicate the clusters of low solvation, color coded by their average ΔG value. Red indicates most positive ΔG among the results. B) Same for the ligand of pdb2jel (same protein partners as in A), the view is rotated to allow the best view on the binding site), shown as cartoon/surface, showing only two clusters of low hydration. The second cluster can be found in pdb entries as a second binding site (e.g pdb3eza).

hydrophilic than the binding site of pdb1buh, which shows the trend found within this test set, that the peak of the binding site correlates roughly with the amount of hydrophilic residues on the surface. Interestingly, it correlates only weakly with the average hydrophobicity of the binding site, but the peak of the pRDFs for the entire surface correlates stronger with the overall hydrophobicity of the entire surface. In any case, none of the peaks of the binding site pRDFs has a higher amplitude than for the entire surface. This might be due to varying concentrations of hydrophobic residues in the binding site, that the amplitude is lower than on the average surface. Hydrophobic residues have been found relevant for the prediction of binding sites (Jones and Thornton, 1997b, de Vries and Bonvin, 2008).

Since small spots of high affinity influence the average distribution, clustering of low affinity areas was used to identify continuous patches of low solvation. Therefore, grid points with a positive ΔG value are clustered together using a distance cut-off of 1.8 Å between each grid point, leaving numerous small clusters of low solvation. Clusters with a size smaller than 10 hydration sites are discarded in order to not take into account single low solvation spots in areas of high solvation. Subsequently, the small clusters are clustered again to bigger clusters, using a distance cut-off of 4.0 Å. Among the larger clusters, those were taken for analysis, that at least cover as much surface as a possible binding partner would cover. In most of the binding sites, at least some overlap with one or more of these clusters of low hydration can be found but also other regions are widely covered by low hydration sites. Since a protein can be involved in more than one binding event, it is possible that those regions are part of other interaction pathways.

Table 3.2 – Prediction statistics of the three clusters with lowest solvation.

	receptor	ligand	average
Sensitivity	0.22	0.63	0.42
Specificity	0.11	0.30	0.20
Accuracy	0.67	0.60	0.64

Definition of Sensitivity: TP/(TP+FN); Specificity: TP/(TP+FP); Accuracy (TP+TN)/(TP+TN+FP+FN) with TP true positive, TN true negative, FP false positive and FN false negative.

After clustering, around 52% of the protein's surface is covered by unfavorable hydration sites on average and also 56% of the known binding site is covered with clusters of low solvation. In some cases, a binding site is covered by more than one cluster of low solvation, while one cluster contributes most to the coverage, other clusters can be found in the rim region, only briefly overlapping with the known binding site. In 12 from 20 binding sites, the binding site is at least, partially covered with the cluster of lowest hydration and 17 are covered by one or more of the lowest three. Except for one binding site (see Figure 3.13A), at least one region with unfavorable hydration could be found overlapping with the binding site. While in some cases it was just a small overlap of 10%, it could be up to a 90% overlap with the binding site in other cases. Taking all unfavorable hydration clusters into account, 56% of the binding sites are covered by all and 42% of the binding sites by the three clusters with the lowest affinity.

In some cases, single spots determine the unfavorable hydration at the binding site (see Figure 3.13B), while in other cases the binding site is subdivided into several clusters of low hydration. The latter is not surprising, since polar contacts play a crucial role in many binding processes and therefore, high affinity spots for water are likely to be found within the protein binding sites. Also, no binding site was found within this test set, that was entirely covered by unfavorable hydration clusters, which is not surprising, since 15 out of the 20 binding sites are more hydrophilic than hydrophobic. In some cases, the amount of hydrophilic residues goes up to 80% within the binding site, allowing only a small overlap with low hydration clusters (e.g. receptor pdb1ay7), while binding sites with more than 50% of hydrophobic residues are covered in a larger scale.

The size and number of clusters generated this way depends mostly on the minimum size and the maximum distance discriminator for the clustering procedure. Therefore, using such cluster analysis for low hydration patterns on the surface is not straight forward and will largely depend on the proteins within the test set. To predict a binding site using low hydration profiles on the surface might also not be successful, since, at least in this study, the top3 low hydration clusters cover 36% of the surface and all hydration clusters 52% (for prediction statistics see Table 3.2). The overprediction can be reduced to lower values by changing the minimum cluster size, the distance discriminator and the minimum ΔG value, but hydrophilic parts of the binding site will probably not be covered and the overall sensitivity reduced.

3.3.2 Binding site solvation using a mixed solvent

Mixed solvents provide different behavior for the different chemical groups enclosed within this solvents. Mixed solvent simulations allow accounting for the density distribution of all chemical groups of the solvents and therefore, different features of the protein's surface can be investigated. In the case of iPrOH/water simulations, three different chemical groups are available: water tends to avoid hydrophobic regions, the Me-group of iPrOH tends to avoid polar and hydrophilic regions, while the OH-group is attracted to hydrophilic regions in principle, but can be found near hydrophobic residues due to the corporate effect of both Me-groups of iPrOH.

To compare the overall affinity of the binding site with the non-binding site, ΔG values of the corresponding residues are summed up and normalized by the amount of counted solvation sites. This value represents the average affinity of a solvation site within the binding site or the non-binding site. It has to be emphasized, that only the known binding site for the given complex is taken into account and other possible binding sites are neglected.

Results including the protein-protein test set show for Me-iPrOH, that the average affinity, in 15 of 20 cases, is higher for the binding site in the order of two to five times than the average value for the non-binding site. For OH-iPrOH, this was found for 13 out of 20 proteins with up to two times the average non-binding site value. Within these mixed solvent simulations, the affinity of water molecules does not differ so clearly from binding site to non-binding site. In ten cases, the affinity was higher in the binding site and vice

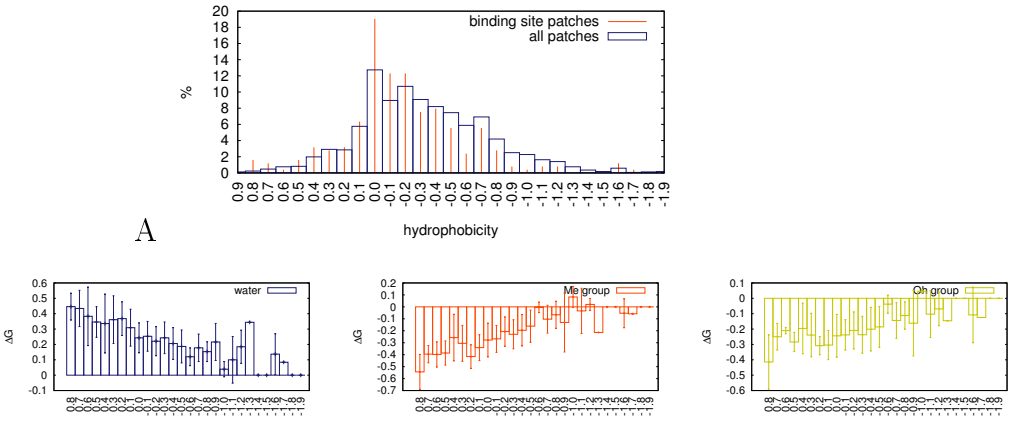


Figure 3.14 – Results are shown for the 20 proteins from the protein-protein binding site test set. A) Distribution of patches of certain hydrophobicity on the entire surface and the binding site. For patches generated for binding site residues, also residues in the proximity of this residue not in the binding site were taken into account. B) Affinity of iPrOH in the binding site. Compare Figure 3.10 for the entire surface.

versa. The difference between binding site and non-binding site is within the range of only 1 to 20%. No clear trend can be observed for the inhibitor test set, including only the binding site of the inhibitor. In the case of Me-iPrOH, the same amount of binding sites have higher affinity than the non binding site and vice versa. OH-iPrOH is slightly more often affine towards the binding site (seven to five), but often the differences are small. Only for water does the binding site seem less affine than the non-binding area, resulting into four proteins with a slightly higher affinity for water than the non-binding areas and in eight cases the non-binding area could be found more attractive.

Random spots on the surface of the protein-protein test set were used to validate the independence of the difference in ΔG between binding site and non-binding site from the size of the observed surface. Therefore, ten random spots with the same number of surface residues as the known binding site were generated for each protein. These spots were created around one random non-binding site residue and are continuous on the surface of the protein while not overlapping with the known binding site. The same protocol was applied as used for binding site versus non-binding site solvent affinity differentiation.

In the case of Me-iPrOH, the binding affinity of one random spot versus the rest of the surface was found tending towards the rest of the surface. Averaging over all ten random spots, 7.2 out of 20 proteins were found with a higher affinity of the random spot compared with the rest of the surface (with a standard deviation of 2.8). This is only half the value as found for the known

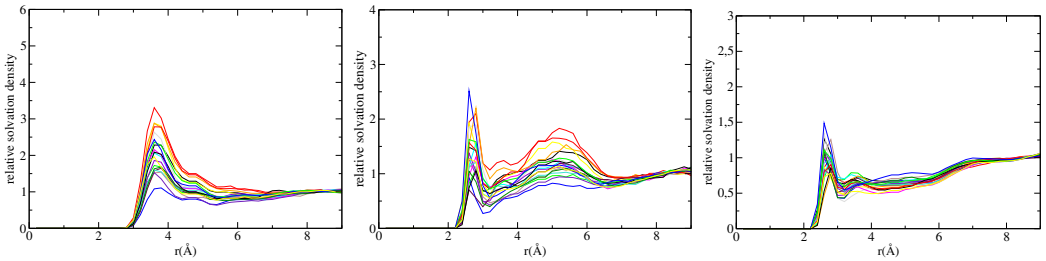


Figure 3.15 – pRDFs for the known binding region are shown. From left to right: pRDFs for Me-iPrOH, OH-iPrOH, water. The The estimated bulk value for the binding site pRDFs is higher than for the entire surface, since solvent detected in 9 or 10Å distance from the known binding site might be in the bulk or at the protein surface, 9 or 10 Å away from the binding site.

binding site versus the non-binding site. The OH-group average was 7.8 (2.6). The differences for water were 8.1 in average (4.8). This shows on one hand, that the higher affinity of iPrOH at the binding site is not simply based on the size, but holds some special features attracting the isopropyl alcohol, and on the other hand, that there are also several other spots on the surface, that have a high concentration of iPrOH bound to the surface.

Interestingly, the affinity of iPrOH within regions of certain hydrophobicity in the binding site does not significantly differ from those of the entire surface. The only difference found was the lower population of hydrophilic patches in the binding site (see Figure 3.14) and therefore, the deviation for those patches was higher, if they existed at all. It is possible, that the higher affinity of Me-iPrOH for the binding site shown above, is indirectly effected by the absence of very hydrophilic patches within the binding site, rather than the overall hydrophobicity of the site.

If higher cut-off values are chosen (-1.2 kcal/mol, Me-iPrOH and OH-iPrOH resulting in about 50% surface coverage in each case), the number of solvation sites are less (Figure 3.17B). However, these sites are still distributed over the surface and therefore, mapping onto the surface would still result in a large percentage of surface covered (77% for iPrOH alone). This also results in a high overlap with the binding site of 85%, while high affinity spots of water cover parts of the rest of the binding site, ending up with 85% surface coverage and 90% binding site overlap. Although most of the surface is covered by solvation sites, the density of the distribution varies strongly along the surface. This difference in the density of the distribution can be used to favor clusters of solvent at the surface.

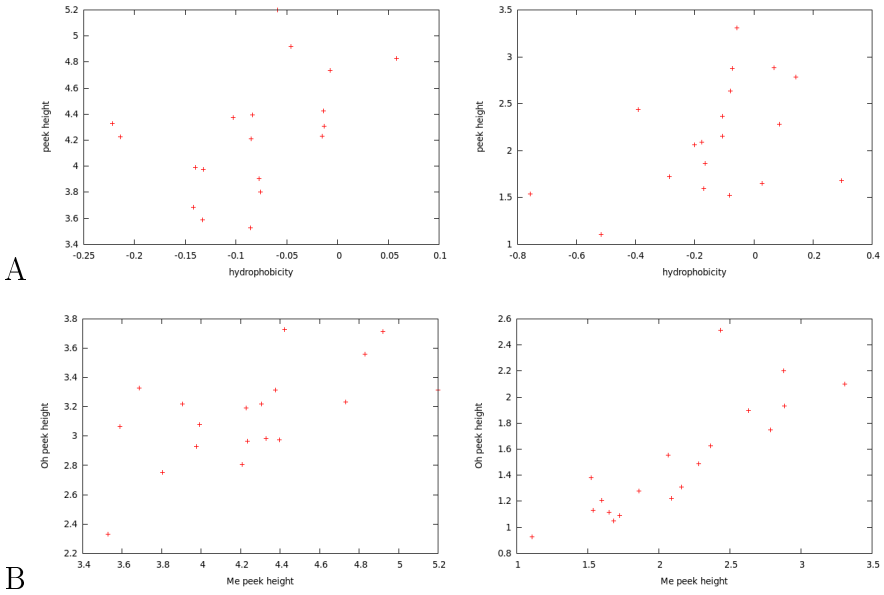


Figure 3.16 – Comparing pRDFs for the entire surface (left) and the binding site (right) A) Me-iPrOH versus the hydrophobicity of the surface. B) Peek of Me-iPrOH versus peak of OH-iPrOH.

Proximal radial distribution functions for the entire surface show a slight variance for each protein, defined by the physicochemical composition of the surface. The pRDFs for the three different solvent groups perform differently for the known binding sites, compared to those for the entire surface (see Figure 3.15). A higher variance in the peak of the relative solvent distribution can be observed, pointing on diverging composition of physicochemical properties within the binding site compared to the average distribution on the surface.

It is found, that the height of the peak of Me-iPrOH only slightly correlates with the overall hydrophobicity at the surface (Figure 3.16A on the left) and the correlation is even weaker if the binding site pRDFs are compared to the hydrophobicity of the known binding sites. The height of the peak of the pRDFs for OH-iPrOH does not correlate with the hydrophobicity at the surface and only slightly if comparing the binding site pRDFs with binding site hydrophobicity. The height of the peak of the pRDFs for any solvent at the binding site is thereby not correlating with the height of the peak of the solvent around the entire protein. Interestingly, the height of the peak of the Me-iPrOH pRDFs correlates stronger with the OH-iPrOH pRDFs peak height for the binding site, compared to the entire surface (see Figure 3.16B).

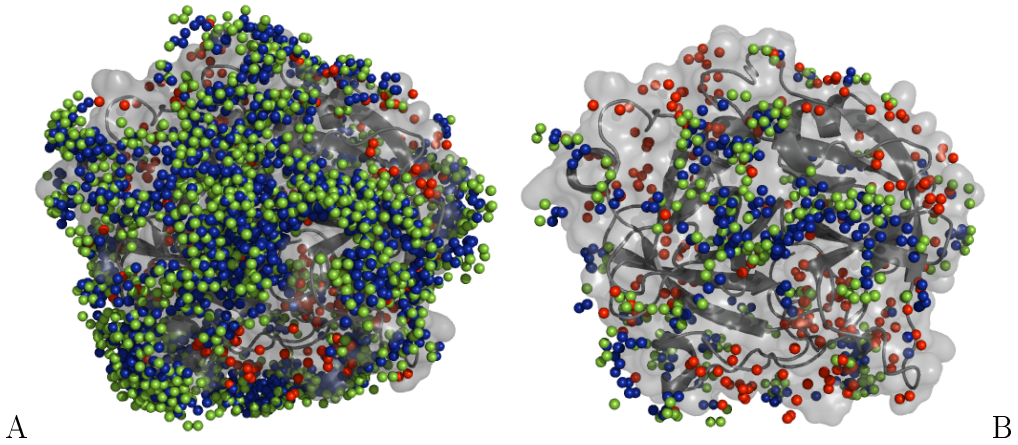


Figure 3.17 – Solvation spots of Me-iPrOH (blue spots), OH-iPrOH (green spots) and water (red spots). The receptor of pdb1acb is shown in surface and cartoon representation. A) Low cut-off values for iPrOH (-0.84 kcal/mol) and water (-0.5 kcal/mol) B) Cut-off values for iPrOH set to -1.2 kcal/mol.

High affinity solvation sites for Me-iPrOH, OH-iPrOH and water can be found spread over the surface of the proteins. Depending on the cut-off value chosen, most of the surface is covered by such spots, while the density of the distribution along the surface varies. As reported before, iPrOH avoids highly polar regions and is therefore only rarely found within this region. Using -0.84 kcal/mol as cut-off value (see Seco et al. (2009)), large areas are covered by iPrOH (about 90% of the surface residues are within 4.0 Å to one solvation site) and water (77% of the surface), as shown in Figure 3.17A.

Binding site prediction for the protein-protein test set

To account for protein-protein binding site prediction with the help of iPrOH, the different chemical groups were clustered separately (see Figure 3.18 for a schematic representation of the procedure). Since it was found, that the binding site has a higher affinity for iPrOH than the rest of the surface, as a first step, small clusters of solvation sites with high affinity were identified. Therefore, all Me-iPrOH solvation sites with a ΔG lower than -1.2 kcal/mol (approximately the eight fold of the expected bulk value, populating half of the protein's surface) were clustered.

The same was done for OH-iPrOH solvation sites, while for water hydration sites, the lower limit for the ΔG was set to -0.5 kcal/mol ($\sim$2.5 fold the expected bulk value), since the ΔG distribution for water is slightly shifted towards more positive values. This cut-off allows accounting for approximately the same amount of solvation sites (compare Figure 3.8C). The minimum clus-

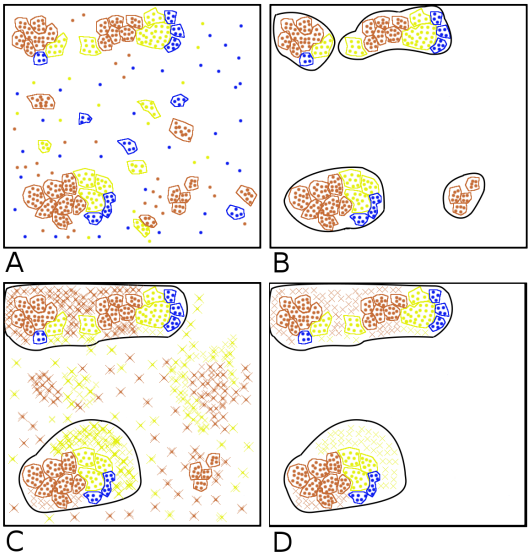


Figure 3.18 – A) High affinity solvation sites of Me-iPrOH (orange spots), OH-iPrOH (yellow spots) and water molecules (blue spots) are clustered separately. Accepted clusters are shown as spots within colored borders. B) Solvation sites outside of the clusters as well as small clusters are discarded. Remaining clusters merged to larger clusters of (mixed) solvation sites (black borders around the clusters). C) Medium affinity solvation sites (colored crosses) are added to the existing high affinity clusters if they are in the close proximity of high affinity clusters (black borders). D) Solvation sites outside of the clusters are discarded and up to three of the remaining clusters returned as prediction.

ter size was set for all three to include at least 2% of all solvation sites, which reduced the amount of solvation sites by about 50% and resulted in a high amount of small high affinity binding sites (Figure 3.18A). These were then merged, which resulted in larger patches of mixed solvation sites for iPrOH and water (Figure 3.18B). The size of each cluster had to be 10% of the left-over solvation sites (reducing the amount of solvation sites by additional 50% on average). Smaller clusters were deselected in this way and only clusters kept, which were populated by several chemical groups or at least by a very dominant one. Since iPrOH avoids highly polar regions and water is attracted by these regions, favored areas were the hydrophobic interaction of iPrOH as well as the polar interactions of water are close to each other, increases the probability of finding the binding site among clusters of solvation. Hydrophobic binding sites would not be discarded, since iPrOH itself can be found in clusters with high affinity in close proximity. Unfortunately, larger polar regions can be overlooked because of the low amount of high affinity hydration sites of water close to each other and the lack of iPrOH.

Subsequently, larger spots of solvation sites with medium high affinity for iPrOH (between -1.2 and -0.8 kcal/mol, eight to four fold the expected value)

Table 3.3 – Prediction accuracy of the top3 clusters from high solvation clustering in mixed solvents.

	receptor	ligand	average
Sensitivity	0.41	0.66	0.53
Specificity	0.20	0.45	0.33
Accuracy	0.66	0.70	0.68

Definition of Sensitivity: TP/(TP+FN); Specificity: TP/(TP+FP); Accuracy (TP+TN)/(TP+TN+FP+FN) with TP true positive, TN true negative, FP false positive and FN false negative.

Table 3.4 – Detailed sensitivity and specificity for the protein-protein test set

pdbcode	partner	type	sensitivity	specificity
1ACB	receptor	enzyme	83	70
	ligand	inhibitor	72	45
1AY7	receptor	enzyme	57	16
	ligand	inhibitor	100	30
1BUH	receptor	other	0	0
	ligand	other	87	45
1AKJ	receptor	other	17	10
	ligand	other	0	0
1D6R	receptor	enzyme	76	27
	ligand	inhibitor	67	21
1KAC	receptor	other	42	14
	ligand	other	61	24
1KTZ	receptor	other	44	6
	ligand	other	79	67
1TMQ	receptor	enzyme	82	59
	ligand	inhibitor	79	63
2JEL	receptor	bound antibody	0	0
	ligand	antigen	92	65
1IQD	receptor	bound antibody	4	2
	ligand	antigen	69	69

Sensitivity and specificity for the proteins in the protein-protein test set. For definition of sensitivity and specificity see legend of Table 3.3.

were clustered around the existing high affinity regions to smoothen the edges of the clusters on one hand and on the other hand to connect separated clusters (Figure 3.18C). Therefore, the medium high affinity clusters had to be close to high affinity clusters of the solvent, with a maximum distance of 4.0 Å. The final clusters were ranked by size and three clusters of the solvation sites were kept as a maximum (Figure 3.18D). Protein residues within 4.0 Å of such high affinity sites were mapped as possible binding sites.

Table 3.3 shows the results for the prediction of the protein-protein binding sites. On average, 50% of the known binding sites were found close to high affinity solvation sites. Among all surface residues, for the larger receptor proteins 33% of the surface was counted as possible binding sites and for the smaller partner protein 38%. This is much larger than the known binding site size (which is 15% of the total surface on average) and therefore, the specificity of the prediction is rather low and only 33% of the predicted residues are actual binding site residues, which reduces the accuracy of the prediction to 68%.

In half of the cases, more than two-thirds of the known binding site is predicted to be involved in binding (Figure 3.19A and C, Table 3.4). For the proteins pdb1ay7 receptor, pdb1kac receptor and pdb1ktz receptor, large parts of the surface were predicted, resulting in some overlap but very low specificity and the binding site is completely missed by the prediction for four proteins (Figure 3.19B, Table 3.4): in the case of the receptor of pdb1akj, only a small overlap with the known binding site could be found, while the peptide binding site of the MHC molecule was highly populated by high affinity solvation sites. For the ligand of pdb1akj, other proteins are also known to bind at a different binding site, but the high affinity clusters of iPrOH were only partially overlapping with these binding sites.

In the case of the receptor of pdb1buh, the ATP binding site was fully covered by solvation sites, while the binding site of the ligand of pdb1buh was found to be highly affine for Me-iPrOH but only on a small area of the surface and was therefore discarded during prediction. In contrast, the known binding site of the ligand of pdb1buh was predicted accurate and other parts of prediction could be found overlapping with an additional binding site (pdb2ass, Hao et al. (2005)). The accuracy and sensitivity for proteins belonging to the class of so called “other“ proteins varies strongly. Often, a large overlap with the known binding site can be found (except for pdb1akj and the receptor of pdb1buh), but also other areas on the protein are predicted.

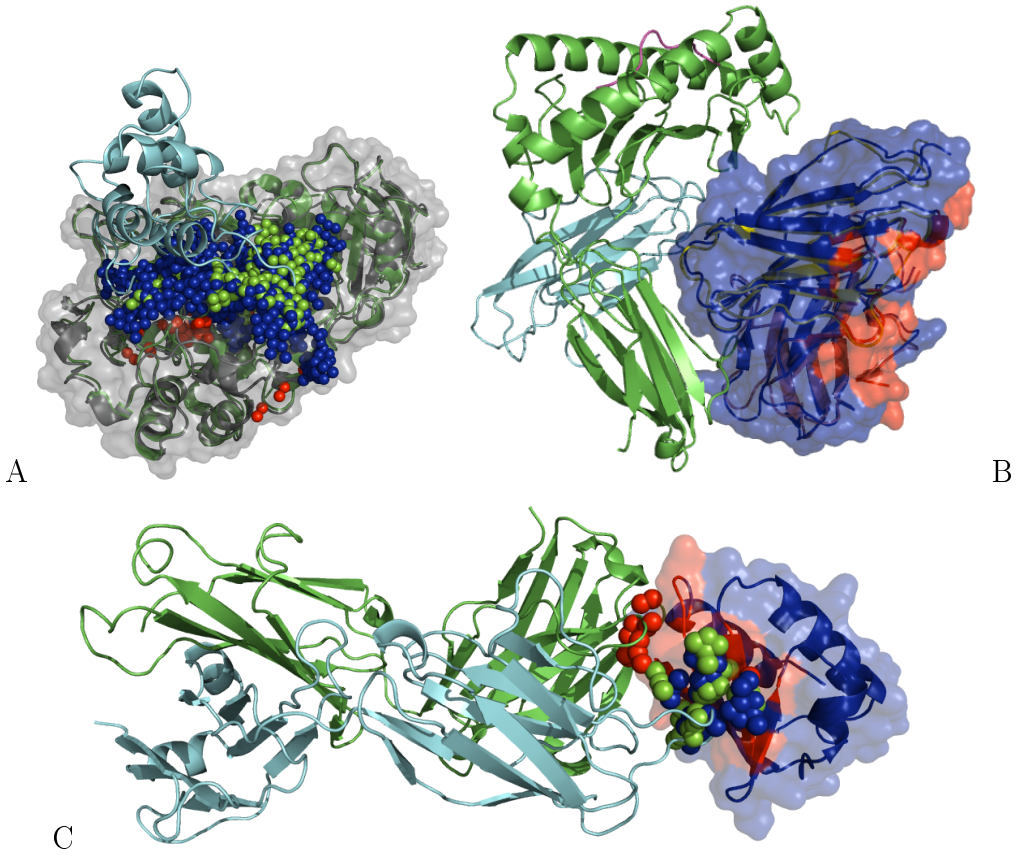


Figure 3.19 – A) Receptor of pdb1tmq shown in surface representation and the ligand in cartoon representation. Small spheres represent the high affinity solvation sites for Me-iPrOH (blue), OH-iPrOH (green) and water (red). B) Ligand of pdb1akj shown in surface representation and the receptor as cartoon. The high affinity solvation sites were mapped onto the surface, indicated by the red surface color on the ligand. Blue surface represents residues not predicted as binding site. C) The ligand of pdb2jel shown as surface and the receptor as cartoon. Spots mark solvation sites as described above.

For the both antibodies in this test set (pdb2jel and pdb1iqd), only a very small amount of solvation sites populated the antigen binding site, while the prediction for the antibody binding site on the surface of the antigens was in good agreement with the known binding sites (for pdb2jel compare Figure 3.19C). For each partner of the four enzyme-inhibitor complexes in this test set, an overlap of the prediction with the known binding site could be found. With the exception of the receptor of pdb1ay7, all predictions were in good agreement with the known binding site and also the prediction accuracy was very good. In case of the receptor of pdb1d6r, the binding site of the inhibitor was largely overpredicted, while for the ligand of pdb1d6r nearly all remaining high affinity sites could be found correlating with another known binding site (pdb2iln, Capaldi et al. (2007)).

Binding site prediction for small protein-protein binding inhibitors

To identify the binding sites of the small drug-like protein-protein binding inhibitors in the second test set, an adjusted approach for clustering was used as described above. The parameters had to be adapted to the differences in protein-protein binding sites compared to protein-small ligand binding sites. Some of the organic inhibitors are bound in cavities, larger ones in very concave regions. In contrast to protein-protein binding sites, these cavities might be much smaller. To account for these differences, the parameters used above had to be adjusted in order not to exclude small cavities.

While steps A) and B) of the clustering procedure was performed (Figure 3.18A and B) also for protein inhibitor binding site prediction, steps C) and D), including lower solvation values, were skipped. Instead, the ΔG value used in step A) was reduced to -1.0 kcal/mol, increasing the size of small clusters. Otherwise, most of the cavities would have been overlooked by default, since these sites are smaller on average than the minimum requisite of solvation sites used to predict protein-protein binding sites. To directly account for the size of small ligand binding sites, the minimum number of elements within one cluster was slightly reduced, while the rest of the settings was used as described above.

Interestingly, within many cavities of this test set, a certain amount of water was found. While high affinity water hydration sites are sometimes only overlapping partially with lower affinity or density, in at least three cases (pdb1m47, pdb2o8t, pdb1cqr) water populates the surface of the binding site by 30-45%. Within eight of the eleven binding sites, water clusters with more than 10% binding site coverage were found, while 11% of the entire surface were covered by water clusters. The specificity of 25% shows a noticeable amount of water clusters was within one of the given binding sites.

Including iPrOH, 36% of the surface of the proteins was covered by clusters of high affinity solvation sites, resulting in a sensitivity of 45%, a specificity of 19% and an accuracy of 62%. In each case, the binding site was at least partially overlapping with high affinity clusters by 25% and six were covered by 60 to 100%. Only about 15% of the highest solvation spots ($<$ -1.4 kcal/mol) of each solvent are to be found within the binding site. Since a protein can be involved in more than one binding process and it is known, that the proteins within this test set bind to other proteins, additional high affinity spots may point to additional binding sites.

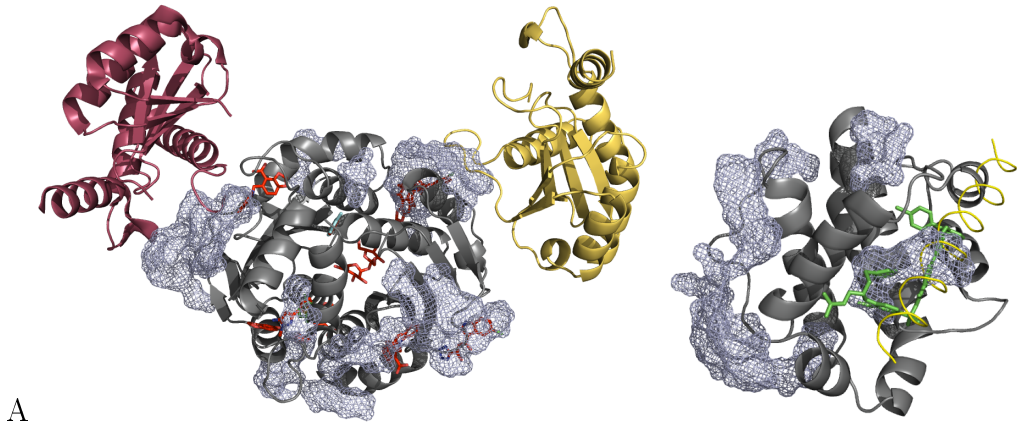


Figure 3.20 – A) pdb1bi4 in complex with ligands and binding partners (pdb3lpu, pdb1qs4, pdb3ao1, pdb3nf6, pdb3ovn). The high affinity clusters are shown in blue mesh representation. B) Protein pdb1r2d shown in gray cartoon and clusters shown in blue mesh. The inhibitor is shown in green sticks (from pdb2yxj), and the protein partner in yellow ribbon at the same position (from pdb2p1l). Parts of the binding site were not available for solvent molecules due to a closing of the upper end of binding site with the short α-helix.

One example for a protein with multiple binding sites is the catalytic domain of HIV-1 integrase (pdb1bi4, see Figure 3.20A), for which several crystal structures exist. Besides several other small ligand binding sites, the partner protein binding site can also be found partially overlapping with clusters of high affinity. Some proposed fragment binding sites near known binding sites (pdb3ao1 (Wielens et al., 2011)) as well as a novel binding site proposed by fragment design (pdb3ovn (Wielens et al., 2011)) were not identified among the predicted high affinity clusters. In case of Bcl-XL (pdb1r2d, Figure 3.20B), the inhibitor binding site is the same binding site as for its partner protein. Besides this region, a large binding site was predicted, for which in several PDB entries crystal contacts were found.

3.4 Conclusion

Explicit solvent simulations enable the investigation of solute solvent interactions and of the behavior of solvents within different parts of the protein. In experimental studies, the population of water was found to vary in individual binding sites and depending on the water content one distinguishes between wet and dry interfaces (Janin, 1999, Chandler, 2005). Nevertheless, several studies showed that binding sites contain more hydrophobic residues than the average surface (Jones and Thornton, 1997a, de Vries and Bonvin, 2008). Therefore, several approaches attempted to identify unfavorable solva-

tion sites on protein surfaces in order to further predict possible binding sites (Barillari et al., 2007, Beuming et al., 2011).

In this study, solvation of protein binding sites with water was investigated using explicit solvent simulations. Proximal radial distribution functions showed, that the solvation varies strongly among the different binding sites, dependent on their composition. Binding sites with a high amount of hydrophilic residues show a pRDF with a higher peak of solvent density near the surface than hydrophobic sites. Also, the form of the distribution depends on the composition of the binding site: the second peak found in pRDFs, and mostly influenced by carbon contacts, decreases around surfaces with a high amount of hydrophilic residues.

Unfavorable aqueous hydration sites could be found widely distributed over the surface with divergent density. Clustering of such unfavorable sites revealed a significant overlap with known binding sites. Nevertheless, highly polar binding sites were rarely covered by clusters of low aqueous solvation, which excludes such binding sites from the prediction. Although the removal of water upon binding may increase the energy barrier for protein binding, high affinity clusters of water are found within many binding sites, resulting in a low sensitivity of 42% for predicting putative binding sites due to clustering of low energy hydration sites. This is not surprising, since it is found that water molecules can mediate the binding between biomolecules (Ladbury, 1996).

To mimic the behavior of molecules with different chemical properties at the protein's surface, mixed solvent simulations were used. Using isopropyl alcohol and water, high affinity spots were identified in protein-protein binding sites as well as small ligand binding sites. Isopropyl alcohol showed a tendency towards hydrophobic regions due to the methyl groups. Nevertheless, the hydrophilic hydroxyl group could be found binding to hydrophilic regions of the protein. Highly polar areas on the surface were avoided by iPrOH, so that water clusters with high affinity were also included for the prediction of binding sites. While the methyl and the hydroxyl group of iPrOH showed only a low tendency to correlate according to their relative density in the proximity of the entire surface (as shown in the peak height of the pRDFs), a strong correlation for the height of the peaks at the binding site was found. The binding sites were favored by iPrOH over the rest of the surface, which could not be found for random spots on the surface in this magnitude.

The clustering of high affinity solvation sites was in good agreement with known binding sites, resulting in a high sensitivity of 53% in case of protein-protein binding sites. In fact, in many cases the binding site was predicted by a large fraction or not at all. Since, for many proteins within this test set, only one partner is known, additional high affinity binding sites might be related to additional protein binding sites. This might be true in some cases, since predictions for proteins with multiple binding partners agreed well with all known binding sites. Also, high affinity clusters of iPrOH and water were found within binding sites of small organic protein-protein binding inhibitors. In those cases a partial overlap of the binding site with one of the clusters could be found in every protein in the test set. However, in some cases larger parts of the binding site were unavailable for solvent molecules, due to a deformed or closed binding site. Moreover, additional binding sites of the partner protein could be found populated with high affinity clusters of iPrOH and water molecules, while other areas on the surface were avoided.

The information gained by high solvation mapping and binding site prediction is in many cases of high value. The high sensitivity of predictions e.g. for pdb1acb, pdb1tmq and pdb1d6r could be valuable if included into docking methods (compare Chapter 5). Since the antigen binding site of antibodies is approximately known, also the iPrOH/Water predictions for the antigens could result in a successful docking. The overall sensitivity is quite high, especially when additional known binding sites are considered. As the very accurate prediction of the peptide binding site of the MHC molecule (pdb1akj) and the ATP binding site (pdb1buh) have shown, the prediction of small ligands should in principle be possible, if the cavity can be accessed by solvent molecules. Therefore, inclusion of flexibility could improve the identification of small ligand binding sites.

3.4.1 Acknowledgment

Many thanks to Prof. Xavier Barril and PhD Student Daniel Álvarez García who provided the iPrOH/TIP3P input file used in the simulations.

Chapter 4

Desolvation properties of small ligand binding sites

The identification and classification of small ligand binding sites plays a crucial role in computer guided drug development. Therefore, detailed information on the binding site of small ligands or drug like molecules is essential for the development of new lead drugs. If the real binding site is unknown, or alternative binding sites or configurations are in focus, computational methods can help to identify those binding sites and guide towards possible binding hot spots and binding geometries (Nayal and Honig, 2006, Wells and McClendon, 2007, Fuller et al., 2009).

Since small ligands have less possible contacts to the partner protein than larger ligands (e.g. other proteins), those contacts have to be strong enough to interact effectually. Matching this requirement, small binding partners can often be found in cavities with less polar and more hydrophobic properties (Miller and Dill, 1997, Liang et al., 1998, Campbell et al., 2003). Henrich et al. (2010) gave a summary of possible detectable characteristics for small ligand binding sites and protein-ligand interactions: the shape of the cavity, amino acid composition, solvation effects, hydrophobicity and electrostatic potentials. Thus, three major classes of prediction methods have emerged: geometry-based cavity detection algorithms try to identify the most favourable cavities among all clefts on the protein's surface by geometrical pocket description or free accessible volume calculation (Hendlich et al., 1997, An et al., 2004, 2005, Weisel et al., 2007), while energy based methods attempt to find energetically favored positions, e.g. by calculation of interactions with different single probes and

the protein's surface (Wade and Goodford, 1993, Laurie and Jackson, 2005, Brylinski et al., 2007) – an overview of available methods from this two classes is given in e.g. Leis et al. (2010). The third class includes fragment based search and docking procedures, which place different chemical fragments on the protein's surface and calculate e.g. the Gibbs free energy of a fragment at the protein's surface. This enables the most attractive binding sites to be discovered for different chemical groups, which, clustered to a meaningful size, propose a certain chemical ligand composition (Brenke et al., 2009, Fukunishi and Nakamura, 2011, Kozakov et al., 2011). Besides these methods, a fourth class of binding site detection algorithms appeared based on structure comparison methods in order to generate templates of similar sequence or functionality. Algorithms of this class try to identify evolutionary conserved regions to rank cavities generated with geometric approaches (Huang and Schroeder, 2006) or superimpose known structures to filter possible binding sites (Brylinski and Skolnick, 2008). Most methods based on geometry calculations, only present the center of the cavity representing the predicted binding site. Some other methods also return the grid, or probes, used for binding site prediction (Hendlich et al., 1997, An et al., 2005, Weisel et al., 2007) but just a few try further refinement, mostly based also on geometrical calculations (Weisel et al., 2007, Yu et al., 2010, Volkamer et al., 2010).

The method presented in this chapter combines a geometry based approach with more meaningful energy calculations, aiming at three characteristics of small ligand binding sites, the shape and volume of the cavity, solvation effects and indirectly the hydrophobicity of the cavity. Computational effort for calculations of energies is reduced to a minimum using an effective and fast cavity detection algorithm: A pure geometry-based approach, related to fundamental techniques (e.g. Brady and Stouten (2000), Kawabata and Go (2007), Yu et al. (2010)), identifies major cavities for which the desolvation free energies are calculated, that is the cost of removing water upon ligand binding.

The cost for water replacement does not only impact binding site detection, but as well ligand optimization in cases of drug design (Michel et al., 2009, Luccarelli et al., 2010) and is therefore a valuable target for investigations – besides the pure cavity detection. Using desolvation free energy calculations also allows discarding probes which represent regions that are unlikely for ligand binding and allow reduction of the prediction area within a cavity towards more native binding areas.

Accessorily to cavity detection and in contrast to most geometry-based approaches, the algorithm tries not only to predict the most favorable cavity but also to contour potential binding areas within the cavities as well as favored polar contacts between groups of the ligand and the protein's surface, without knowledge of the ligand. The restriction of calculating energies for probes only in or within the proximity of cavities, reduces the computational cost compared to other energy based methods, which often carry out calculations for probes on the entire surface. The method was optimized for a small set of proteins (used to classify binding site predictors in Leis et al. (2010)) and validated on a test set of bound and unbound proteins, widely used to benchmark the performance of several binding site predictors (amongst others: Huang and Schroeder (2006), Yu et al. (2010)).

4.1 Settings and statistical measurements

A published test set was used to evaluate the method. This set contained 48 bound and unbound structures including several proteins widely used to benchmark binding site predictors ((Yu et al., 2010)). It was used to determine the performance of several geometry-based and energy-based methods (see Yu et al. (2010) for a list of a subset of methods evaluated on this test set). Here, evaluation principles were applied which are commonly used and established statistical measurements were employed, in order to be comparable with the performance of previous methods.

4.1.1 Structure preparation and prerequisites

The PDB files were downloaded from the protein data base (Berman et al., 2000) and separated into protein and ligand files. Hydrogen atoms were removed from the protein files and missing atoms added by the tleap program from the AMBER package (Weiner et al., 1984). The ligand files were cleaned and if multiple ligands at the same position were presented, only one was held. Unbound structures were superimposed on their bound complement so that the ligand could be placed in the binding site for statistical calculation and visualizations. All calculations were performed in the absence of the ligand, while the ligands are shown in the figures for illustration purposes.

The prediction consists of two steps: in the first step the cavity detection algorithm places probes on the surface of a protein's cavities. The second step includes the calculation of the cost to place one of these probes onto the surface in the presence of water. To calculate this penalty, the generalized Born (GB) model included in the Sander program from the AMBER 11 package was used (Hendlich et al., 1997).

4.1.2 Statistical measurements

Chen et al. (2011) summarized three different measurements for the quality of small ligand binding site predictions. One option is the measurement of the distance of the center of a predicted cavity to any atom of the ligand with a certain cut off, named D_{CA}. This is a fundamental measurement, since many predictors result in the prediction of a complete cavity or directly in a single point in the center of the cavity and is therefore widely used. D_{CC} defines the distance between the center of the prediction and the center of the ligand and reveals whether a prediction is more in the periphery of the real binding site or near the center of the ligand. A third measure is called O_{PL} and tries to identify the accuracy of the prediction by calculating the overlap of the prediction with the ligand atoms and vice versa. This is only possible if the predictor offers more information than just the center of the cavity.

To evaluate the method presented here, D_{CA} and D_{CC} values were calculated while for O_{PL} a different measurement was chosen. The overlap of ligands with predictions and predictions with ligands were calculated separately, to directly account for sensitivity and specificity. A ligand heavy atom (a) is counted as covered, if the distance to a probe (p) is smaller than the sum of the ligand atoms van der Waals radius, plus the radius of the probe (see equation 4.1).

$$f(a) = \begin{cases} 1, & \text{if } dist(a,p) \leq (a_{vdw} + p_r)\ \forall p \text{ in the prediction} \\ 0, & \text{else} \end{cases} \tag{4.1}$$

These values indicate the ability of the method to identify the correct position of the ligand atoms and can be perceived as sensitivity when accounting

for all heavy atoms of the ligand (see equation 4.2):

$$O_{sens} = \frac{\sum_{i=1}^{n} f(a_i)}{n}, \text{ with n the number of ligand atoms} \tag{4.2}$$

Analog a probe is considered as a correctly predicting probe if - within the same distance as above - a ligand heavy atom can be found, giving the specificity of the prediction method (equations 4.3 and 4.4):

$$f(p) = \begin{cases} 1, & \text{if } dist(a,p) \leq (a_{vdw} + p_r)\ \forall a \text{ of the ligand(s)} \\ 0, & \text{else} \end{cases} \tag{4.3}$$

$$O_{spec} = \frac{\sum_{i=1}^{n} f(p_i)}{n}, \text{ with n the number of probes} \tag{4.4}$$

The number of ligand atoms is defined by the number of heavy atoms belonging to one or more ligand/s found in the crystal structure of the bound complex, whereas the number of predicting probes is restricted to the method explained in the next chapter and is usually the sum of all probes found in the top1 or top3 cavities.

4.1.3 Evaluation of prediction quality

Two different criteria are chosen to evaluate the performance of the method: top3 indicates that the three best ranked cavities are used for the prediction. Top1 indicates that as many top ranked cavities are taken into account as ligands are found bound to the protein. This was recommended by Chen et al. (2011) and also describes a more feasible treatment of predictions than e.g. counting the prediction of one of several binding sites with 100%. To calculate the specificity and sensitivity of the method, only those probes are taken into account that are in the actual prediction (top3 cavities or top1 cavities). The specificity and sensitivity of all probes generated for prediction are shown separately. The method presented in this chapter does not generate a distinct number of predictions, which means that cavities are not marked as possible binding sites if they do not fulfill certain criteria, just to present a number of cavities demanded. Therefore, it is possible that less than three cavities are in

the output of the predictor. In this case, all remaining cavities are taken into account for the top3 statistics.

In contrast to other predictors such as LIGSITE, Roll, etc. the method presented here does not fill the entire cavity with probes (as those methods do with grid points), but covers only the surface. As a consequence, not all ligand atoms can be overlapped, even if a probe is placed at every possible position on the protein's surface. Another implication is that the estimated center of the cavity might be shifted more towards the protein's surface, depending on the curvature of the cavity, reducing the performance while using above mentioned statistics, compared to methods trying to propose the center of the cavity.

4.2 Geometry-based cavity detection

Cavity detection using geometrical criteria is a widely used tool to identify cavities on a protein's surface and to estimate its size. Cavities are often ranked according to the size of the identified cavities, since studies have shown that the ligand can often be found in the largest cavity (Campbell et al., 2003). Besides the size of a complete cavity, which might be larger than the binding ligand, methods attempt to find additional scoring methods (e.g. by ranking according to a depth value (Yu et al., 2010)) or better sampling of possible interaction regions (e.g. the identification of subpockets (Volkamer et al., 2010)). The presented method follows the assumption that the ligand is found within the largest cavities.

4.2.1 Principles of the cavity detection algorithm

The cavity detection procedure starts by rolling a small probe over the van der Waals surface of each protein atom. The position is kept, if no collision with other atoms is detected and if it was not closer than half its radius to another already placed probe. Similar to many other cavity detection algorithms, a larger probe is used to discriminate between probes placed in and outside of cavities by discarding intercepting small probes. Therefore, the large probes as well as the small probes are first placed on the protein's surface. Depending on the size of the probes, small cavities can be detected (small radius for the small probe, large radius for the large probe) as well as been overlooked (large radius for the small probe, small radius for the large probe). The choice of

the probe size not only influences the ability of the algorithm to cover every possible ligand in this initial step, but also accounts for the computational demand for further energy calculations. Nevertheless, further calculations are fruitless if the probes do not cover the important cavities.

Subsequently, if a probe is farther away than its radius to any other probe, single probes, e.g. probes in the interior of the protein, are deleted. Resulting probes are clustered roughly to generate continuous patterns of probes on the protein's surface – which is illustrated schematically in Figure 4.1A and demonstrated on a sample protein in Figure 4.1B. As a distance discriminator the radius of the probe is taken (this value can be changed on demand to values between 1.0 and 2.0 times the radius, so that probes within one patch at least touch one of the other probes in the patch). Clusters with a size lower than an expectable ligand are discarded, whereas the default value of the expected size can be substituted by a parameter given to the algorithm.

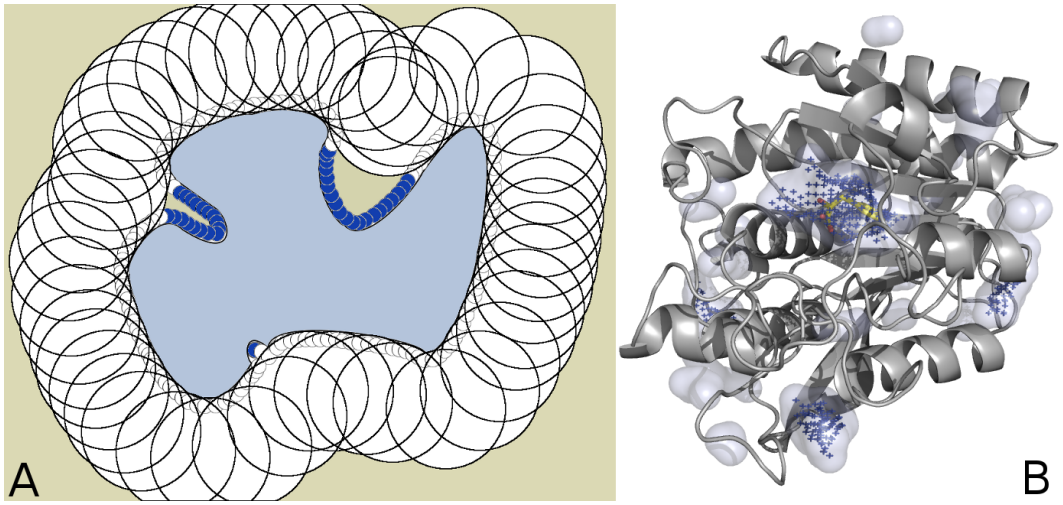


Figure 4.1 – A) Schematic representation of the geometric cavity detection method. Small probes are placed on the protein's surface as well as large probes with a larger radius. All small probes intercepting with the large probes are deselected (white small probes) and the rest is kept (blue small probes). B) Showing the three different stages of the algorithm around a protein (pdb2ctb). Raw data from the cavity detection procedure in A) is shown as transparent grey surface area, blue crosses show the resulting probes after clustering. Ligand is shown in sticks.

4.2.2 Performance of initial cavity detection

The presented cavity detection method has the same restrictions as most other geometry-based approaches and is sensitive to the orientation of the protein in space as well as to the chosen settings. Basic parameters are in this case the size of the small probes as well as the size of the large probes. While for

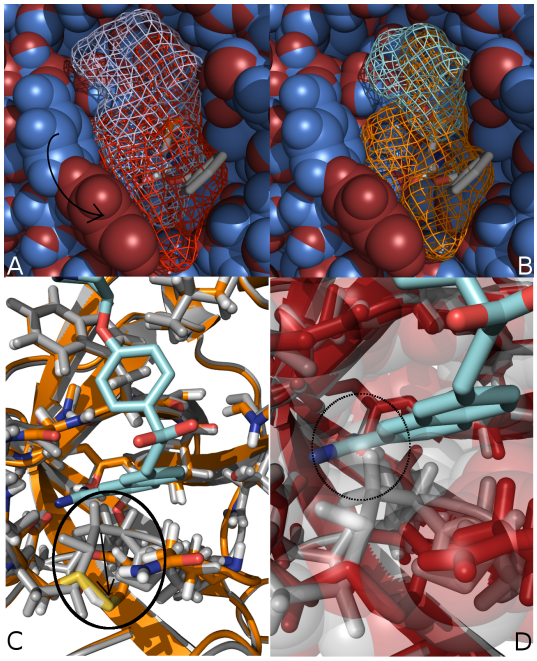


Figure 4.2 – A) The bound structure pdb1blh shown in red spheres with the initial probe placement shown in red mesh. Blue spheres and mesh represent the unbound structure pdb1djb and the prediction, respectively. The ligand is shown in stick representation. The movement of one residue closes the pocket upon binding with the ligand and allows a larger overlap of the real binding site for the placed probes. B) After desolvation free energy calculation and further selection and clustering the prediction for the unbound structure was reduced (mint mesh) while the prediction for the bound was kept (orange mesh). C) and D) In the unbound structure (shown in grey, pdb2tga) the backbone at one residue position is flipped inside, completely closing the binding site (see D) grey transparent spheres overlap with the ligands atoms). In the bound conformation (pdb1mtw) the residue is flipped out as shown in yellow/orange cartoon and sticks in C) and red transparent spheres in D), widen the binding site and offer enough free space for ligand binding.

this study, and the use as settle points for further energy calculations, the size of the small probe was fixed to 1.4 Å (which is the approximated radius of a water molecule), but the size of the large spheres can vary. Small radii for the large probes (like 4.0 Å) result in a very low but precisely described number of cavities but does not cover parts of the known ligands at more exposed surface areas. Larger spheres (e.g. 6.0 Å) allow probes to overlap with nearly all ligand atoms, but come with increasing demand for desolvation free energy calculations.

Table 4.1 shows the impact of varying probe size on the results. Larger large probes enable the sampling of more spots on the protein's surface resulting in a higher number of cavities as well as more probes per cavity. Taking all cavities into account, the larger probes show better performance for sensitivity values, which therefore indicates a large fraction of ligand atoms overlapped with probes, at the cost of an increasing total number of probes. As a starting

point for further calculations, a high sensitivity is the prerequisite for the identification of all possible interaction points, however a large number of probes increases computational time.

The geometry-based approach performed differently well for bound and unbound protein structures. While for bound complexes cavities are already formed, unbound structures sometimes differ significantly from the bound form (see Figure 4.2). Either cavities are too narrow to be detected with the small probes, or they are opened widely (e.g side-chains are rotated, opening the cavity unfavorable) and large probes with large radii are necessary not to eliminate the small probes. The performance on the bound set can be reduced when closing of the binding site of a small ligand upon binding reduces the free accessible volume to a minimum, and is discarded due to the reduced size of the detected cavity. The minimum size describing a cavity is set in this study to a volume of a minimum of 200 $Å^3$ per cavity, which is at the lower range found for ligand binding sites (100 to 1000 $Å^3$ (Liang et al., 1998)).

For further energy calculations the size of the small probes was set to 1.4 Å (as reported above to approximate the size of a water molecule) and for the large probes a size of 4.5 Å was chosen. This combination showed good results in case of the widely used DCA values for top1 and top3 predictions, on the small test set used to optimize the method, with an acceptable amount of probes to be calculated.

Compared to other predictors in the field which have been tested on the same test set, the geometry-based cavity detection can be found among the best performing methods. Yu et al. (2010) collected results reported by the authors of different binding site prediction methods on this test set (listed in Table 4.2). According to the ordering used by Yu et al. (2010), this method is second best for the bound test set. For the unbound structures the performance is slightly worse, resulting in third place for the top1 predictions and fourth for the top3, respectively.

Table 4.1 – Prediction statistics for geometry-based binding site prediction

unbound		overlap	sensitivity	specificity	DCA	DCC	nr	probes
top1	BP 4.5	72.92	66.91	47.92	61.11	46.53		
	BP 5.0	73.96	69.45	45.30	61.11	41.32		
	BP 5.5	78.12	73.77	44.47	62.15	41.32		
	BP 6.0	75.00	72.54	43.10	58.33	45.29		
top3	BP 4.5	87.50	80.80	38.14	75.69	59.03		
	BP 5.0	90.62	84.99	36.04	75.69	53.82		
	BP 5.5	94.79	89.05	34.43	76.74	53.82		
	BP 6.0	94.57	89.99	33.72	73.55	56.16		
All	BP 4.5	93.75	85.06	33.77	81.94	63.19	4.66	119.58
	BP 5.0	93.75	87.33	31.13	79.86	55.90	5.35	130.18
	BP 5.5	97.92	92.51	29.35	80.90	57.99	5.94	139.86
	BP 6.0	97.83	94.62	28.79	77.90	60.51	6.30	135.80
bound								
top1	BP 4.5	82.29	79.65	61.38	75.00	57.29		
	BP 5.0	80.21	77.73	54.78	75.00	51.04		
	BP 5.5	77.08	75.61	48.21	65.62	47.92		
	BP 6.0	76.04	74.66	45.07	60.42	44.79		
top3	BP 4.5	94.79	90.90	46.27	87.50	69.79		
	BP 5.0	92.71	90.02	41.68	87.50	63.54		
	BP 5.5	95.83	93.16	38.64	84.38	64.58		
	BP 6.0	91.67	89.41	35.23	76.04	58.33		
All	BP 4.5	95.83	91.73	41.74	88.54	70.83	4.54	123.87
	BP 5.0	93.75	90.85	36.04	88.54	64.58	5.33	128.27
	BP 5.5	97.92	95.24	32.92	86.46	66.67	6.21	131.52
	BP 6.0	95.83	93.57	29.26	80.21	62.50	7.02	132.01

Statistics for geometry-based cavity detection for the test set of 48 unbound (upper part) and 48 bound (lower part) structures for the *top1* and *top3* prediction as well as for all cavities. BP indicates the size of the large probe in Å. Other parameters are kept the same for all statistics. Overlap gives the percentage of ligands for which at least one probe of the prediction (within *top1* cavity or *top3* cavities or *All* cavities) fulfills the criteria given for sensitivity (see chapter 4.1.2). Sensitivity, Specificity, DCA and DCC values are given in percent and calculated following procedures in chapter 4.1.2. Additionally, for *All* cavities the average number of cavities as well as the number of probes per cavity are given.

Table 4.2 – Results for the 48 bound and 48 unbound structures

Method	top1		top3	
	Unbound	Bound	Unbound	Bound
POCASA	75	77	88	94
dPred^{geo}	61	75	76	88
PocketPicker	69	72	85	85
LIGSITE^{cs}	60	69	77	87
LIGSITE	58	69	75	87
CAST	58	69	75	83
PASS	60	63	71	81
SURFNET	52	54	75	78
Q-SiteFinder	51	80	86	97

Data shown in this table is taken from Yu et al. (2010). All methods are ranked according to the ranking in the original paper by the bound top1 value. dPred^{geo} is the geometry-based binding site detection method presented in this chapter. For details of statistical calculations see chapter 4.1.2 and for possible differences in top1 definition see chapter 4.1.3. Values for Q-SiteFinder are taken from Laurie and Jackson (2005) for a subset of the presented test set. Also, the declaration of a successful prediction is different from the other methods. Results are given with reservation therefore.

4.3 Energy calculation procedure and statistics

Areas on a protein's surface that are known to bind small ligands are often found to be cavities with a more hydrophobic character and less polar interactions. Since the geometry of cavities reduces possible entry areas, the necessary unbinding of bound water molecules upon ligand association describes a large energy barrier which has to be overcome. To determine the cost for replacing water molecules, small neutral probes are placed in the protein's cavities, representing a position where a ligand atom might be in contact with the surface. Therefore, the generalized Born (GB) approach of the Sander program was used to calculate the energy difference within an implicit water model.

4.3.1 Calculation of desolvation properties in cavities

The calculation of desolvation penalties in order to determine the hydrophobic core of protein-protein interaction sites using the finite difference Poisson Boltzmann equation had previously shown success on a variety of complexes (Fiorucci and Zacharias, 2010a). Probes were placed at a 3 Å distance to each other on the complete surface of the protein. In the approach presented here, more probes are placed per Å^2 of surface area and are afterwards treated independently, while in the approach of Fiorucci and Zacharias (2010a) probes

within 10 Å are averaged to generate smooth transition from favorable to non-favorable areas, reducing the influence of locally high penalties. This may be beneficial for the investigation of large surface areas that become buried upon protein-protein complex formation, but is not suitable if the number of contacts is small as it is for protein-ligand binding.

To estimate the change in solvation free energy upon binding of a small neutral probe, a reference energy is calculated for the protein alone. Each probe is then placed separately on the surface and the change in energy is calculated as

$$\Delta EGB_i = EGB_i - EGB_{ref} \tag{4.5}$$

which is an approximation of the energy required to displace water at the given position, where EGB_{ref} is the reference energy of the isolated protein and EGB_i is the energy of the protein plus the probe i. The difference is stored on the probe and represents the penalty of placing a neutral ligand atom at the position of the probe.

4.3.2 Desolvation free energy statistics and generalized Born settings

Initial optimization on a small test set of bound and unbound, as well as homology modeled, protein structures (see Leis et al. (2010) for the list of proteins) revealed the best parameters: a small probe size of 1.4 Å was used to mimic a water molecule, while for the large probes a radius of 4.5 Å showed the best compromise between sensitivity and number of probes. Smaller large probes (3.5 Å or 4.0 Å) resulted in a very good performance for DCA values but had a low sensitivity and vice versa for bigger large probes (a range of 3.0 Å to 7.0 Å was tested, data not shown, results for large probes on the evaluation test set for varying large probes between 4.5 Å and 6.0 Å are shown in Table 4.1). Variations of small probes were also tested, showing better DCA results with a probe radius of 1.2 Å to 1.3 Å, as well as good sensitivity. Despite better results for the pure geometry-based approach, for desolvation penalty calculations only the small probe size of 1.4 Å was used.

To calculate the desolvation free energies with the generalized Born model, the Sander program of the AMBER11 suite was used with the OBC method (setting igb=5, based on Onufriev et al. (2000, 2002)). The electric constant for the outer region was set to $\epsilon = 80.0$, while for the interior of the protein $\epsilon = 4.0$ was used, which is related to the dielectric constant of the buried regions of the protein (Schutz and Warshel, 2001). Other parameters were tested: the default for the GB calculation of $\epsilon = 1.0$ (also used for hot spot detection of transient small ligand binding pockets in protein-protein binding sites by Metz et al. (2011)) was tested as well as $\epsilon = 10.0$, which was used for desolvation calculation of protein-protein binding sites (Fiorucci and Zacharias, 2010a). The ΔEGB values for the different settings are given in Figure 4.3C, showing differences in magnitude and amplitude: the general distribution of values within the probes differs only slightly, resulting in approximately the same predictions with adjusted selection criteria, while the values for the penalty significantly differ. This is also reflected in the prediction performance, differing only slightly from the ones obtained using $\epsilon = 4.0$.

Since enlargement of the small probe lead to a significant drop in the performance of the cavity detection, larger small probes could only be tested by resizing probes only for GB calculations, resulting in overlaps with the protein. As an outcome, energy differences enlarge, but do not change the overall picture dramatically (distribution is shown in Figure 4.3D). Following the procedure developed for 1.4 Å probes, DCA and DCC values decrease slightly which is mostly affected by single probes with extremely large penalties for 2.8 Å probe radius. In fact increasing the radius slightly to 2.1 Å can help to distinguish between different values of low penalties more effectively, but was not tested systematically however in this study.

The desolvation penalty for probes near the ligands binding site and probes in other areas differ slightly (see Figure 4.3A, see also for definition of ligand contact area), but a significant larger amount of probes with a low desolvation penalty could be found at probes near the known ligand. Also, the number of probes associated with a high penalty is slightly larger within cavities where ligands are found, indicating possible polar contacts between the ligand and the protein. For probes outside the ligands binding site, a higher amount compared to binding site probes could be found beyond the maximum (Figure 4.3A). Figure 4.3B shows that the difference between the bound and unbound test set concerning the ΔEGB distribution is negligible.

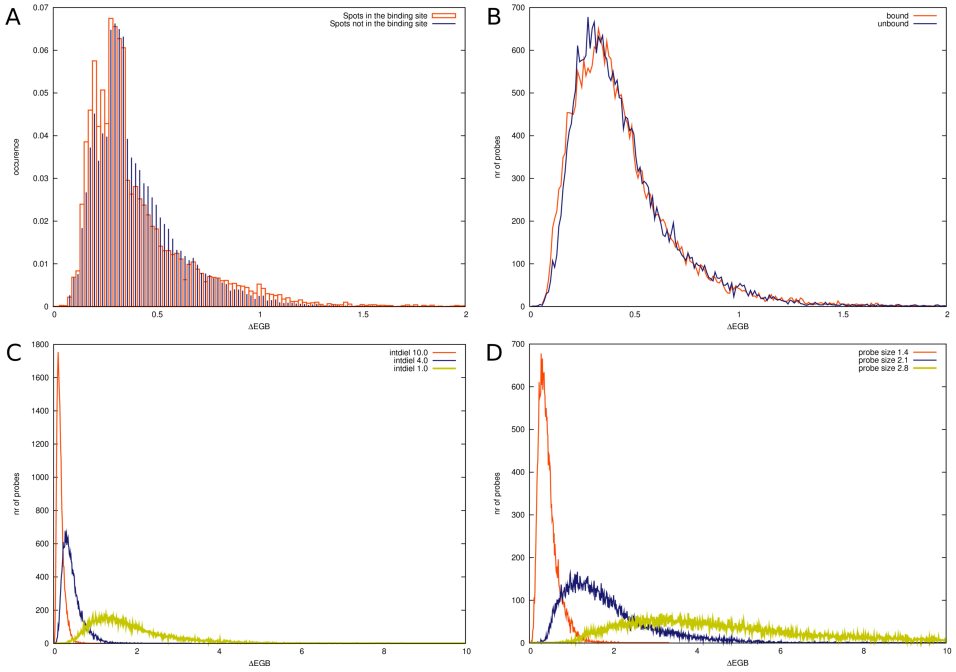


Figure 4.3 – A) Shows the distribution of desolvation penalties for the unbound test set (bound similar, data not shown). Desolvation free energy for probes near the known ligand (within a maximum distance of 1Å from the vdW surface of the ligand heavy atoms to the probes surface, resulting in probes with at least touching vdW and probe radii) in orange and for other probes in blue. B) Distribution of the desolvation penalties in the bound (orange) and unbound (blue) test set. C) Showing the differences in the distribution for different internal dielectric constants of the interior of the proteins. D) Varying probe sizes during desolvation free energy calculation. All probes had been placed at 1.4Å in respect to the vdW surface of the protein. Larger probe sizes result in an overlap of the probe with the protein during energy calculations. The internal dielectric constant of 4.0 was used.

4.4 Predicting binding site geometries using desolvation penalties

The distribution shown in Figure 4.3A indicates, that ligands are more often found in cavities containing probes with low and high desolvation penalties. Therefore, the filtering of probes for cavity selection and deselection is based on clusters of probes with high and low ΔEGB, representing regions of high hydrophobicity as well as stabilizing polar contacts (illustration of the algorithm in a schematic view can be found in Figure 4.4).

4.4.1 Clustering of desolvation penalties

If within the result list only one cavity is given, a basic correction function is called, deleting single probes in the rim region of the prediction as well as narrow regions within the cluster. The latter is done to partition a pocket

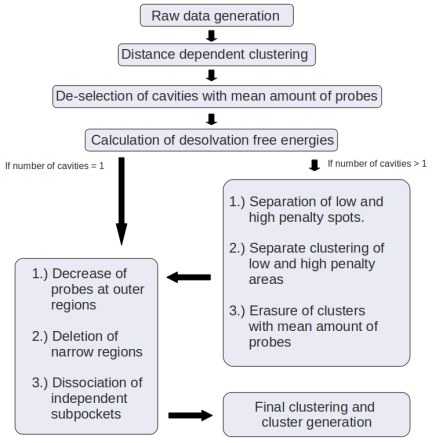


Figure 4.4 – Sequence of the algorithm: The raw data is generated by placing small probes on the surface and discarding those intercepting with large probes on the surface. Then single probes and small clusters are discarded and desolvation free energies are calculated for the remaining probes. If only one cavity is given, narrow regions are deleted and probes at the rim region are trimmed. Otherwise, if more cavities are present, high and low desolvation penalties are separated from each other followed by clustering of the probes in the two probe sets. Clusters that are large enough, are kept and finally clustered together to form binding pockets or subpockets.

into subpockets by deleting contacts between independent regions connected via thin accumulations of probes, if possible. A probe is deleted if it has less neighbors as requested by a threshold value and less than half of the neighbours of this probe have less neighbors as requested by themselves.

The procedure is augmented for those predictions containing more than one cavity: the probes are separated into two lines, one containing the lower 70% of all probes and the other containing the 30% with the highest desolvation penalties. Both lines are clustered separately resulting in a neglecting of small high or small low penalty clusters. This discards single extreme values within a homogeneous environment on one hand and on the other hand, completely refuses pockets interspersed with varying penalty values. Since the border between low and high desolvation penalty is artificially generated from the average distribution of penalties in this particular test set, it might not suit perfectly for a specific protein but works well in the general case. The output is ranked by the size of the clusters but energy values are given additionally, so that a compartmentalization and re-ranking based on the penalties is possible.

Figure 4.5B illustrates the procedure of clustering of single cavity predictions: As a first result of the raw probe placement and deletion, the initial probes are set (grey surface representation) and after clustering reduced to one large cluster (blue crosses, partially overlapped by small red spheres). After desolvation penalty calculation and further clustering, the number of probes

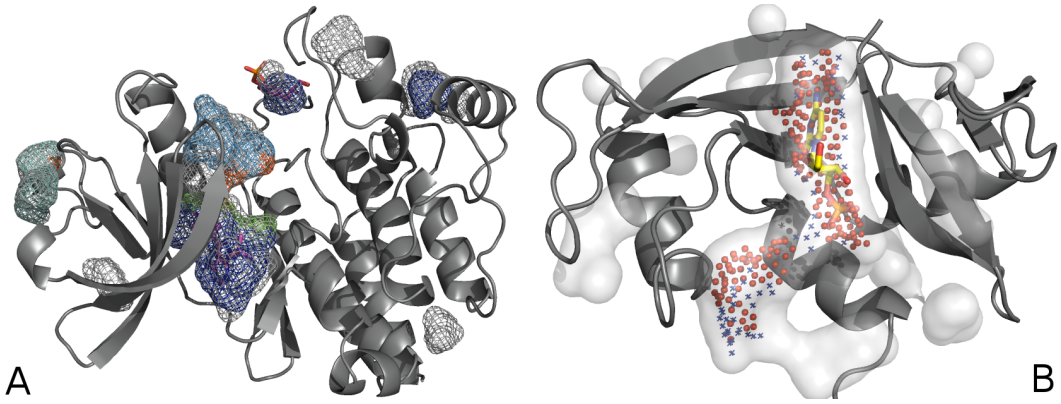


Figure 4.5 – A) Resulting clusters and cavities for pdb1qpe. Grey mesh indicates all cavities found after cavity detection using geometric criteria and clustering and colored mesh those left after desolvation free energy calculations and clustering. B) Shows the three different stages of the algorithm around a protein (pdb8rat). Raw data from the cavity detection procedure (see Figure 4.1a) is shown as transparent grey surface area, blue crosses show the resulting probes after clustering. Red spots mark the remaining probes after desolvation calculation and further clustering. Ligand is shown in sticks.

(blue crosses) is reduced at the edges and a narrow region in the middle of the two subpockets deleted, revealing that the large cavity can be divided into one larger net of spheres encapsulating the ligand, and one smaller cluster beside the known binding site (shown in red small spheres). For predictions with more than one cavity, the previously defined procedure, using separation of desolvation penalties, is used. As the example in Figure 4.5A shows, all cavities are taken into account (see Figure 4.5A, grey mesh representation) for desolvation free energy calculations and after clustering and removal of improper cavities, the number of cavities reduce from seven to five. The largest cavity found in the first stage is thereby divided into two – now independent – subpockets, with the larger of the remaining two pockets enclosing the known ligand binding site. The second ligand binding site is also detected, but enclosed in the fourth largest cluster, marked as failed in top1 and top3 statistics.

4.4.2 Performance of the binding site prediction

The initial approach highlighted, that it is in principle possible to at least overlap with most of the ligands (for the unbound test set 93.75% of the ligands were overlapped by at least one probe, 95.83% for the bound test set, respectively, values are shown in Table 4.2 with settings 4.5 Å large probes and all cavities). Sensitivity for the unbound structures with a value of 85% and 91.73% for the bound ones indicates, that in most cases the overlap is

Table 4.3 – Prediction statistics for dPred^{GB}

		overlap	sensitivity	specificity	DCA	DCC	nr	probes
unbound	*top1*	72.92	66.82	41.97	69.44	50.69		
	top3	87.50	78.96	41.97	81.94	59.03		
	All	91.67	82.10	39.65	86.03	61.11	3.69	116.80
bound	*top1*	86.46	82.65	68.55	85.42	64.58		
	top3	94.79	89.95	51.11	93.75	72.92		
	All	95.83	90.78	48.10	94.79	73.96	3.69	119.69

Statistics for energy based docking for the test set of 48 unbound (upper part) and 48 bound (lower part) structures for the *top1* and *top3* prediction as well as for all cavities. Sensitivity, Specificity, DCA and DCC values are given in percent and calculated following procedures in chapter 4.1.2. For *All* cavities additionally the average number of cavities as well as the number of probes per cavity are given.

not only partial but also dominant. Values for the top1 and top3 predictions denote, that the ligands can not always be found in the largest or among the largest cavities. Nevertheless, taking all cavities into account illustrates that a rescoring of cavities or a resampling of probes can be successful.

Reducing the amount of possible probes, and therewith the size of a cavity as well as its relative center, resulted in a reordering of the geometrically detected cavities as well as the removal of irrelevant ones. For the unbound test set, the number of detected cavities on average is reduced from 4.66 (with 119.58 probes per cavity) to 3.69 cavities (with 116.80 probes per cavity). Using this procedure, more than 20% of the probes can be discarded successfully, resulting in an increase of specificity from 38.14% to 41.97% for the top3 predictions while the sensitivity decreases slightly (compare Table 4.3). Taking all cavities into consideration, the sensitivity decreases by 2.96% while the specificity increases by 5.88%. The same is true for the bound test set, which performs comparably: specificity values increase for top1 (7.17%) and top3 (4.84%) predictions. Sensitivity for the top3 predictions decreases slightly (0.95%), but increases for the top1 predictions (3%), due to a change in the ranking of the cavities. Half the probes within the top3 predictions are overlapping with a ligand atom and almost nearly 70% in the top1. Since this includes all probes from successful as well as from the unsuccessful predictions, this value indicates that for all true predictions most of the probes are actually marking protein-ligand contact areas. On the other hand 82.65% of the ligand atoms were

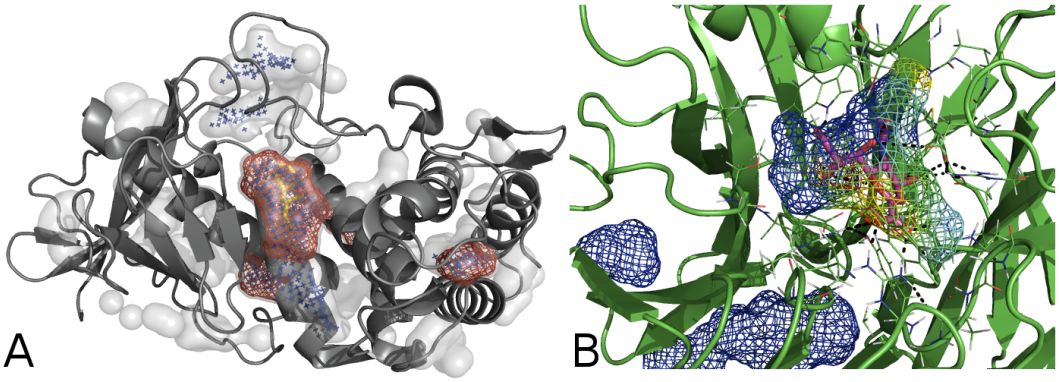


Figure 4.6 – Two structures where the ligand could not be found with the ROLL algorithm (Yu et al., 2010). A) Initial (grey raw surface points, blue geometry-based clustering) and final prediction (red mesh) for pdb1npc (unbound test set) resulting in three cavities with one larger cavity enclosing the ligand (yellow sticks) and three empty smaller cavities. Before reduction, due to desolvation penalty calculations, two additional cavities existed and the known binding site was nearly twice as large as the final cavity. B) Shown are all resulting predictions for pdb2sim (bound test set) in mesh representation. The mesh is color coded from low penalties (blue) to high penalties (red). The ligand (shown in stick representation) is found in the largest predicted cavity with only slight overprediction.

Table 4.4 – Results for the 48 structure test set

Method	top1		top3	
	Unbound	Bound	Unbound	Bound
dPred^{GB}	69	85	82	94
POCASA	75	77	88	94
PocketPicker	69	72	85	85
LIGSITE^{cs}	60	69	77	87
Q-SiteFinder	51	80	86	97

See Table 4.2 for details. dPred^{GB} is the method presented in this chapter using geometry-based cavity detection and subsequent desolvation free energy calculations.

in touch with probes (top1, 89.95% for top3, respectively), including missed ligands, these values point out, that if a ligand is in a predicted cavity, mostly all its contacts to the protein's surface are near probes.

The binding site prediction performance itself improves as well: for the unbound test set the top1 and top3 prediction rate gain 6% to 8% (DCA values) compared to the pure geometry-based approach (see Table 4.1). Within the top3 predictions, 82% of the ligands are within 4 Å from the prediction center (DCA value) and 69% of the ligands can be found in the top1 prediction sites. For the bound test set, top1 results increased by 10% (75% to 85%) and the top3 by 6% to 93.75%. The protocol, used here for the prediction of small ligand binding sites, can be found among the best performing methods. The

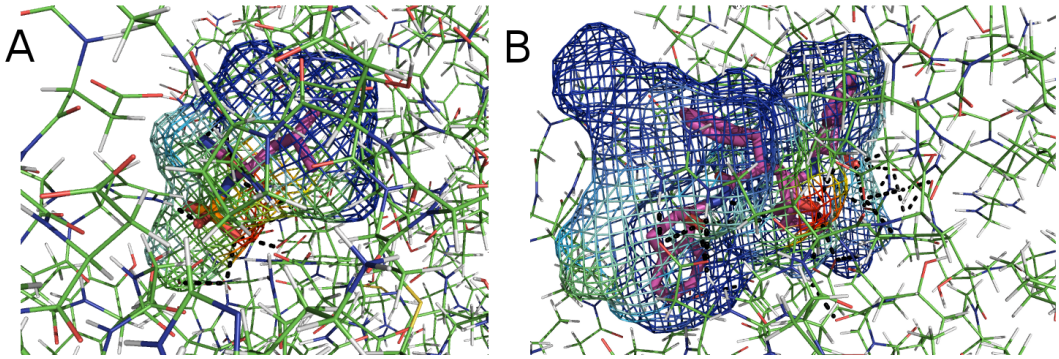


Figure 4.7 – Possible polar contacts between the ligand and the protein shown as black dashed lines. Prediction is shown in mesh representation and with color codes for low desolvation penalties (bluest) and high desolvation penalties (yellow to red). Ligands are shown as sticks. A) shows results for pdb2tmn and B) for pdb4phv.

top3 value is comparable to the best performing method (see Table 4.4 and Figure 4.6 for examples not found with the ROLL predictor), while the top1 prediction performed prominently, compared to all other methods.

The difference in the local structure of bound and unbound proteins, indicated by Figure 4.2, is also reflected in the prediction values: the geometric probe placement presented in this method is not always able to identify unbound protein ligand binding sites perfectly. Therefore, the performance for unbound structures does not outperform other binding site predictors, but is among the methods with the best results. The general procedure of using smaller probes to also define all known binding sites in the unbound form geometrically, is hardly applicable and would result in an inefficient sampling of desolvation energies and therefore, in less distinguishable penalties. Other methods using grid based approaches (like LIGSITE, ROLL, etc.) define grid spacing down to 0.5 Å, also enable the detection of narrowed binding sites.

Besides the prediction of the binding area, the method presented here also attempts to predict polar contacts between the protein and the ligand. High desolvation penalty regions within, or at the border of larger low penalty areas, often give a hint of whether a possible polar contact can be established in this region. Figure 4.7 shows the prediction for pdb2tmn (Figure 4.7A) and pdb4phv (Figure 4.7B). For pdb4phv most polar contacts are within a high area of desolvation penalty (1.7 kJ/mol) while others are in a lower area (0.5 to 1.0 kJ/mol). For probes in regions where no polar contact of the ligand and the protein can occur, values between 0.1 and 0.3 kJ/mol are proposed, fulfilling the expectations. Also for pdb2tmn high values of desolvation penalty

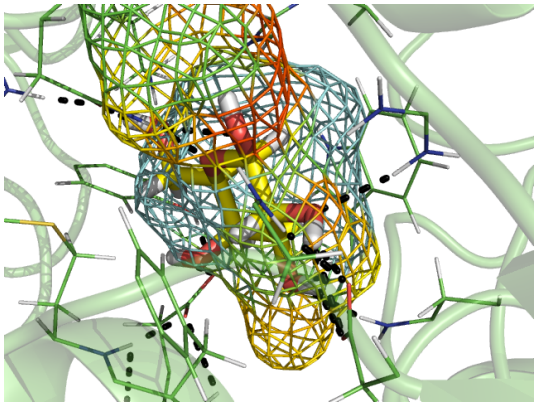


Figure 4.8 – The prediction covering the known ligand binding site for pdb1gcg does not show low desolvation penalties. All possible polar contacts are within high desolvation free energy difference regions.

can be found in regions where polar contacts may occur (up to 1.4 kJ/mol) while for regions slightly further away placement of probes is less penalized (0.2 kJ/mol).

Besides predictions with a mixed state of desolvation penalties, cavities are among the predictions that do not include low desolvation penalty spots. For pdb1gcg the lowest spot in the correct prediction has a value of 0.5 kJ/mol covering the ring of the ligand. High penalties are accounted for those probes covering ligand atoms that can introduce polar contacts with protein atoms (1.0 kJ/mol for the yellow lower spot and 1.4 kJ/mol for the red spot as shown in Figure 4.8).

On the other hand, the two additional cavities predicted had a homogeneous distribution of desolvation penalties of 0.33 kJ/mol for one and 0.4 kJ/mol for the other predicted binding site. Compared to other results, 0.33 kJ/mol is at the upper border of lowest penalties sampled for this test set (see Figure 4.9A). In each case, the difference between the lowest and highest desolvation penalty is in the mean of the test set (compare Figure 4.9B), although the lowest values are higher than on average.

A big difference between lowest and highest desolvation free energy difference is only found in single predictions. On average, the differences are within 0.75 to 2.0 kJ/mol and high values of desolvation free energy hardly exceed 3.0 kJ/mol. These values are noticeably lower than the values found scanning the protein surface by Fiorucci and Zacharias (2010a). One reason for this fact might be the size of the probes (2.0 Å for scanning the protein surface vs. 1.4 Å for scanning the cavities), another the difference in the test set and the

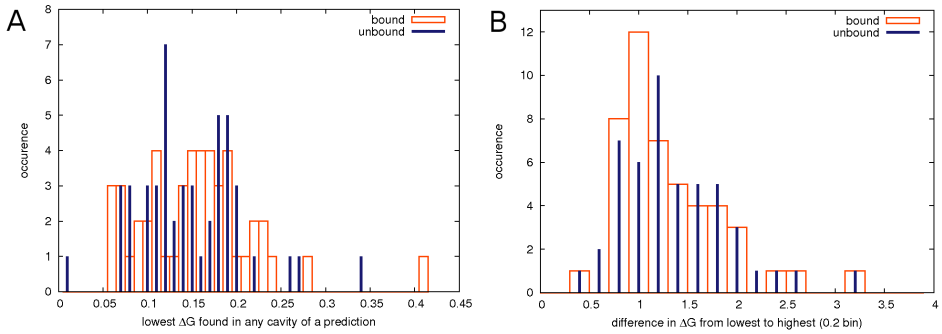


Figure 4.9 – A) Occurrence of lowest penalties found in any cavity of one prediction for the bound and unbound test set. B) Difference between the lowest and highest value within one prediction (taking all cavities in one prediction into account).

search area. Since probes at the more polar surface are omitted during this procedure, the high amount of hydrophobic contacts within the cavities might lower the overall penalty.

4.5 Conclusion

The field of small ligand binding site detection is dominated by approaches describing cavities geometrically, while energy based methods are often disregarded because of their computational demand (Leis et al., 2010, Chen et al., 2011). Fast methods based on comparison of sequences or known binding sites appeared, demonstrating good performance, but are restricted to cases were a large knowledge base exists. More sophisticated methods, using Molecular Dynamics simulations or fragment docking, can often give more insight into binding affinities or the druggability of a cavity (Seco et al., 2009, Kozakov et al., 2011), but come with a high computational demand.

In this chapter a method was presented, based on the calculation of penalties arising from occupying solvent accessible surface areas with a neutral probe. The general procedure has shown some significance for the detection of hydrophobic patches within protein-protein binding sites (Fiorucci and Zacharias, 2010a), but has not yet been applied to small ligand binding sites. Since solving the Poisson-Boltzman equation for thousands of probes on the protein's surface requires a significant amount of computational power, the presented method achieves a remarkable acceleration for the calculation of favorable protein-ligand binding sites, due to the usage of the generalized Born model.

The fast and accurate geometrical cavity detection protocol results in a significant reduction of meaningful probes, for which time consuming energy calculations must be performed. This assumption is based on the observation, that small ligands are often found in shallow areas of the protein's surface (Miller and Dill, 1997, Liang et al., 1998) and enables the breaking down of GB calculations to some hundred per protein. For small proteins with a few hundred probes, prediction can be finished within seconds or take up to a few minutes.

Binding site prediction statistics show, that calculation of desolvation penalties is a valuable method to detect small ligand binding sites and performs well amongst the best methods. Similar to more complex methods, this method can identify hydrophobic, as well as polar binding sites. This makes it valuable for the detection of binding poses, since in many cases the approximated orientation of the ligand in the cavity can be estimated by the distribution of polar and non-polar contacts. Many of the predictions show a good overlap of placed probes and known ligand atoms, indicating that not only binding site cavities can be distinguished from non binding site cavities, but also relevant from non-relevant areas within a cavity.

One drawback is the lower performance on unbound structures compared to bound structures. Possible enhancement could be the introduction of flexibility into the probe placing algorithm or a differing probe placement, e.g. an adjusted grid based approach. Nevertheless, for bound structures, the sampling as well as the scoring of the binding cavities showed a good performance for the identification of relevant binding sites, as the top1 prediction values show. Additional information, e.g. sequence conservation or statistical evaluation, could help to further discard cavities which do not bind small ligands and therefore increase the number of top1 predictions.

Chapter 5

Optimizing unbound protein-protein docking

The realistic prediction of protein-protein complex structures (protein-protein docking) is of major importance as only a small fraction of real and putative protein-protein interactions in a cell can be experimentally determined. Some interactions are only transient or weak such that experimental determination of a complex structure is difficult or sometimes even impossible. Computational protein-protein docking methods are becoming of increasing importance in order to generate, at least, model structures of possible protein-protein complexes. Protein-protein docking approaches can also be helpful to evaluate newly designed protein-protein interactions which are of increasing interest in the area of biotechnology and medicinal chemistry. Several classes of docking methods have been developed (reviewed in Pons et al. 2010, Moreira et al. 2010, Zacharias 2010a; Schneider & Zacharias, 2010, see Chapter 2.3).

These methods systematically search for putative protein-protein binding geometries using various surface matching approaches or employ force field based energy minimization or related optimization procedures. As a first step, rigid docking is performed typically to generate a high amount of possible binding poses. Afterwards, a rescoring of the rigid solutions, followed by one or more refinement steps, often including some flexibility, is applied to generate complexes in closer agreement with the native geometry in a prominent position on the result list (Pierce and Weng, 2007, 2008, Pons et al., 2010, Zacharias, 2010b). It is also possible to re-evaluate the docked complexes according to available experimental data on a putative binding surface or based

on predictions from bioinformatics binding site prediction approaches (Ben-Zeev and Eisenstein, 2003, Gottschalk et al., 2004, Zhang et al., 2005, Liang et al., 2006, de Vries and Bonvin, 2008, Huang and Schroeder, 2008, Kowalsman and Eisenstein, 2009, Liang et al., 2009). Alternatively, experimental or prediction data can also be included directly, during the docking search, as restraints in force-field based docking approaches (de Vries et al., 2006, Melquiond and Bonvin, 2010). An example of this type of approach is the HADDOCK program that employs data-derived restraints during molecular dynamics and energy minimization to drive the docking towards a target region (Dominguez et al., 2003). A drawback of including experimental data or binding site predictions during docking is the restriction of the search such that at least some of the data-derived restraints must be fulfilled. If the data is incorrect or too inaccurate it may interfere with the success of the docking procedure (de Vries et al., 2010).

Another option is to include external data as weights on putative interacting (or non-interacting) atoms of the partner molecules. In the case of force field based docking methods, a weight larger than 1 on an atom implies a stronger interaction with atoms of the partner resulting in an improved score of solutions including the labeled (weighted) atoms and in addition, it also enhances sampling of the labeled region during the search. Compared to restraint-driven approaches the weighting of interactions does not exclude the sampling of regions not covered by the external data and an additional advantage is the possibility to only use available data on one partner structure. The possibility to include external data as weights has already been used in combination with Fast-Fourier-Transform (FFT) correlation-based docking methods (Ben-Zeev and Eisenstein, 2003). However, since in this case the sampling resolution is independent of the type of scoring function the approach is similar to re-evaluation of docked complexes including external data (after completion of the docking search).

This chapter presents the possibility to include either experimental data or bioinformatics predictions in the form of force field weights during docking using the ATTRACT docking program (compare chapter 2.3.3). The force field parameters of the ATTRACT docking program have been optimized recently to optimally identify near-native docking minima among a large set of incorrect decoy complexes (Fiorucci and Zacharias, 2010a). To further evaluate the performance of the parameters in this study, systematic docking was per-

formed on a set of complexes with available partner structures in the unbound conformation (Mintseris et al., 2005).

5.1 Settings and proceedings

Various types of weighting interaction potentials were evaluated. This includes the increase of all the interactions of selected pseudo atoms of a residue by a factor of 1.5 or 2.0 relative to the unbiased force field as well as weighting (scaling) only the attractive interaction. To evaluate the behavior of weighted residues, artificial binding sites were created, representing possible types of prediction. For the artificially created binding sites the center of the known binding site was often taken as a starting point to generate growing binding sites. To emulate predictions from experimental methods in one case, central binding site residues were selected and in an other case the residues were randomly selected out of all binding site residues. The meta-PPISP server (Qin and Zhou, 2007) delivered the information for the docking run with predicted binding sites. Details on scaling only selected residues or patches around a putative binding site are given in the following subsections as well as in the Figure legends.

5.1.1 Binding site prediction, benchmark set and acceptance criteria

Protein binding site predictions were obtained using the metaPPISP-Server. The metaPPISP server collects predictions from three different methods (cons-PPISP (Chen and Zhou, 2005), PINUP (Liang et al., 2006) and Promate (Neuvirth et al., 2004)) and forms a consensus (meta) prediction. It is considered as one of the most successful protein-protein binding site prediction methods (Zhou and Qin, 2007).

In order to evaluate the performance on unbound partner structures all systematic docking searches were performed on a benchmark set containing of 82 complexes in the unbound form from a published benchmark set (Benchmark 2.0) (Mintseris et al., 2005). This test set consisted of 63 rigid complexes (interface RMSD between bound and unbound on average 0.82 Å), 13 medium difficult (interface RMSD 1.63 Å) and 8 difficult cases (interface RMSD 3.67 Å), 23 of those complexes are enzyme inhibitor or enzyme substrate, 10 anti-

body antigen, 12 antigen and bound antibody and 39 other complexes. The structures pdb1n2c (medium difficulty, other type) and pdb2vis (rigid case, antigenâantibody complex) were omitted because of difficulties with binding site predictions using the meta-PPISP binding site prediction server, due to limitations in the maximum number of atoms (Qin and Zhou, 2007).

Docked complexes were clustered, and acceptable solutions according to the CAPRI criteria (Lensink et al., 2007) were collected and ranked only according to the ATTRACT score. The clustering of solutions was performed beginning with the lowest energy complexes (best scoring) using an RMSD of the ligand protein after superposition of receptor proteins (RMSDlig) cutoff between any two solutions of 5 Å. An acceptable solution is defined as an RMSD of the ligand (RMSDlig) $<$ 10 Å and a fraction of native contacts relative to the native complex larger than 0.1 or only a fraction of native contacts of more than 0.3.

5.1.2 Forcefield performance on the unbound test set

A recently designed ATTRACT force field based on optimizing the scoring of native interfaces relative to a large set of decoy surfaces gave a very good performance on bound partner structures with near-native solutions in the top 10 ranked complexes in 90% of the test cases (Fiorucci and Zacharias, 2010b). The unbound docking run followed the exact protocol (Fiorucci and Zacharias, 2010b) of the bound docking run. After clustering of the results in 55 out of 82 docking cases (around 65%) at least one acceptable solution was found in the top100 solutions. Other methods typically achieve this in only around 50% of the benchmark cases and typically only after rescoring including a variety of physicochemical and/or bioinformatics parameters (Cheng et al., 2007, Pierce and Weng, 2008, Liang et al., 2009, de Vries and Bonvin, 2011). Here, no knowledge on putative binding regions of the partner structures was included. However, for many docking cases of biological relevance experimental data on a putative protein binding interface in the form of mutagenesis data or evolutionary conservation are available. Such data can be included during docking in the form of data-driven restraints (e.g. used in the docking program HADDOCK) or for re-evaluation after an unrestrained docking search (Huang and Schroeder, 2008, Pons et al., 2010).

5.2 Docking performance evaluation using artificial binding sites

Docking with artificially generated or known binding sites served as proof of concept and to determine the feasibility of biased force field docking. Additionally, binding sites of varying overlap with the known binding site were created in order to estimate the behavior of the force field on weighted regions, as was expected to be proposed by binding site prediction or experimental methods.

5.2.1 Creation of artificial binding sites for known binding site docking

The known binding site was estimated by defining a heavy atom in the binding site, if within a distance of 5 Å to one heavy atom of the partner protein, for a proof of concept docking run. The residues allocated to those atoms are taken as binding site residues for weighted docking. To account for the impact of weighting, several different scaling methods have been tested. One is the magnitude of the scaling factor which is set to either 1.5 or 2.0 of the original forces. Another, the handling of attractive and repulsive forces, where either both or only the attractive forces are regulated. Besides the kind of scaling, the impact of the size and position of the weighted region needs to be observed. Several different conditions have been tested to monitor the behaviour of the ATTRACT docking program concerning changes in specificity and sensitivity. To create those artificial surface residue patches, the surface residue closest to the geometric center of the known binding site was used as an anchor. Additional residues are then added, beginning with the closest surface residues to the anchor resulting in rather circular binding spots.

5.2.2 Docking with known binding sites

The possibility of including data on putative binding sites in the form of weights on the force field contributions is ideal for the ATTRACT docking method. This is based on docking energy minimization in a force field that drives the interaction between two proteins. In order to test different combinations of weights for binding and non-binding regions it was first tested for the case of known protein binding regions for both partner structures. For the

weights on those atoms that are known to be involved in binding, 2.0 was used. Although this seems a small weight in the case of a known binding region it was found that higher weights were not beneficial in the case of less reliable data on a binding site (see below) and a weight of 2.0 was therefore chosen as an upper limit. As expected, inclusion of data on the known binding protein sites significantly enhances the docking performance and increases the number of successful docking cases with acceptable solutions in the top1 or top2-10 category. Scaling up the weights in the known binding region as well as the additional downscaling of the non-binding region reduces the space of acceptable complexes concerning the energy to complexes near to the native ones. Methods scaling attractive and repulsive forces results in a top1 rate of around 55% and a top10 rate of 88% (for weights of 2.0 for binding site residues and 0.75 for non binding site residues) while scaling only attractive forces leads to a top1 rate of around 67% and a top10 rate of 94% (same weighting as previous). Best results were achieved with an up-scaling of only attractive interactions involving pseudo-atoms of residues in the known interface by factors 1.5 or 2.0.

In addition, the failure rate drops from five runs without an acceptable solution to one in the run scaling only attractive forces. The fact that there are still cases were acceptable solutions where only found in the top10-100 category (or not at all) is due to the large conformational differences between unbound and bound partner structures in these cases. Increasing the attraction of a known interface region does not restrict the relative orientation of protein partners upon binding. Including weights did not only affect the scoring but also produced overall more acceptable solutions. While in the reference (unbiased) run 952 acceptable clusters of solutions were found (for all runs together) this number more than tripled in the case of including weights on known interface residues (e.g. 2886 in the case of weighting attractive interactions by 2.0). The improved sampling can also be observed in the funnel plots (Figure 5.1) which show a significant shift towards lower ligand-RMSD for sampled complexes with the lowest energy. Counting all receptor ligand contacts of the top 1000 complexes reveals that structures with high amount of native contacts are scored high even if the RMSD is not acceptable, due to rotation and/or tilts of the ligand within the receptor's binding site (all results of the best ranked 1000 structures have native contacts > 0.0).

In addition, the number of native contacts for the best acceptable solution increased from on average 51% (standard deviation (SD) = 23%) in the refer-

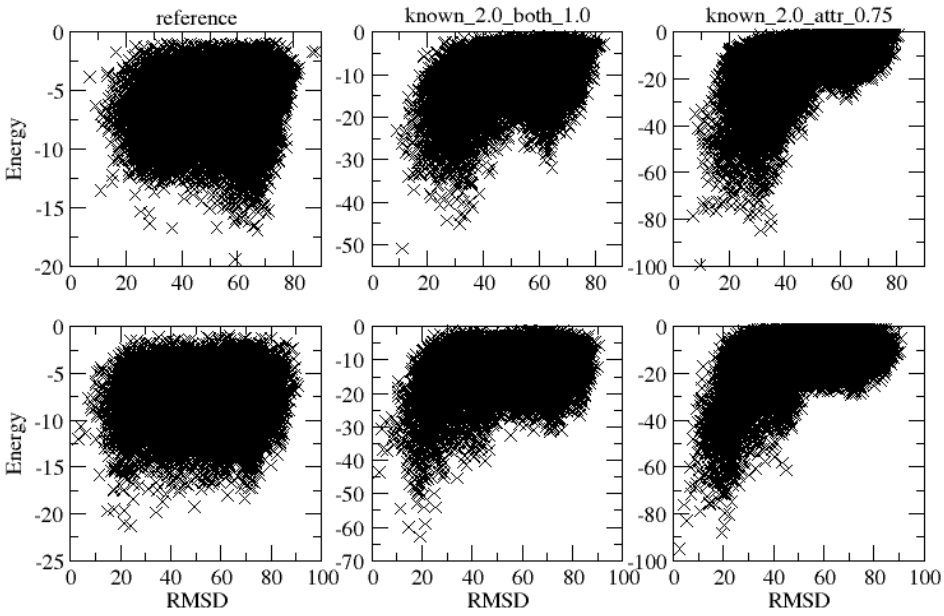


Figure 5.1 – Upper plots: pdb1m10 with a reference rank of 9159 for the first structure with ligand-RMSD of < 10.0 Å, weights of 2.0 for all forces rank 510 and scaling down non binding site residues to 0.75 while only attractive forces are scaled rank 1. Lower plots: pdb1tmq with reference rank 889, rank 33 and rank 1 for same weighting as above.

ence search to 58% (SD = 23%) in docking searches including weights. The top 10/top 100 results included structures with on average significantly more native contacts (top10: 25% +/- 11% top100: 12% +/- 5%) compared to the reference docking search (top10: 4% +/- 5% top100: 2% +/- 2%). Instead of including weights during docking, it is also possible to rescore solutions from the reference docking search using the force field weights. Rescoring can change the number of acceptable solutions within a given range of ranking but cannot improve the quality or number of acceptable solutions. Acceptable solutions were found in 56 of the 82 cases on rank 1 for docking including weights (2.0) on known binding site residues. Whereas only 49 top 1 solutions were found using rescoring of unbiased docking runs. No acceptable solution could be found in the reference run for three complexes (and of course also not upon rescoring). The weighted docking runs found acceptable solutions for two of those complexes (pdb1h1v rank 108 and ligand RMSD 8.9 and pdb1ibr rank 22 and RMSD 8.6, both cases claimed as difficult in the benchmark). Overall, using weights during docking resulted in an average rank of acceptable structures of 5.4 and for rescoring 10.3. Especially for difficult and medium difficult cases, the use of weights during docking improves results (compared with a rescoring of the structures obtained by the unbiased searches). Docking

with weights performed better in nine medium and difficult cases, and for two structures acceptable solutions could be found that had not been found during unbiased docking.

In realistic docking cases the protein binding site is often only approximately known. To mimic such conditions, artificial binding sites (representing possible predictions or known regions) were generated, which represented different scenarios in terms of sensitivity and specificity of the approximately known regions of interaction.

5.2.3 Variations of binding sites covering possible binding site prediction motifs

"Interface center" indicates inclusion of force field weights on only the central 50% of the known binding sites on both partners and performed similar to "Double interface" which consists of twice the size of the known binding site and including the known binding site completely. In both cases ca. 80% of acceptable solutions scored among top ranks or top2-10 ranks. Increasing the mean weights on 50% of the protein's surface residues including the known binding region (named "50% of total") still results in acceptable solutions within the top10 in 63% of the cases, while weights on random surface patches with partial overlap with the known binding sites on both partners, lead to a significant performance drop comparable to the results of the unbiased run ("Overlapping Site"). Since predictions or experimental knowledge is often of limited accuracy the results of the test runs indicate that an over-prediction of the binding site (too large but including the complete native binding site) is more promising than an attempt to exactly replicate the shape or size of the binding site.

In comparison "Random assignment" of putative binding regions on the protein surface excluding the known binding site (prediction with nearly zero sensitivity and specificity) resulted in a significant drop of the docking performance, even below the performance of the reference docking search. However, still in the majority of cases an acceptable docking solution within the predictions was found but ranked very badly (only four within the top100). This indicates a distinct advantage of the present technique of including putative binding site data as force field weights instead of restricting the search to predicted interaction regions (e.g. as restraints). Inclusion of incorrect predictions (zero sensitivity and zero specificity) as distance or contact restraints for ex-

Table 5.1 – Artificial binding site statistics

method	receptor fraction	ligand fraction	sensitivity	specificity
Interface center	4	8	40	100
Doubled interface	18	36	100	35
Overlapping Site	10	20	35	35
50% of total	48	47	100	33
Random assignment	8	16	< 1	< 1
Known binding sites	10	20	-	-

Receptor fraction and ligand fraction correspond to the average fraction of protein surface that included weights on surface atoms. For definition of sensitivity and specificity see legend of Table 5.2.

ample, would have resulted in complete failure of the docking searches for all complexes. An overview of the performance of the different test cases investigated in this chapter, also including the known binding site, can be found in Figure 5.2.

5.3 Single and multiple residue experiments

Analogue to former, the different methods were explored as well for inclusion of knowledge on single residues. A single residue was chosen randomly from the core of the above determined binding site residues of one of the binding partners (in all cases the receptor). For the docking runs with multiple residues three out of all binding site residues were selected without any restriction. Additionally, residues in the neighborhood of a given residue were also weighted reducing the weight distance dependent of $1.5 - (0.035 * distance)$ in case of single residue. This means that in a radius of $\sim$14 Ångstroms around the known residue the forces are raised while in a larger distance constantly reduced down to a minimum of 0.75 times the normal forces. For the run including three residues surrounding atoms were weighted with $1.5 - (0.0625 * distance)$ resulting in artificial binding site spots of 8 Å radius around each residue including 13% of all surface residues. Beyond the up-scaled binding site the attracting forces are linearly reduced to a minimum of 0.75 leaving the weighted residues outstanding as attractive regions among a surrounding of gradually more repulsive surface residues.

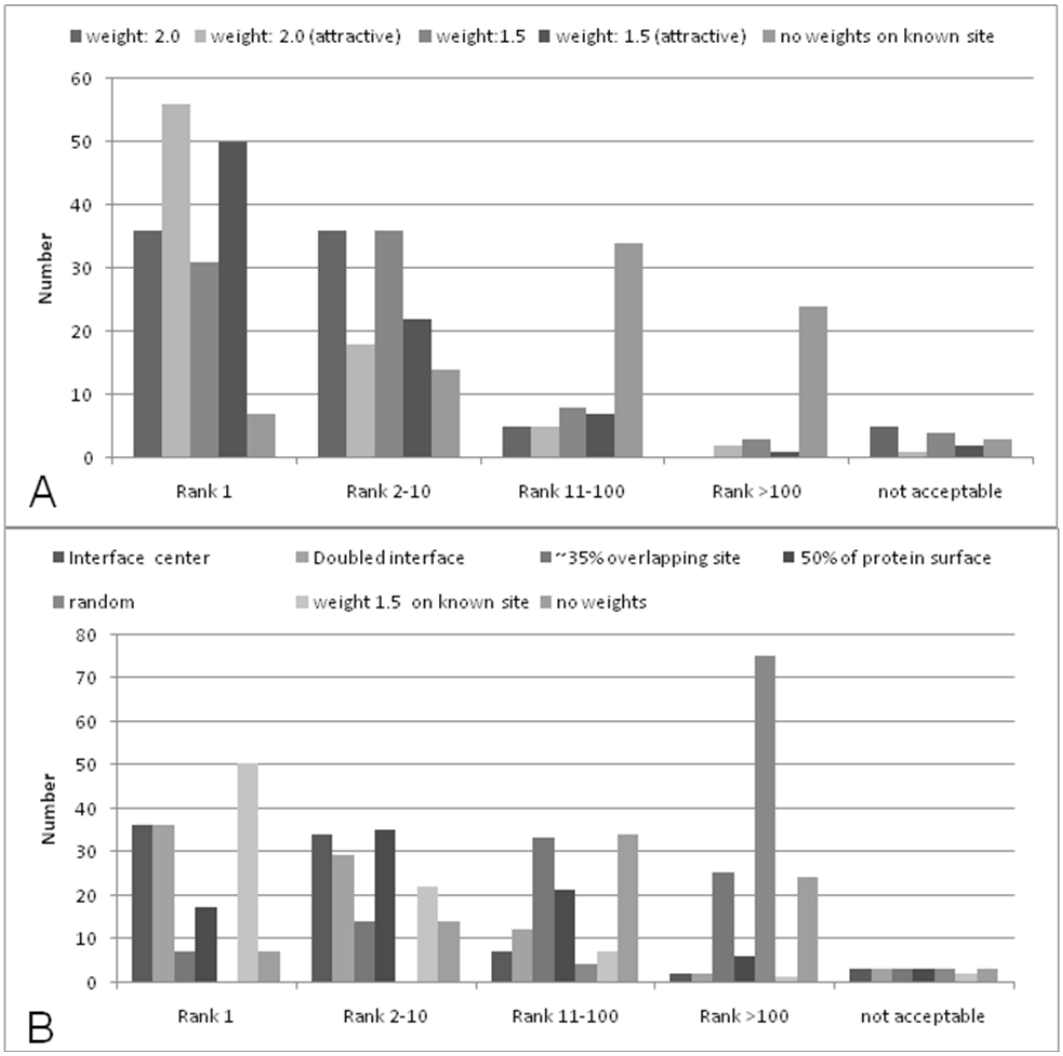


Figure 5.2 – A) Comparison of various types of weighting (scaling) interactions that involve coarse-grained-atoms that belong to a known interface. The term attractive refers to scaling only the attractive contribution (otherwise both repulsive and attractive) by a factor of 2.0 or 1.5, respectively, compared to the unbiased docking run. The ranking (x-axis) indicates the rank of the best acceptable solution (according to the CAPRI criteria). The y-axis indicates the number of solutions (out of 82) in each category. B) Effect of modifying the part of the surface for which attractive interactions were increased by 1.5. Interface center indicates weighting of only half of the known interface residues (interface center) whereas doubled interface means weighting twice as many residues but including the known interface residues. Boxes for ∼35% overlap indicate weights on an artificial random protein surface area that includes on average ∼35% of the known interface whereas 50% of protein surface indicate weights on 50% of both proteins total surfaces including the known binding region. Random indicates weighted random artificial binding sites that do not overlap with the known binding site.

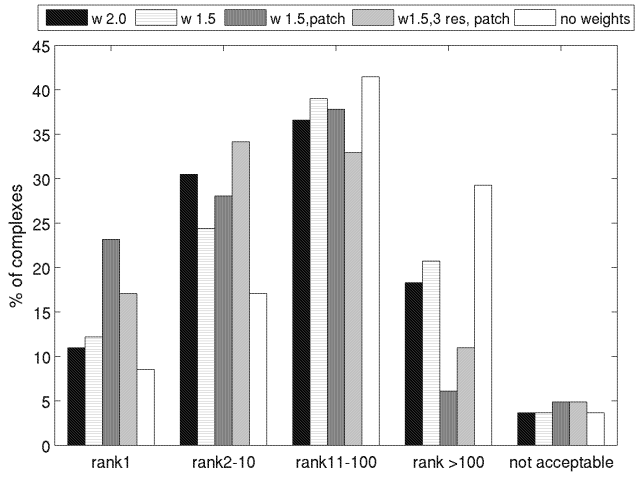


Figure 5.3 – Docking performance in case of scaling the interactions of single residues located approximately at the center of a protein binding region (only attractive interactions by 1.5). For comparison the results of scaling not only a single residue but also residues within 14 Å of the selected residue (in a linearly decreasing manner according to $w = 1.5 - (0.0035 * distance)$ from the selected residue: termed patch) and for three randomly selected residues in the known binding region are also shown

5.3.1 Inclusion of knowledge on single residues

Frequently, mutagenesis experiments on proteins can identify residues that are likely to be part of a protein-protein binding interface. Docking searches including weights only on one partner (receptor) and on one central core residue were performed. This shifts the score distribution of acceptable solutions already significantly towards better ranking solutions (Figure 5.3). Knowing and weighting one single residue of one of the partners increased the number of acceptable solutions in the top10 from ∼22% in the reference run to ∼36%. Increasing the weights also for residues in the neighborhood of the selected residue (as described above) further improved the docking results.

Similarly, picking three random positions in the known interface (instead of a central interface residue) and increase the interaction weight around them resulted in acceptable docking solution in the top10 category in 51% of the cases (80% in the top100 category, see Figure 5.3). Placement of those three residues beside each other at the edge of the binding region can lead to a high affinity complex generation where the ligand is shifted or tilted towards those high attractive residues. This causes a shift towards non-native complexes in terms of sampling and in terms of scoring an increase in energy for structures out of alignment with the known ligand protein. If the residues are distributed equally within the binding site or at least one or more are in the core, docking

results are improved significantly. Using this kind of information leads to 9 docking runs with a loss of ranks (776 in total 86.22 on average), 11 stay the same (all are either already top ranked or no acceptable structure was found) while 62 are improved resulting in a "win" of 5042 positions (81.32 on average; counted are only those results where the reference run and the new run are below a best rank of 1000, including all the bettering in terms of ranks is 30733, 495.69 on average).

5.4 Handling information from protein-protein binding site predictions

Several methods for predicting putative protein binding sites on proteins have been developed (reviewed in de Vries and Bonvin 2008). The metaPPISP-Server combines the results of three separate approaches and forms a consensus prediction (Zhou and Qin, 2007). The statistical accuracy of the predictions was evaluated for the benchmark set of partner structures (using the default threshold for predictions, Table 5.2 and Figure 5.4). On average a sensitivity of ∼37% was observed indicating that on average about 37% of the predicted residues overlapped with the native binding site.

To account for the limited precision of predictions two variants of interaction weighting were tested. In a first approach predicted residues were scaled with values of either 1.5 or 2.0. Residues were chosen following the proposal of metaPPISP to count residues as predicted with prediction values above 0.34 (for the distribution of prediction values compare Figure 5.4). In a second set of test runs the prediction values were directly taken as values in the range of 1.0 to 1.5 and 1.0 to 2.0 respectively, resulting in smoother transition to high prediction values accounting for uncertainty of predictions with values slightly above the cut-off of 0.34.

5.4.1 Inclusion of predictions on protein-protein interaction regions

Information from predictions are not as reliable as experimental knowledge and often lead to a diffuse binding site spot so that in most cases it does not perfectly fit the real binding site but is rather shifted or sometimes misplaced

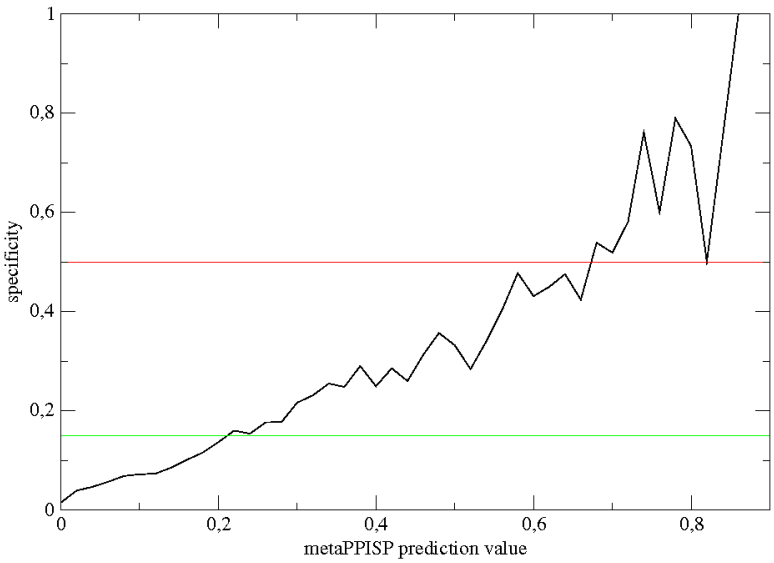


Figure 5.4 – Specificity of single metaPPISP (Qin and Zhou, 2007) prediction values within a bin size of 0.05. The red line indicates a specificity of 50%. The green line represents the percentage of binding site residues among all surface residues. Values above the green line are better than random values, since the fraction of binding site residues among all surface residues is 15% (10% for receptor and 20% for ligand) for the complexes in the test set of complexes.

Table 5.2 – Prediction accuracy of the metaPPISP-Server on the test set of unbound protein structures.

	receptor	ligand	average
Sensitivity	0.34	0.40	0.37
Specificity	0.36	0.45	0.40
Accuracy	0.87	0.76	0.81

Definition of Sensitivity: TP/(TP+FN); Specificity: TP/(TP+FP); Accuracy (TP+TN)/(TP+TN+FP+FN) with TP true positive, TN true negative, FP false positive and FN false negative. Prediction values >= 0.34 are counted for statistics, lower values neglected.

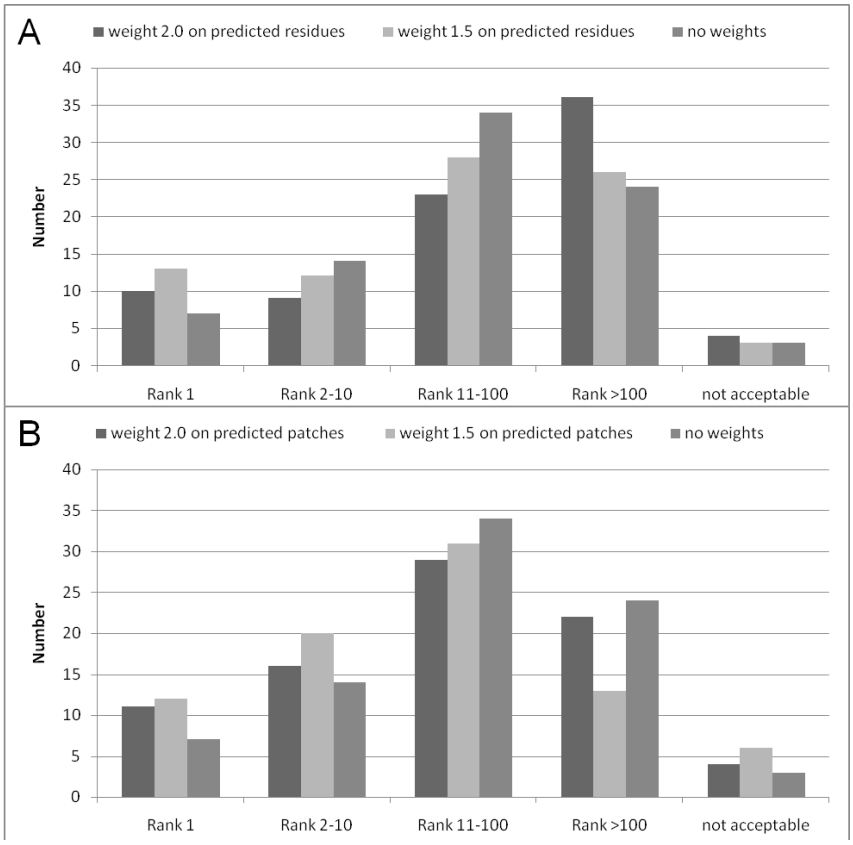


Figure 5.5 – Effect of including weights on residues predicted to be part of protein-protein interfaces using the metaPPISP approach (Qin and Zhou, 2007). A) All interactions of pseudo atoms with a metaPPISP prediction threshold of 0.34 (recommended significance threshold of the metaPPISP server) were uniformly assigned a weight of 1.5 and 2.0, respectively. B) The weight on predicted interface atoms was linearly weighted according to the prediction of the metaPPISP-server using w=0.5*PPISPscore + 1.0 with a maximum weight of 1.5 or 2.0, respectively.

completely. Additionally the size of the prediction and/or the number of spots predicted might deviate from the requested native binding geometry. The latter might be more a feature of binding site prediction since some proteins can bind to more than one partner but also intense the difficulties for the prediction of bound complexes. This lack of accuracy is the reason why scaling predicted values injudicious and high often leads to bad results. As the shifting of the known binding site showed, binding site spots overlapping with but not covering the entire binding site often result in very bad binding geometry prediction (compare run Overlapping site in Figure 5.2B) equal to the unbiased docking.

In the first method all residue atoms above a prediction threshold of 0.34 were assigned a constant weight. In this case a weight of 1.5 or 2.0 (only for attractive interactions) assigned to predicted residues resulted in improved dock-

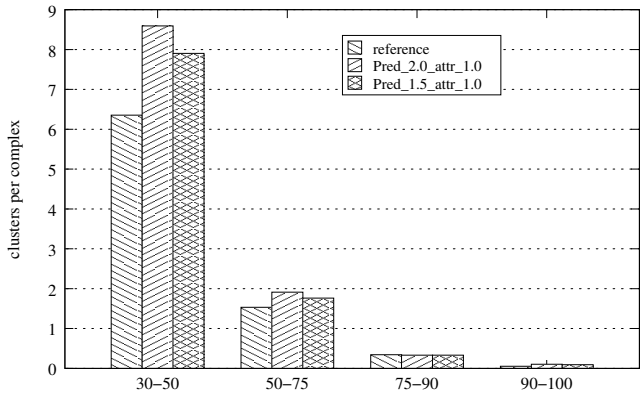


Figure 5.6 – Comparison of the quality of acceptable docking solutions (in terms of native interface contacts) after systematic docking on all test cases. Native contact calculations included all inter protein coarse-grained atom contacts within a distance of < 7 Å. This corresponds closely to the < 5 Å criterion used at atomic resolution. This includes also complexes that were not treated as acceptable solutions.

ing results (Figure 5.5B). However, the improvement is modest and concerns mostly the docking of complexes already in the top categories. In a second set of systematic docking searches the prediction values for each residue (from the metaPPISP-server) were directly (linearly) translated into attractive weights (weight range 1.0-1.5 or 1.0-2.0). This gave a further improvement of docking results with ∼40% acceptable solutions in the top10 and 77% in the top100 categories (weight 1.5) compared to 22% and 67% in the reference case without bias (Figure 5.5B).

The inclusion of binding site predictions resulted not only in an overall improved ranking of acceptable solutions but also in an improvement of the quality of the docking solutions in terms of the number of native contacts (Figure 5.5 shows the final rankings; Figure 5.6 and Figure 5.7 show the improvement in terms of native contacts and ligand RMSD). The improvement compared to the unbiased docking runs is only modest and, as expected, less than the improvement found in the case of including weights on known binding regions. Figure 5.8 illustrates successful predictions and docking results for one example (pdb2sic). In this case with an accurate binding site prediction (Figure 5.8A) dramatic improvement of the docking performance can be observed (Figure 5.8B-E).

However, due to the limited accuracy of the binding site prediction, the data set contained also many examples where the prediction was completely wrong (illustrated in Figure 5.8F for pdb1qa9) which resulted in weighting of predicted regions with no overlap to the native interface at all. As a result the scoring of incorrect complexes was enhanced compared to the solution

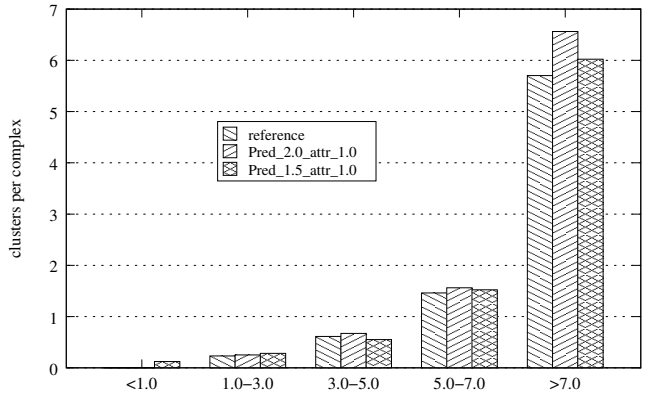


Figure 5.7 – Comparison of the quality of acceptable docking solutions (in terms of ligand RMSD) after systematic docking on all test cases. Values signed as greater than 7.0Å are in the range of 7.0 to 10.0Å. Only acceptable structures were considered.

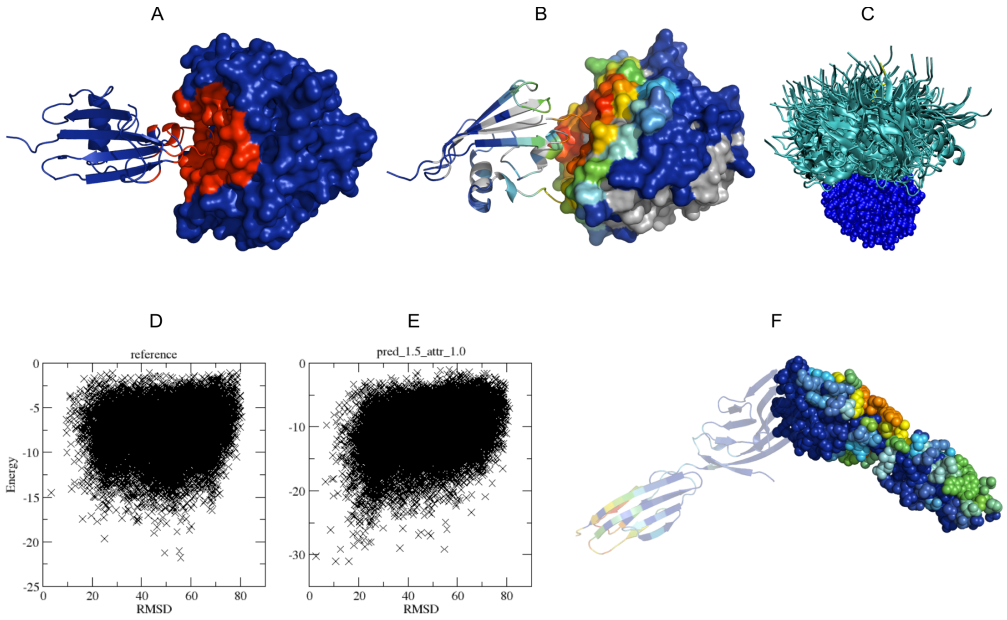


Figure 5.8 – A) Meta-PPISP based predicted interface residues (red) on the surface of the protein partners forming the complex pdb2sic (residues not weighted for docking in blue). The receptor protein is shown as molecular surface (ligand as cartoon model). B) Docking sampling (mapping contacting residues in docked complexes) of the top1000 solutions including weights on PPISP-predicted interface residues (shown in A). Grey indicates no sampling at all, red indicates dense sampling. C) Top100 ligand protein placements (cartoon) from the same docking run (receptor as blue molecular surface; native ligand placement in yellow).D) Score of docked complexes vs. deviation (only ligand) from the known placement of the ligand protein relative to the fixed receptor protein during an unbiased systematic docking run. E) Same as in D) but for a systematic docking run with weights of 1.5 on all meta-PPISP predicted interface residues. F) Meta-PPISP-predicted interface residues mapped on the surface of the binding partners of the complex pdb1qa9 (same color coding as in B), receptor shown as cartoon, ligand in van der Waals sphere representation).

close to the native complex. For example, pdb1qa9 was predicted with 4.2 Å ligand RMSD on rank 70 within a cluster of 11 structures and 28 structures with a ligand RMSD of less than 10.0 Å at all. The weighted run lead to a dramatic decrease of sampled structures as well as scoring. The structure that was sampled in the unbiased run was then ranked 5409 (in a run with a hard prediction spot) with an energy of 1/40 of the top ranked. Reducing weights to the softest method tested lead to a rank of 660 (with weights up to 1.5) up to rank above 1000 (with maximum weight of 2.0). One has to say that despite the fact that the prediction was as wrong as it possibly could be for both proteins (which is not often the case) the weighted run was able to find a solution.

All in all, the number of structures with an RMSD lower than 10.0 Å increases with weighting (see Figure 5.7) as well as the scoring. On average the rank of the best ranked acceptable structure in the unbiased run was around 650 while 250 using weights. The funnel plot in Figure 5.8D and E shows the impact of weights, on the entire sampling and ranking, as a shift towards weighted regions. Whether this results in better RMSD is determined by the quality of the prediction. To overcome the unfavorable shift towards lower ranks for runs with bad prediction, reducing the weights to a maximum of 1.5 in linear scaling provides an excellent compromise between supporting with good predictions and conserving acceptable results when predictions are not beneficial.

5.5 Flexible docking using normal mode relaxation

The ATTRACT program enables inclusion of minimization of docking partners not only in translational and orientational degrees of freedom but also in a set of normal modes of the binding partners (or other collective degrees of freedom) based on an elastic network model of proteins (May and Zacharias, 2005, 2008).

The minimization in normal modes can account approximately for induced fit effects during the protein-protein binding process. The effect of including conformational relaxation of both partners in the 5 softest normal modes during systematic docking was also tested (Figure 5.9). In some docking cases

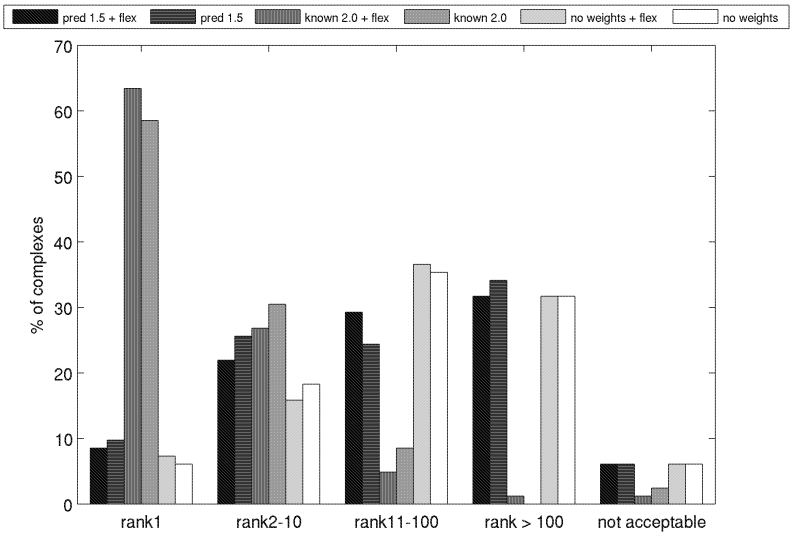


Figure 5.9 – Systematic docking search including minimization in 5 pre-calculated soft normal modes of both binding partners (indicated as +flexibility). Performance was evaluated including elastic network derived normal modes (May and Zacharias, 2008) and for rigid docking on unbound partner proteins employing otherwise identical docking search conditions. Added force field weights are indicated as known 2.0 (weights of 2.0 on known interface atoms), prediction 2.0 (weights of 2.0 on residues predicted to be at the interface using the metaPPISP server)

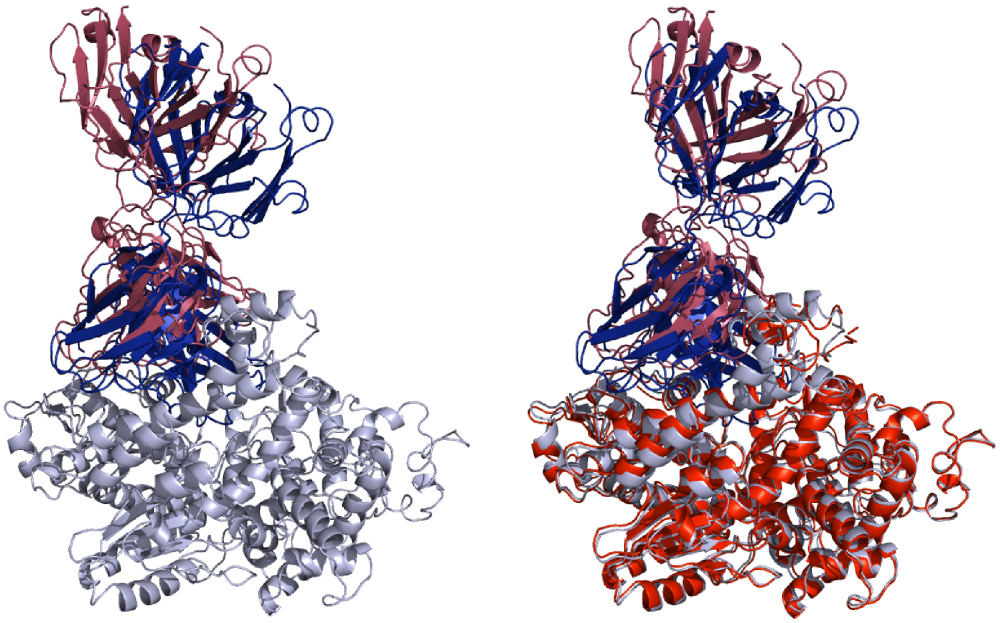


Figure 5.10 – Docking of the unbound partner structures of complex pdb 1BGX (light blue: receptor and dark blue the ligand). On the left in darker red the docking solution closest to the native complex (12.6 Å and not acceptable) from a systematic ATTRACT docking. On the right same docking search approach but including minimization in the 5 softest normal modes of both partners during docking. Darker red shows the solution with lowest ligand RMSD from the native complex (ligand RMSD=9.6 Å, acceptable solution) and as light red cartoon the deformed receptor structure.

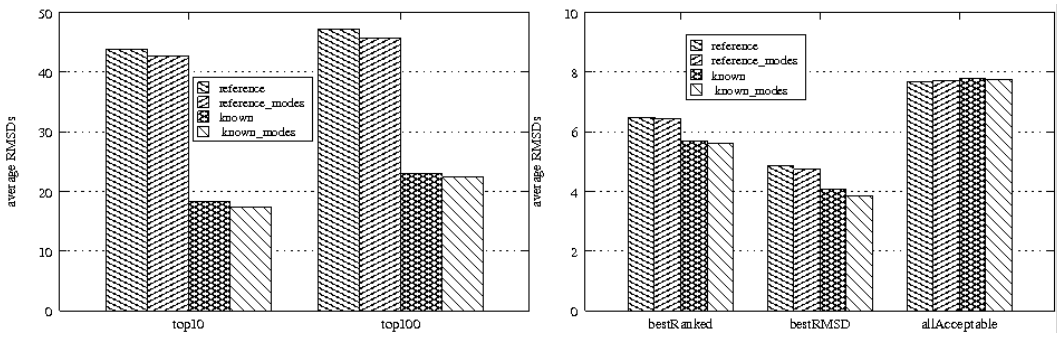


Figure 5.11 – Left: Average RMSD of the top10 and top100 results including runs where no acceptable structure was found at all. Right: Average RMSD of the best ranked, the best RMSD and of all acceptable structures (< 10.0Å ligand RMSD) for docking runs with modes and without.

significant improvement of the results could be obtained (e.g. see Figure 5.10). However, in several other cases the inclusion of normal mode minimization during docking was also of benefit for incorrect docking solutions. Hence, it improved also the induced-fit and scoring of non-native solutions such, that overall the docking performance, in terms of acceptable solutions, did not improve for unbiased docking or docking including binding site predictions. In the case of known binding sites the percentage of acceptable solutions in the top10 category increased from ∼84% (rigid) to 91% in the case of including normal mode relaxation. Figure 5.9 shows the results for all docking runs. For the run without any bias, the performance does not result in global ranking changes, except some improvement on the one side and some worsening on the other. Only docking runs with weights on the known binding sites improve in terms of ranking and sampling. Due to the impact of weights on docking, the results are already very good in the rigid case so that improvement using flexibility is moderate. On the other hand, some clashes could be avoided which resulted in an increasing number of top1 positions and a reduction of >10 Å results. Additionally, using modes it was possible to create an acceptable structure for complex pdb1bgx (see Figure 5.10) which was not possible in all the other runs. The ranking of this result is with rank 569 rather poor but within the top10 there are some results just missing the criteria of less than 10.0 Å ligand RMSD. These one might become acceptable after some refinement steps.

Docking with modes does not only sample additional or differing solutions but can also improve the aggregation of structures that could also be found in the unbiased (reference) rigid docking runs. In most cases the effect is modest and the average RMSD is just slightly decreased. As Figure 5.11

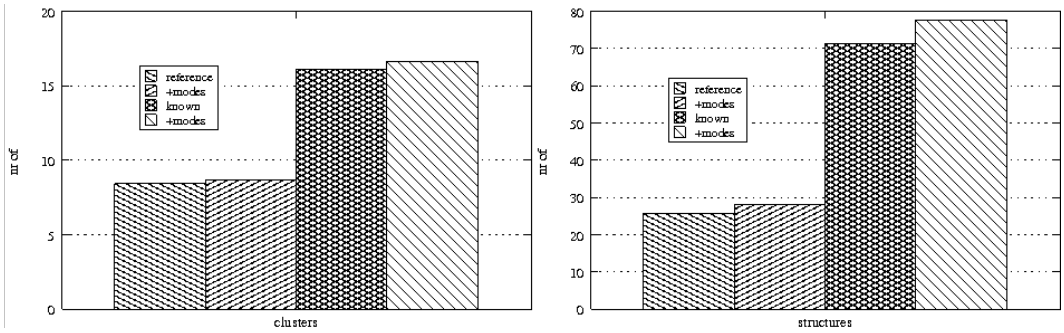


Figure 5.12 – Average number of clusters (left) and clustered structures (right) found for a docking run with acceptable structures.

shows, the improvement is nearly negligible for the best ranked acceptable solution but noticeable for the best RMSD solutions. Overall, the average RMSD of all acceptable structures does not change significantly, because the average number of acceptable structures increases (compare Figure 5.11 and 5.12). Some structures not acceptable in rigid docking became acceptable due to the introduction of flexibility.

Plus to the decrease of RMSD for best ranked and best RMSD structures the overall composition of top10 and top100 changes, with reference to the average RMSD. While in the global view of a complete benchmark this might be a small decrease in RMSD, it might be significant for single runs (see Table 5.3). Since this also includes results where no acceptable structure was found at all – with a significant larger amount of those results for the unbiased docking runs – it can be found as an additional benefit of flexible docking.

Besides the RMSD of the structures, also the ranking is important for further selections of complexes for refinement. The complete improvement in ranks for the reference run is thereby remarkable: for the 82 docked complexes with modes 1415 ranks could be "won". This includes only solutions where in the unbiased reference run as also in the unbiased run with modes the rank was below the 1000 mark. Neglecting this, the total win in places is 10733 resulting in an average rank of best solution of 646.54 for the reference and 525.64 for the run with modes (to compare fairly runs for pdbs 2btf, 1k5d, 1ibr, 1ib1, 1h1v, 1bgx were not included in this analysis because either one or both runs did not find an acceptable solution). Hence, docking with modes might not be favorable for each docking problem but in general a decrease of RMSD and average rank as well as an increase of the number of acceptable structures and clusters could be observed.

Table 5.3 – Examples for flexible docking RMSDs

pdb	best rank structure	best RMSD structure	top10 structure	top100 structure	nr clusters structure
1ATN	9.27 (rank 102)	8.76 (rank 103)	69.15	61.72	2
1ATN modes	9.29 (rank 129)	5.16 (rank 372)	52.25	53.55	4
1EER	6.90 (rank 686)	6.90 (rank 686)	41.17	42.87	2
1EER modes	8.23 (rank 161)	6.79 (rank 807)	37.56	38.67	3
1IQD	9.19 (rank 5)	3.09 (rank 93)	46.46	54.53	3
1IQD modes	2.62 (rank 11)	2.62 (rank 11)	54.40	59.17	5

For the best ranked structure and best RMSD structure the RMSD is shown while for the top10 and top100 results the average RMSD of the structures is shown. The last column shows the increase of acceptable clusters due to flexible docking.

5.6 Binding site prediction using top ranked docking results

As found in the previous chapters, the ATTRACT force field itself is often able to find near native complexes within the top scoring results. Investigating the residue-residue contacts between receptor and ligand for the best ranked results reveals which contacts are favored by the ATTRACT force field and therefore, which regions on the proteins' surfaces are favored binding regions. Since no restraints between residues or any additional information is used during ATTRACTs minimization step except the force field, it is possible that the partners find the right binding region but not the right binding mode (as noted for docking with the known binding site in chapter 5.2.1). However, these docking results can give hints as to where the partners favor to bind and therefore, offer the possibility to extract additional information useful for further docking. Analysis of energy changes upon complex formation (like e.g. the change in desolvation energy as indicator for the optimal docking area (ODA) (Fernández-Recio et al., 2005)) have been used before and proved good performance as a predictor.

5.6.1 Binding site prediction using the ATTRACT score

To extract information out of docking runs for the given test set (Benchmark 2.0 (Mintseris et al., 2005), compare chapter 5.1) the top1000 results of an unbiased run were extracted and analyzed. For each residue, the contact with

Table 5.4 – Prediction accuracy of the ATTRACT scoring function on the test set of unbound protein structures.

	average
Sensitivity	0.60
Specificity	0.35
Accuracy	0.75

For definition compare table 5.2. Only residues with a probability of 20% are used. The fraction of the predicted surface residues among all surface residues was 13% in the case of the receptor and 46% in the case of the ligand proteins (average 29%).

a residue in the partner protein in a docking solution was counted for each of the 1000 results and then normalized to gain a probability of finding a certain residue in contact with the partner protein. A contact is fulfilled when the distance between the residue and the partner protein is lower than < 5.5 Å in coarse grained resolution, which is more rigid than the < 5.0 Å criterion normally taken for all atom contact determination. This ensures, that only essential close contact residues are taken into account. Values are mapped on the residues and compared with the known binding site (Table 5.4).

The fraction of predicted binding site residues for the receptor fits quite nicely to the known fraction of binding sites (see table 5.1) while for the ligand nearly half of the surface is predicted to be in the binding site. Due to the systematic docking approach many of rotations of a ligand are sampled during the docking procedure, combined with the lower number of surface residues a high overprediction occurs. Since overprediction showed good performance during systematic docking with weights (see Figure 5.2) and the size of larger proteins binding sites is approximated properly, a residue in contact with the partner protein in more than 20% of the results is treated as a high affinity residue and therefore counted as predicted binding site residue.

These results are comparable to the results from the meta-PPISP server (compare Table 5.2) and thus, can be used for weighted docking in a similar manner. One drawback of the method is that for complexes where no near native complex is found in the top results, the prediction will lead a second docking run towards even worse results. Also the size of the receptor protein may influence the results since more non native conformations might emerge between native conformations among the top results because of the systematic

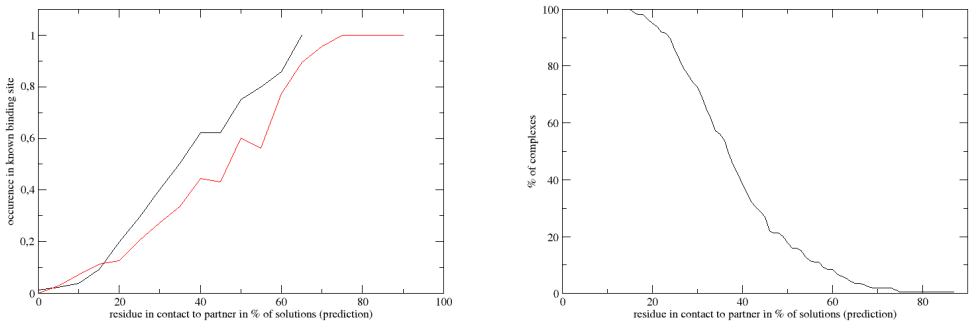


Figure 5.13 – Left: The occurrence in the binding site for specified sampling values are given for the receptor (black line) and ligand (red line) proteins. Right: Probability to find a value or a higher value mapped on a structure.

search along the protein's surface, so that the rate of the highest probability is lower than for small molecules. Even if these residues are significantly more often sampled than other residues, in many cases the upper limit is around 20% for the highest sampled residues (see occurrence of values in Figure 5.13). This method indeed identifies residues in contact with the partner protein during docking but is not sensitive to the arrangement of binding sites, so that predicted residues might be distributed over the complete surface (see Figure 5.14), which is largely found in the case of ligand molecules.

5.6.2 Weighted docking using ATTRACT score as predictor

Docking runs have been performed following the procedure given in chapter 5.1. A maximum weight of 2.0 was chosen. Such high values are rarely reached (compare Figure 5.13) so that in fact the weighted residues between 1.0 and 1.2 can be found in every protein while residues weighted with 1.4 or higher are only found in less than half of the structures. This should make it approximately comparable to the results of the docking runs using meta-PPISP prediction with maximum weights of 1.5 (see Figure 5.5).

The docking runs using the ATTRACT score as a bias, show that the results drift further apart from each other. While complexes which had a high amount of native contacts within the top1000 structures (and in most cases a noticeable amount of acceptable structures) in the reference run make further improvement, already badly ranked solutions are ranked worse. In single cases

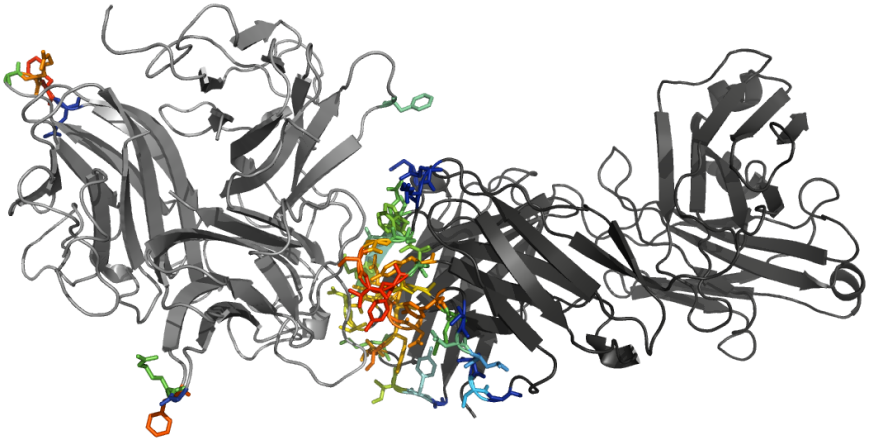


Figure 5.14 – On the left the ligand of pdb1i9r is shown (lighter grey), receptor on the right (darker grey). Residues which had contact to the partner in more than 20% of the results are shown as sticks. Colors represent the occurrence with blue the lowest amount and red the highest. Grey cartoon representation indicates sampling of contacts in less than 20% of the cases.

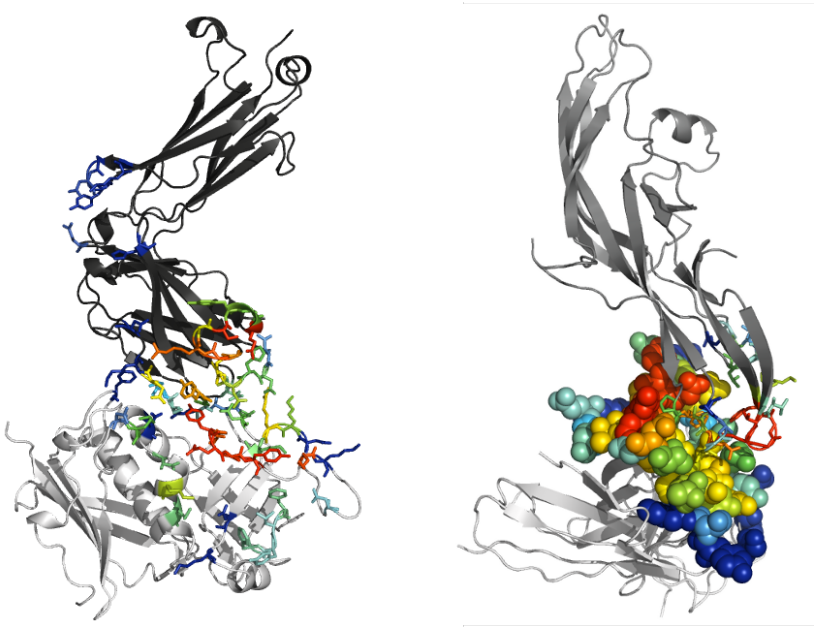


Figure 5.15 – Left: pdb1sbb Right: pdb2qwf. Residues which had contact to the partner in more than 20% of the results are shown as sticks for the ligand and in sticks (pdb1sbb) or sphere representation (pdb2qwf) for the receptor. Colors represent the occurrence with blue the lowest amount and red the highest. Grey cartoon representation indicates sampling of contacts in less than 20% of the cases.

even well ranked solutions in the reference run are ranked worse after redocking. For example pdb2sic was listed in the reference run on rank 178 with a 3.34 Å ligand RMSD structure and high amount of native contacts. Unfortunately, this cluster of complexes was together with one other complex the only acceptable solution in the top1000 while nearly all other solutions had zero native contacts. In contrast pdb2hmi was listed on rank 5946 in the unbiased run and in the biased rank 208 was achieved for the best acceptable solution. The top1000 of the reference run of pdb2hmi were populated by structures with native contacts of 20% or more, which is slightly below the acceptance criteria of 30%. Additionally, complexes with the ligand at the right position but in wrong orientation in the reference run could be improved, due to the weighted docking, in terms of scoring and sampling of near native structures.

For complexes with already excellent sampling in the top solutions of the reference run, the ranking and ligand RMSD improves further e.g. pdb2qwf (see Figure 5.15 on the right) which was rank 27 with 9 Å ligand RMSD in the reference and rank 3 with 7.6 Å ligand RMSD in the biased run. For medium quality sampled solutions (e.g pdb1sbb see Figure 5.15 on the left) the rank, RMSD or native contacts for the best ranked acceptable solution vary only slightly. Overall the ranking is constantly improved, so that the average rank for best ranked acceptable solution was in the reference run 647 (51 complexes ranked in the top100) and was reduced to 430 (55 ranked in top100) in the biased run. Using meta-PPISP predictions as a bias instead of the ATTRACT docking score, for 63 complexes an acceptable solution was found within the top100 solutions (10% more acceptable solutions in the top100 compared to the ATTRACT score bias).

However, as Figure 5.13 shows, the prediction quality for the fixed receptor is higher than for the rotated ligand, so that a combination with information from binding site predictors can help to overcome this limitation. In this approach, all residues above a given threshold were weighted during docking, regardless of the neighborhood. The creation of continuous patches can also improve the docking procedure and can be combined with other prediction methods to increase the sensitivity of the prediction.

5.7 Conclusion

One way of improving computational methods for prediction of protein-protein binding geometries is to include experimental data. Simple restriction of the search to a predicted binding region or inclusion as restraints during the search can be a drawback in as much that incorrect data limits the docking run or is guided to the incorrect binding region. Inclusion of experimental or prediction data, in the form of force field weights on predicted interface residues, creates a bias towards the predicted binding region, without excluding other possible binding regions and can be used even if only data for one partner protein is available. Inclusion during the docking itself may also enhance the number of sampled docking solutions near a predicted binding region.

A systematic evaluation of the approach in combination with the ATTRACT docking method indicates, that indeed it can significantly improve the docking performance if reliable data on putative binding regions is available. Even single identified interface residues can significantly improve docking results. However, including prediction data on putative protein binding sites, gave only modest improvement of the docking performance, compared to the unbiased reference search. This is due to the limited precision of such predictions and fully consistent with test cases on designed artificial binding sites that included on average, only part of the known binding site. Although it was tested on one prediction server only it is expected that the result will be similar to other prediction methods since the performance of the best available methods differs only slightly in terms of prediction accuracy (de Vries and Bonvin, 2008).

The inclusion of binding site predictions as ambiguous restraints was recently tested on a similar benchmark (de Vries and Bonvin, 2011). Acceptable solutions were found in the top100 docking solutions for ∼41% of the cases (in the case of including predictions) and ∼15% for unbiased (ab initio) docking. This compares to 77% (including predictions) and 65% (unbiased docking) in this work. Hence, for the ATTRACT method the scoring without additional data performs quite well but the gain upon inclusion of prediction data is smaller compared to the study using predicted data as ambiguous restraints.

The scoring of the ATTRACT force field showed a tendency towards native contacts even in the unbiased docking run. Among the residues found in contact with the partner protein in the 1000 best ranked solutions, a high amount is part of the known binding site. Nevertheless, using the contacts

retrieved from the unbiased run to bias a second run, resulted only in a slight improvement. This was mostly due to the fact, that the ligand was docked systematically from a variety of starting points and in different rotations, so that high contact residues were spread over the surface.

It should be emphasized here that even a modest improvement of the docking performance and a shift of some of the docking results from the top10-100 category to the top10 category can be very helpful. The focus was on the first systematic search without inclusion of any refinement or re-scoring of the docking predictions. Typically, a limited set of docking solutions will enter a refinement stage at atomic resolution using molecular mechanics modeling methods and possible rescoring of the solutions. This can make a further improvement in docking prediction accuracy and specificity. A restriction of the refinement step on a small set of putative docking solutions is an important prerequisite for the success of the refinement and rescoring of predicted complexes.

5.7.1 Acknowledgment

Parts of this chapter have been published in Schneider and Zacharias 2012a.

Chapter 6

Prediction of the antibody Fc interaction with the C1q component

To solve a structure of a biomolecule in atomic resolution, high quality methods like X-ray crystallography or NMR Spectroscopy are used in the first place. Although these methods, among others, can clarify many questions concerning the three dimensional structure of biomolecules, certain limitations restrict the applicability, for example the size of a system or the stability of a transient complex (Davis et al., 2008, Kobilka and Schertler, 2008, Claudio and Dalvit, 2009). Other methods attempt to indirectly state assumptions on the three dimensional structure of macromolecules by e.g. mutagenesis experiments (Moreira et al., 2007b). This allows to identify, whether certain changes in the amino acid composition of the protein cause local or even global changes, facilitating the analysis of function and structure of the system (Moreira et al., 2007a, Bradshaw et al., 2011, Kelly et al., 2012). Complexes unable to be resolved by X-ray crystallography (and also other methods) because of their size, stability or availability e.g. in a crystal, are often target of such mutagenesis experiments. In such experiments, residues assumed to have functional impact are mutated to selectively lower certain features relevant for the observed process, e.g. hydrophobicity or electrostatics of the region (Williams et al., 2006, Bostrom et al., 2009). Although these methods yield valuable information, they can not provide the three dimensional description of a complex.

Computational methods can enlarge the experimentally obtained indications towards three dimensional structures, using molecular docking methods. As shown in the previous chapter, even least amount of information can help to generate near native structures, while for medical relevant systems often a noticeable amount of trustworthy informations can be found. After refinement, employing molecular dynamics (MD) simulations in explicit solvent, docked complexes can serve as a working model to better understand the details of the investigated system and to plan future mutagenesis studies for refining the structural model or for improving binding specificity and affinity.

In this chapter, a model in atomic resolution of a pharmaceutically relevant complex is presented, which was generated using previously presented docking protocols including relevant data from mutagenesis experiments (see Chapter 5 for the protocol). This data allowed to guide the complex prediction process towards experimentally documented interactions, while it also allows to highlight unknown or unnoticed contacts between both partner proteins.

6.1 The C1q-IgG1 complex

One interesting case with uncertain knowledge of the three dimensional structure is the C1-IgG1-Fc complex of the complement system. Several binding modes have been proposed, relying on a wide range of mutagenesis experiments, but without identifying the complete interaction spectrum in the binding site or prospects of all atom models. The high amount of experimental data and the medicinal relevance of the complex made it a worthwhile target for sophisticated docking and simulation techniques. Besides the interest of fundamental research, an all atom model of this complex would allow further experiments and affinity analysis, supporting clinical research on the complement system.

6.1.1 Function of the C1q-antibody-Fc binding in the complement system

The C1 complex of the complement system is a multi-domain protein that triggers activation of the classical pathway of complement which is a major part of the innate immunity (Ehrnthaller et al., 2011). The C1 component consists of six C1q subunits with an overall shape of a bouquet of flowers

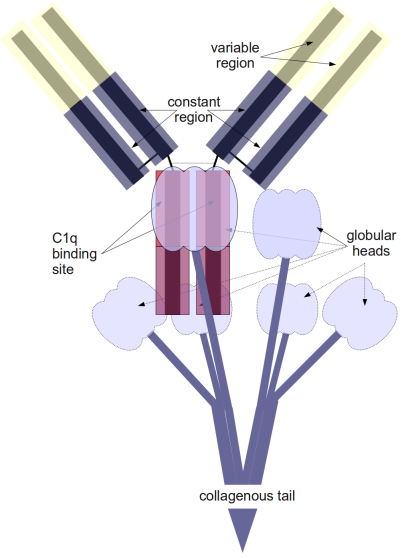


Figure 6.1 – Schematic presentation of the rough binding mode of the C1-IgG1-Fc complex. The upper part shows the antibody structure in rectangular shapes and in the lower part the trimer of the C1q subunit as part of the C1 complex

(illustrated in Figure 6.1). Complement activation is triggered by the initial binding of the globular domains of the C1q subunits to the Fc portion of IgG or IgM antibodies bound to antigens on the bacterial surface (Sarma and Ward, 2011, Ehrnthaller et al., 2011). Multiple binding is required to stabilize this initial molecular complex and leads to a cooperative response which can distinguish between antibody molecules bound to a bacterial cell from those free in solution. The process of multiple C1q binding to a bacterial cell eventually results in complement activation and elimination of the bacterial cell. Stable binding of two or more of the six globular heads of C1q has to be established in order to activate the complement C1 component. The binding event triggers the activation of the C1r and C1s protease subunits of C1 which, in turn, activates the classical complement cascade (Zlatarova et al., 2006, Trouw and Daha, 2011, Ehrnthaller et al., 2011).

The C1q globular domain is a heterotrimeric complex with an arrangement of protein modules that has been found in several other proteins (Ghai et al., 2007) and is not exclusively binding to antibodies but also to a variety of other ligands including C-reactive protein, pentraxin 3 and several bacterial surface proteins (Lu et al., 2008). The structure of the C1q subunit and the structure of antibody Fc regions have been determined experimentally (Gaboriaud et al., 2003, 2004). In addition, residues important for the interaction of the

two proteins were identified by mutagenesis studies (e.g. Kojouharova et al. (2003, 2004), Kishore et al. (2004), Moore et al. (2010)) and an approximate arrangement was suggested (Gaboriaud et al., 2003), but the complex structure between the two proteins has so far not been determined experimentally in atomic detail. Putative interaction regions between C1q and the antibody Fc domain have been investigated by mutagenesis studies with the aim to enhance the complement activation of therapeutic antibodies (Idusogie et al., 2000, Moore et al., 2010). Engineering the antibody Fc region to enhance the cytotoxic activity of antibodies offers the potential of designing monoclonal therapeutic antibodies with greatly increased destructive capabilities against specific target cells. Such target cells include not only invading pathogens but also cancer cells.

6.1.2 Residues involved in C1q-antibody-Fc binding

The putative interaction region of the globular C1q trimer with the IgG1-Fc domain of antibodies has been characterized by several mutagenesis studies (Kaul and Loos, 1997, Kishore et al., 2002, 2004, Kojouharova et al., 2003, 2004, Roumenina et al., 2006, Zlatarova et al., 2006, Gadjeva et al., 2008). It was possible to identify key residues in C1q that when substituted significantly reduce the binding affinity to Fc (Figure 6.2). These residues include an Arginine at position 162 in chain A (in short ArgA162), ArgB108, HisB117, ArgB129, ArgB163 and ArgC156. Several mutations in Fc were identified that influence binding to C1q including also some substitutions that increase the affinity (Thommesen et al., 2000, Idusogie et al., 2001, Michaelsen et al., 2006, Presta, 2008, Moore et al., 2010). For example, the substitution of Asp270, Lys322 and especially Pro329 and Pro331 on Fc is known to reduce C1q binding (Burton et al., 1980, Idusogie et al., 2000, 2001, Oganesyan et al., 2008).

It has also been possible to increase the affinity of C1q to the Fc region of IgG1 after substitution of the amino acid Lysine at position 326 to Tryptophan (in short Lys326Trp) and Glu333Ser (Idusogie et al., 2001). In addition, Moore et al. (2010) found that the substitution of Ser267Glu, His268Phe and Ser324Thr significantly increased the binding affinity of C1q to Fc. Interestingly, the putative interaction regions on the surface of C1q and Fc show also a high degree of electrostatic complementarity (illustrated in Figure 6.2).

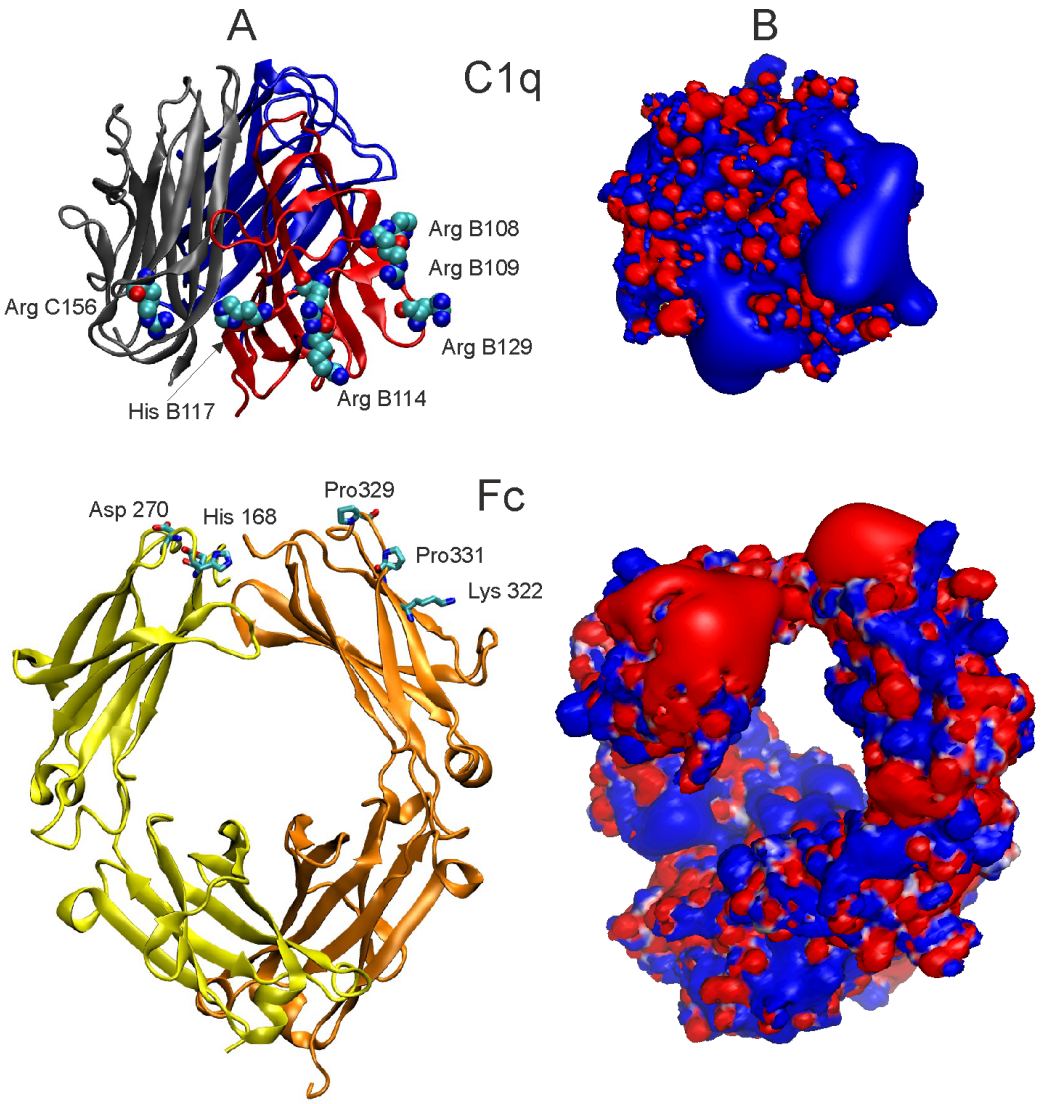


Figure 6.2 – (A) Cartoon representation of the globular C1q subunit (upper panel) and the IgG1-Fc subunit (lower panel). Several amino acid residues that have been found experimentally to contribute to the C1q-Fc interaction are labeled and indicated as van der Waals spheres (in case of C1q) or as stick model (for Fc). The cartoon colour coding is blue for chain A, red for chain B and grey for chain C, respectively, of each subunit of the C1q trimer. In case of Fc chain A is in orange and chain B in yellow. (B) The electrostatic potential around C1q (upper panel) and Fc (lower panel) contoured at the 2 kT level (k: Boltzmann constant and T; Temperature: 300 K). Negative potential is shown in red whereas regions of positive potential are indicated in blue. The view is approximately the same as in (A) indicating electrostatic complementarity of the proposed interaction regions.

6.2 Computer-aided all atom complex prediction

In this approach the prediction of the all atom complex consisted of three steps. The foundation of the complex prediction was the fast coarse grained docking procedure which allowed to rapidly generate binding modes in a reduced atom model. Although the number of starting points in the systematic docking procedure could be lowered by neglecting starting points at areas far away from the assumed binding region, using minimization in a variety of orientations results in thousands of independent structures. Translation of the coarse grained to all atom model followed by a short high temperature simulation in vacuum was used to filter possible binding configurations prior to the more sophisticated and time consuming refinement with explicit solvent MD simulations.

6.2.1 Docking and refinement scheme

Binding mode prediction using coarse grained systematic docking

The protein data bank (pdb) entries 1PK6 and 1HZH served as unbound C1q and antibody Fc partner structures, respectively, for all protein-protein docking searches. Only the IgG1 Fc part of the antibody structure (residues 245-475 of chains H and K) in the 1HZH-entry was used for docking. Systematic docking searches were performed using the ATTRACT docking program (Zacharias (2003, 2010a), Fiorucci and Zacharias (2010b); also compare chapter 2.3.3 for the ATTRACT docking program and 5.3 for the protocol used). A weight of 2.0 for the attractive interactions was included for residues that have been shown experimentally to be important for C1q-Fc interaction. Surrounding residues have also been weighted which has been found on many test cases as a good bias to include experimental data on protein binding sites (compare chapter 5.3.1). Systematic docking was started from several starting points (spaced by 8 Å) and orientations of the C1q partner near the roughly known binding region of the antibody Fc subunit. Each docking run consisted of a set of energy minimizations in translational and orientational variables following the previously used protocol (see chapter 5). Comparison of known structures of the Fc portion of IgG antibodies in complex with different partner proteins indicates that global adjustments may play a role during interaction. To

account for such global conformational changes in the Fc segment during docking, minimization in soft global collective modes were included in the docking procedure (May and Zacharias (2005, 2008) and chapter 5.5).

Structure refinement at atomic resolution

Atomic models of the docked complexes were obtained by superposition of the atomic resolution partner structures onto the docked complexes in coarse-grained resolution. The refinement steps were performed with the AMBER11 all-atom molecular modelling package employing the parmff03.r1 force field (Case et al., 2010).

The refinement procedure consisted of two steps: the first involves an initial energy minimization to remove sterical overlap after coarse grained to all-atom translation (1500 steps, using a distance-dependent dielectric constant: $\epsilon = 4r$, r is the distance between atoms). A further optimization step includes the heating of the system in vacuum using three consecutive MD simulations at temperatures of 550 K (6 ps), 400 K (4 ps) and 300 K (4 ps) followed by 1500 EM steps. During the MD-simulations harmonic distance restraints between Cα atoms within each partner protein were applied. This allows full flexibility of side chains and limited flexibility of the backbone of each partner and flexible adjustment of the interacting partners.

The second and more time consuming refinement step involved MD simulations in explicit solvent. After heating up during 1 ns to 300 K and removal of positional restraints on the two proteins, each docked complex was equilibrated for 11 ns. For the equilibration procedure three different treatments of restraints covered possible scenarios: Simulations were run either with no restraints at all, low restraints on the Cα atoms (0.25 $kcal/mol{-}$Å^2) or with distance restraints between the approximated interface residues as used in the first refinement step. The final structures were again energy minimized (2000 EM steps) and served as model structures for the C1q-Fc-complex.

6.2.2 C1-IgG1 complex prediction

The knowledge of the putative C1q binding region on the Fc dimer allowed focusing the docking search by starting from six initial placements close to the putative binding region. Each starting placement included 300 different starting orientations that were docked using the ATTRACT energy minimization

approach (Zacharias, 2003, May and Zacharias, 2008, Fiorucci and Zacharias, 2010b). During docking the flexibility of the Fc protein was approximately included by energy minimization in the 5 softest normal modes obtained from an elastic network model (ENM) of each partner protein. The simultaneous optimization in translation and orientation and in the softest collective degrees of freedom allowed for an induced fit during docking. Putative residues involved in binding to Fc have also been investigated on C1q by mutagenesis studies (Kojouharova et al., 2003, 2004, Kishore et al., 2004, Roumenina et al., 2006) which resulted in the identification of several basic (mainly Arg and His) residues in the B as well as C subunits of the C1q heterotrimer important for binding (Figure 6.2).

Selection of residues based on mutagenesis experiments

Experimental data was included into the docking following previously established procedures as force field weights. Inclusion of such data can significantly enhance the docking performance resulting in improved ranking of near-native docking solutions and more realistic docking geometries (compare chapter 5.3 or Schneider and Zacharias (2011)). During docking force field weights for attractive forces were doubled for several residues involved in binding as suggested by experimental mutagenesis studies. For the IgG1-Fc domain this included the following residues: Pro329, Pro331,Glu318, Lys320, Lys322, Asp270) and for the C1q molecule residues ArgB114, HisB117, ArgB129, ArgB163. The importance of these residues for binding was shown in several independent experimental studies giving it high confidence.

Docking and structure refinement

Clustering of all docked structures resulted in 1210 solutions (that differed in the placement of C1q relative to Fc by an $Rmsd_{lig}$ C1q of > 5 Å after best superposition with respect to Fc). Further filtering of the docking solutions involved both experimental data on putative residues at the interface as well as the scoring of the docking geometries. A residue was counted as forming part of the interface if it was in contact with one or more residue of the partner protein. Two residues were considered in contact if any pair of coarse-grained pseudo-atoms of two residues was within a distance of 7 Å. The docking scoring function in ATTRACT was shown to be quite effective in identifying near-

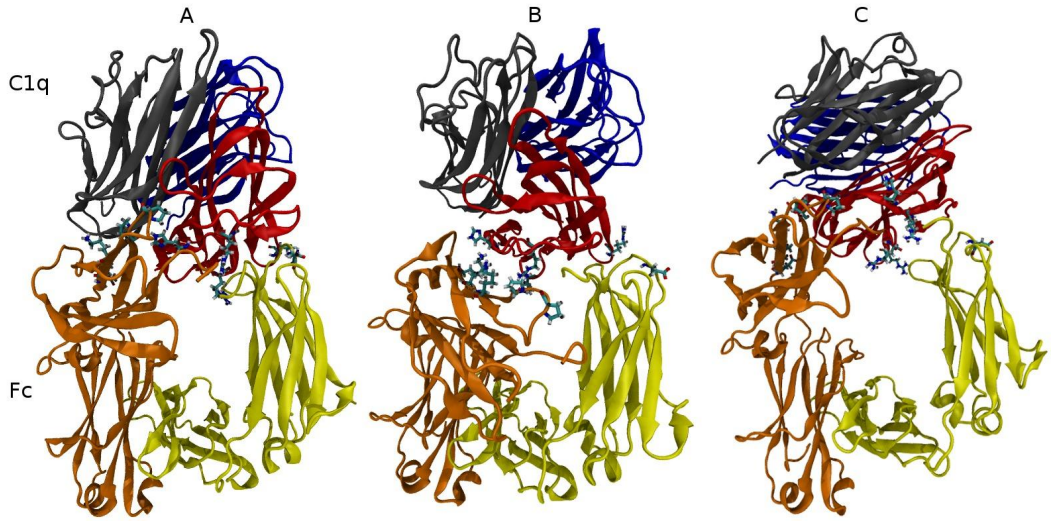


Figure 6.3 – A) Cartoon representation of the best scoring docking model that agreed with experimental results on C1q-Fc interaction after refinement simulations in explicit solvent. Several residues important for binding that were weighted during docking are indicated as sticks. B), C) Representative structural models for two alternative C1q-Fc docking models that showed larger deviations of the Fc-structure from the start structure and agreed only partially with available experimental data. Colour coding is the same as in Figure 6.2

native docking solutions especially in combination with force field weights on putative residues involved in protein-protein interaction (see chapter 5 and (Schneider and Zacharias, 2011)). However, in order not to overlook a putative native-like solution, not only the best scoring docking solutions but also all solutions with 25% of the top score were considered. Of these complexes a subset of 20 complexes that fulfilled 80% of the experimental data on putative interface residues was selected for further refinement using in vacuum high temperature MD simulations. This cutoff was chosen since test simulations indicated that the refinement procedure can still change the percentage of interface contacts by ∼20% such that the refinement procedure may result in perfect (100%) agreement with available experimental data.

Clustering of these initially refined 20 structures identified seven complexes that were selected for further refinement by MD simulations at all-atom level, including explicit solvent. After refinement using explicit solvent MD, resulting structures could be assembled to three clusters of complexes which kept more than 80% of the contacts that involved residues identified by experiment to influence binding. The two smaller clusters of complexes (named model 2 and model 3) showed large conformational changes of the Fc antibody subdomain (Figure 6.3, backbone RMSD in the binding region > 7 Å with respect to the Fc start structure). These changes are larger than what was observed in

previous complexes with different protein partners (Ghai et al., 2007).

Selection of representative model for the C1q-IgG1-Fc complex

Structural model 1 agreed with all experimental mutagenesis data and showed the least deviation of the partner proteins in the complex from the corresponding unbound conformations (overall $Rmsd_{backbone} < 1.5$ Å with respect to the corresponding experimental crystal structures of Fc and C1q, respectively). Model 1 also represented the highest populated cluster of docked complexes for which all members agreed to 80% to 100% with the experimental data on interface residues hence a representative structure of this cluster serves as model structure for further analysis. Interestingly, all three final models contact mainly one of the Fc monomers (chain A) that contain the Pro329 and Pro331 residues. Contacts to the second Fc monomer (chain B) involve residues Asp270 and Glu272 of Fc (Figure 6.3). Michaelson et al. (2006) found that mutagenesis of Pro329 to Ala (Pro329Ala) resulted in loss of complement activation if the mutation is introduced in both Fc monomers but that activation is still possible if only one monomer has been mutated. The model structure of the C1q-Fc complex explains this observation since contacts predicted involve Pro329 and Pro331 of just one Fc monomer and the same residues in the other monomer remain fully exposed to the solvent. As indicated in Figure 6.4 the structural model is also sterically fully compatible with the presence of the carbohydrate structure connected with the Fc antibody part.

6.3 Identification of surface residues that mediate the interaction

The structural models of the C1q-Fc interaction allow the identification of possible contacts that mediate the interaction between the two proteins which have not been considered in previous experimental studies. This can be helpful for designing new mutations to prove or disprove a predicted contact, hence to validate the model, or can help to design mutations that may even enhance the interaction between the two proteins, which can help to improve the efficiency of therapeutic antibodies. In contrast to mutagenesis studies that can help to identify putative interface residues that are important for C1q-Fc interactions the final model of the C1q-Fc-complex helps to assign one or more

interaction partner residues to each residue implicated in mediating C1q-Fc interactions. On the C1q side several mostly basic residues have been identified to participate in the interaction with C1q. In case of residues ArgB108, ArgB109, ArgB114, GluB162, ArgB129 and ArgC156 partner residues in hydrogen bonding or salt bridge forming distance are observed in the final model structure (Figure 6.4). This involves an electrostatic contact between ArgB108 and Asp270 (on the B-chain of Fc).

Decrease of the C1q-Fc binding was observed after substitution of residue Fc:Asp270 (Idusogie et al., 2000, Thommesen et al., 2000). For residue Fc: Lys322 (chain B) contacts to GluC187 and Glu352 (same Fc chain) and also the nearby ArgC156 on C1q was found as a possible network of contacting partners. This interaction is located at the rim of the binding region and may depend on other residues nearby. It has been found that the involvement of Lys322 in C1q-Fc interaction depends on the subclass (Thommesen et al., 2000). In the model residue ArgC156 of C1q can form a salt bridge with Glu333 of Fc. However, in this configuration Glu333 is located at the interface and is partially buried which may destabilize the binding. Indeed, it has been found that substitution of Glu333 by Ser (or Ala) increases the affinity of the complex presumably due to the removal of an unfavorable buried charge in the protein-protein interface (Idusogie et al., 2001).

Substitution of Ser267 by a Glu was found to significantly enhance the C1q-Fc interaction (Moore et al., 2010). This result fits to the proposed model very well. In the model structure Ser267 (chain B of Fc) is located near LysB123 and ArgB108 of C1q and introducing a Glu at 267 strongly increases the electrostatic interaction with these two basic residues (see Figure 6.4). Since most of the electrostatic contacts are mediated by residues at the rim region of the C1q-Fc complex interface, the stability of these interactions were monitored during the MD simulations in explicit solvent. Many of the electrostatic interactions of the rim region form transient contacts during the simulations with hydrogen bonds in rapid exchange with solvent. However, for some contacts hydrogen bonding with long life time were observed as indicated in Table 6.1.

The structural model also gives an explanation for the importance of the two Pro329 and Pro331 residues for the interaction with C1q. In the model these two residues fit into a pocket that is located at the interface between chain B and C of C1q (Figure 6.4). In contrast to most interactions with C1q and the chain B of Fc, the interactions of the two Pro residues (located

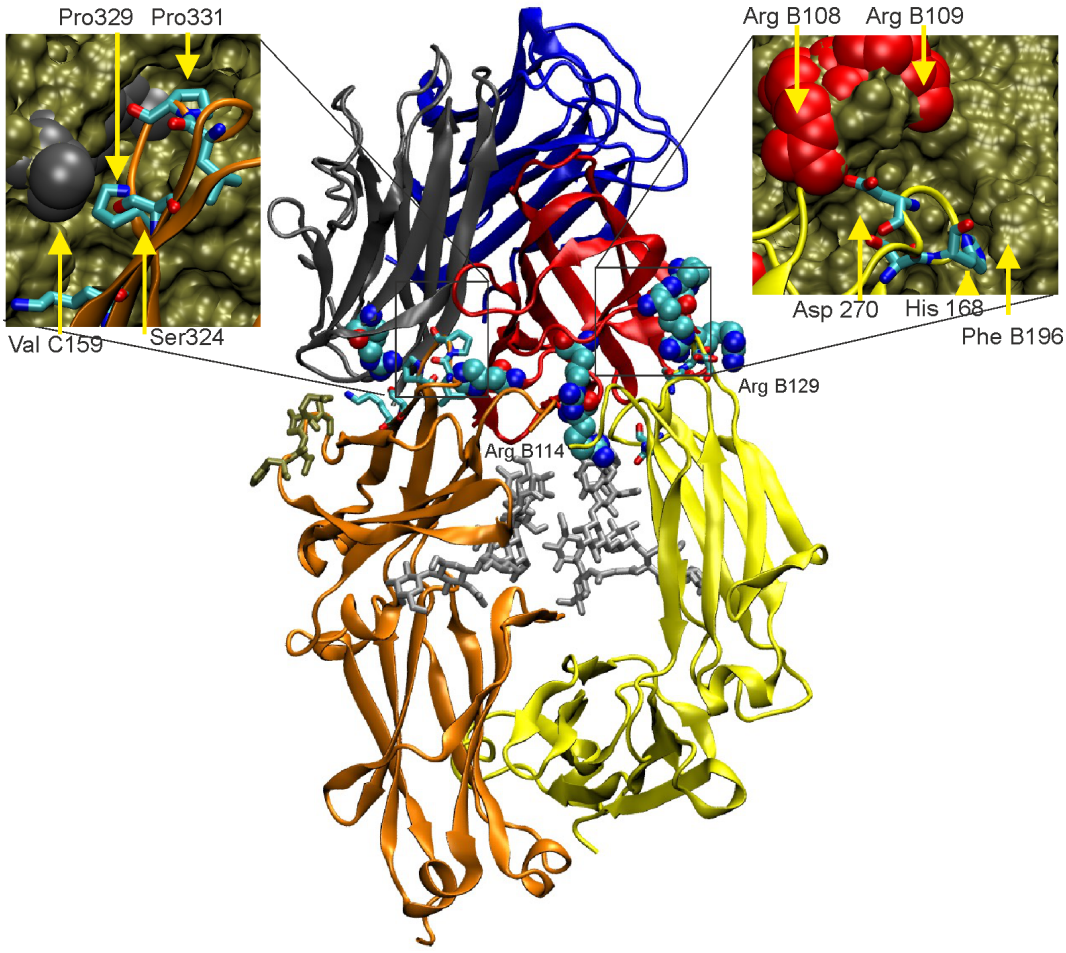


Figure 6.4 – Cartoon representation of C1q-Fc model complex structure in best agreement with available experimental data (same colour coding as in Figure 6.2 and 6.3). Interface residues important for binding are indicated as sticks (Fc) or van der Waals spheres (C1q). Two proposed important binding regions are highlighted in the insets with C1q in surface representation and contacting Fc residues as stick model. The positions of key residues for the interaction are labelled.

Table 6.1 – High affinity contacts of the C1q-IgG1 complex

residue C1q	residue IgG1-Fc	Occupancy
ArgC156	GluA333	100%
ArgC182	GluA318	100%
ThrB134	AsnB297	80%
ArgB161	GluB294	70%
AsnB194	HisB268	50%
ThrB112	GluB269	50%

Hydrogen bonding across the C1q-FC of interface residues during refinement simulation of model 1. Occupancy describes the hydrogen bonding over time.

in chain A of Fc) with C1q involves several non-polar contacts. It predicts contacts of Fc:Pro329 to residues ThrB100, ThrB102, IleB119, ValC162 and HisB117. Substitution of the latter residue was shown to significantly reduce C1q-Fc binding affinity (Kojouharova et al., 2004). Residue Pro331 contacts ValC159 and ValC161, but shares the contact with ValC159 with SerA324 of the Fc structure. This contact offers an explanation for the observation that the substitution Ser324 by a slightly larger but more hydrophobic residue (Thr) with an additional methyl group increases the affinity of the C1q-Fc complex (Moore et al., 2010).

In addition to explaining the importance of Pro329, Pro331 and of the Ser324Thr substitution for C1q-Fc interaction, the model also suggests a hydrophobic contact between both partners at the interface between chain B of C1q and chain B of Fc which involves HisB268 of the Fc molecule. It has been found by Moore et al. (2010) that the substitution His268Phe actually increases the affinity of the complex. In the present model residue HisB268 (of Fc) is located in a binding pocket formed by the hydrophobic residues (of C1q) PheB196, MetB158 and AlaB164 (Figure 6.4). Residue PheB196 is in close proximity such that it probably could form efficient stacking interaction with a Phe268 which could explain the increase in C1q-Fc binding observed experimentally.

Several residue substitutions have been found on the surface of the IgG1 Fc antibody fragment that did not influence binding. This includes residues Lys276, Tyr278 an Asp280 (Thommesen et al., 2000) as well as Val282, Val284 and His285 (Moore et al., 2010). In the final model, all these residues are out-

Table 6.2 – Predicted contacts at the interface between C1q and IgG1-Fc

C1q trimer	IgG1-Fc dimer
ArgB108	AspB270
ArgB109	GluB269
ArgB114	GluB233
HisB117, ValC162, ThrB100, ThrB102, ValC179	ProA329
ArgB129	AspB265
ArgB161	GluB294
ArgC156	GluA333
GluC187	LysA322
ValC161, IleB119	ProA331
ValC159	SerA324
MetB158, PheB196	HisB268
LysB132, ArgB108	SerB267

All residue-residue contacts between the C1q trimer residues and the IgG1-Fc dimer found in the model structure are shown. A contact is defined if the distance between two heavy atoms belonging to pair of residues is < 5Å in the final energy minimized structural model of the complex.

side of the predicted interface (indicated in Figure 6.4, all residue contacts found in the model structure are listed in Table 6.2). The placement of the C1q in complex with the Fc-antibody segment predicts a location on the opposite side of the Fab subunits for the arrangement of the complement C1 molecule (illustrated in Figure 6.5). This corresponds to a sterically ideal arrangement for a productive and simultaneous interaction of antibodies with an antigenic surface and with the C1 complement (Figure 6.5). The model does not exclude the possibility of additional contacts between C1q and the Fab fragments themselves or the linker between Fab fragments and Fc.

6.4 Conclusion

The atomic resolution docking structure of the C1q globular head domain of the complement C1 factor with the IgG1 Fc domain provides a working model for rationalizing available biochemical data on this important interaction. It explains experimental findings based on mutagenesis of putative interface residues and also the fact that only one Pro329/Pro331 motif is sufficient for C1q-Fc interaction. Due to the direct assignment of contacting residues within the C1q-Fc interface of the structural model it offers the possibility to specifically substitute interface residues in order to confirm or to disprove

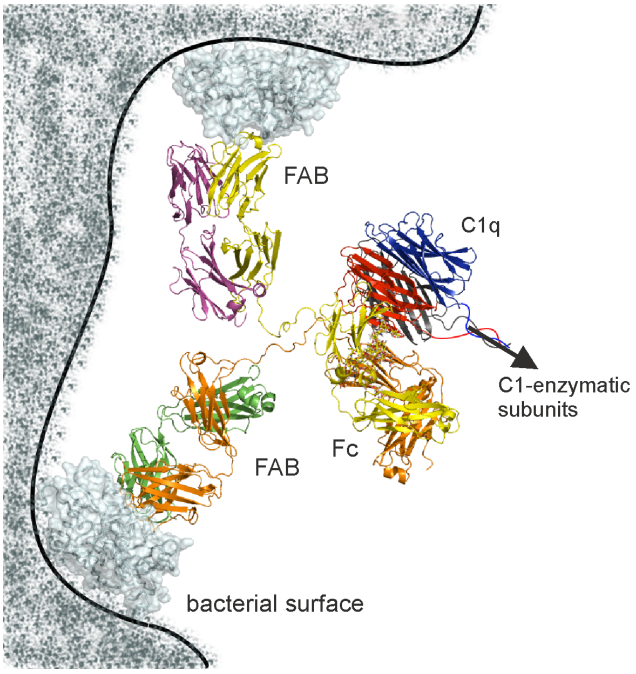


Figure 6.5 – Arrangement of C1q bound to the Fc domain of an IgG antibody in the context of the full antibody structure bound to a bacterial surface.

suggested contacts. It is also sterically fully compatible with the carbohydrate moiety of the Fc fragment and the arrangement of the Fab fragments. In addition to explaining available mutagenesis data, the structural model suggests several intramolecular contacts that could form the basis for the design of new Fc variants with a greater capacity to activate the complement system for example on binding to cancer cells or other target structures. This includes not only residues in the vicinity of the important Pro329/Pro331 motif but also predicts hydrophobic regions on C1q that interact with residue His268 and surrounding amino acids as well as the region around residue Glu333 which is partially buried at the interface. The globular head domain of C1q interacts also with the Fc regions of other antibody types (e.g. IgM). It will be of interest to use the same docking and refinement approach for generating structural models of these interactions in the future.

6.4.1 Acknowledgment

This work has been published in Schneider and Zacharias 2012b

Chapter 7

Conclusion and outlook

7.1 Conclusion

Characterization of the putative binding site of a protein is of key importance in order to investigate the functionality and dynamics of the protein's activity within the living cell. Since most of the biological function of a protein involves the interaction with other biomolecules, transiently or obligatory, the three dimensional structure, as well as the key interactions involved in complex formation, are of great interest for the understanding of biological processes.

All interactions that occur in a cell are embedded in as aqueous environment and water is essential for many processes in the cell. Indirectly, the distinct properties of water in general can be found as driving forces for protein folding and the formation of obligatory protein complexes. Single water molecules have also been found to play a crucial role for the stability of proteins as well as they were found to be a mediator of complex formation. Water molecules within a binding site can shield unfavorable residue-residue contacts and can stabilize transient protein-protein complexes.

In this study, the interaction of proteins with the aqueous environment was investigated using explicit as well as implicit solvent models to account for the direct and indirect interaction of water with biomolecules. Explicit water simulations were used to identify the behavior of water in the binding site compared to the rest of the surface. It was found, that binding sites of the proteins investigated in this study are always, at least partially, overlapped by clusters of unfavorable water hydration sites. An implicit solvent model was used to identify the binding sites for small organic binding partners. Calcu-

lating the energy necessary to replace a water molecule at the surface with a neutral ligand atom enabled differentiating between hydrophobic contacts between protein and ligand and regions of polar or water mediated contacts. In combination with a geometrical cavity detection algorithm, the calculation of these desolvation penalties enabled distinguishing between cavities binding a small ligand and non-binding cavities for a large fraction of proteins and was among the best performing algorithms available.

In experimental studies as well as computer simulations, small organic solvents, such as isopropyl alcohol, have widely been used to identify the binding sites of small organic solvents. The isopropyl alcohol and water molecules account for different chemical contacts available upon complex formation. Isopropyl alcohol can be found at hydrophobic as well as hydrophilic areas on the surface while it avoids polar regions. Water is less often found in hydrophobic areas and can therefore be found in polar regions. In this study, high affinity solvation sites have been used to investigate protein-protein binding sites and in many cases a large amount of solvent could be found in the binding site. Nevertheless, in some protein-protein binding sites neither water nor isopropyl alcohol was found to bind strongly to the binding surface. Additionally, it was found for a set of small organic protein-protein binding inhibitor complexes, that the binding site of the inhibitor was in every case partially filled with clusters of high solvation in the absence of the ligand.

Besides the knowledge of the binding site, knowing the three dimensional arrangement of two binding proteins is of fundamental interest. Often, no three dimensional structure of a protein complex is available, but experiments revealed information on possible contact residues. Many computational approaches include knowledge gained by experiments or, if this information is not available, by bioinformatics binding site prediction methods to dock two or more structures and propose three dimensional models of the unknown complex. Information on interacting residues can be used to restrain the docking procedure to suppose contact residues or to re-evaluate docking results afterwards. Restraint docking approaches do not allow the generation complexes without inclusion of the proposed regions and a re-evaluation only affects the scoring of previously generated structures but does not allow improved sampling of complexes. In this study, a force field based docking procedure (ATTRACT) was optimized for the inclusion of experimental, as well as predicted information, on possible binding sites. Using a force field, instead of restraints

or re-evaluation, overcomes the limitations of the former methods. Increasing the attraction of proposed binding regions does not neglect complexes which do not include the proposed contacts so that near native solutions can be found even if the proposed binding site does not overlap with the correct binding site. If the proposed binding site correlates with the known binding site, additional structures can be sampled or advantageously scored. The protocol presented in this study has shown some remarkable advantages compared to other methods. On a widely used test set, this protocol showed a better performance in terms of sampling and scoring compared to other protein-protein docking methods.

Often mutagenesis experiments are performed for complexes, for which no three dimensional structure can be determined experimentally. Using the previously developed protocol for the inclusion of external information into a docking procedure enabled proposing an all atom three dimensional model of a protein complex of high importance for the activation of the complement system. The main interactions proposed by mutagenesis experiments could also be found as relevant in the computer model. Additionally, further interactions between residues of the partner proteins could be proposed which have so far not been investigated by mutagenesis experiments. Furthermore, this model allows for future investigations to be carried out on the affinity and specificity of the complex and can be the basis for the design of new variants of the partner protein with a greater capacity to activate the complement system.

7.2 Outlook and future work

The state of the methods presented in this thesis is only a snapshot of the development of the methods and therefore future work can improve the presented techniques.

Binding site prediction using explicit solvent simulations

Mapping of the high affinity solvation sites of mixed solvent simulations was in good agreement with the known binding sites. Nevertheless, the test set used for this study was limited in size. For two classes of proteins, enzymes as well as antigens, the method showed a high success rate. In the case of antigens, only two structures were included in this test set, so that a further application

of the method on an extended set of proteins could reveal, whether the success can be related to individual features of the proteins' surface or to a general trend of proteins within this classes.

Among the prediction with the highest affinity, an ATP-binding site as well as the peptide binding site of a MHC molecule was found. The precision of the prediction of the small organic inhibitors binding in comparable cavities as the above mentioned molecules was rather low. However, in any case high solvation sites could be found within the binding site. Many of the binding sites were found to be partially closed or deformed in the sampled conformation. Allowing full flexibility could improve the accessibility of those cavities but might reduce the sampling of solvation sites at exposed and flexible residues, so that a scaling factor might have to be introduced to account for this disproportion.

Binding site prediction using desolvation penalties

The performance of the prediction of small ligand binding sites using desolvation penalties showed to be very successful and accurate in the bound case. However, in the unbound case, the cavities are often narrowed so that a lower amount of probes could be placed to calculate the desolvation penalty. In some cases after the clustering of the probes according to their penalty values the predicted cavities were found to be too small and were discarded. This is also reflected in the lower sensitivity values calculated for the prediction in the unbound case.

To overcome this limitation, two extensions are planed to be introduced. In a first approach, small probes with smaller radii will be used to preferentially describe the cavities. Smaller probe radii result in a smaller difference between highest and lowest penalty values which increases the challenge to discriminate between favorable and unfavorable binding sites. Therefore, overlapping patches of probes will be used instead of single probes to calculate the desolvation penalty and hopefully allow a precise classification of the cavities.

As a second attempt, flexibility will be introduced during the probe placing procedure using normal mode analysis. This allows to place probes on the surface of the deformed structure, which cannot be placed on the unbound structure. Additionally, the rotation of single side-chains in the proximity of the cavities could be introduced to open binding sites.

Weighted protein-protein docking

The test cases with artificially generated binding sites and shifted binding sites have shown, that the success of the docking procedure depends on the quality of the prediction. For the known binding site the success rate was very high and the accuracy decreased for complexes only which undergo a large conformational change upon binding. This shows, that nearly optimal parameters have been found to bias the force field of the ATTRACT docking program towards near native solutions, if the binding site is known or can be predicted precisely. Nevertheless, the performance of this procedure can be further evaluated on larger test sets including more difficult cases or using different binding site predictors.

Additionally, the predictive power of the scoring function of the docking procedure could be further improved in combination with other binding site predictors. This is especially true in the cases of the rotated ligands, for which no continuous predictions could be extracted from the docking results.

Appendix A

List of publications

Below is a list of papers, which were published in journals and books during the PHD as well as a list of papers which will be published in the near future.

Published papers and book chapters included in this thesis

- Sebastian Schneider and Martin Zacharias. Flexible protein-protein docking. *Selected Works in Bioinformatics* (Xuhua Xia), pages 161-176. InTech, 2011.
- Sebastian Schneider and Martin Zacharias. Scoring optimisation of unbound protein-protein docking including protein binding site predictions. *Journal of Molecular Recognition*, 25(1):15-23, 2012.
- Sebastian Schneider and Martin Zacharias. Atomic resolution model of the antibody fc interaction with the complement c1q component. *Molecular Immunology*, 51(1):66-72, 2012.

Published papers and book chapters not included in this thesis

- Simon Leis, Sebastian Schneider, and Martin Zacharias. In silico prediction of binding sites on proteins. *Curr Med Chem*, 17(15):1550-1562, 2010.
- Sebastian Schneider, A. Saladin, S. Fiorucci, C. Prevost, and M. Zacharias: ATTRACT and PTOOLS: Open Source Programs for Protein–Protein Docking, *Methods in molecular biology* (Clifton, NJ), volume 819, Springer, 221, 2012

Prepared to be published in journals

- Sebastian Schneider and Martin Zacharias. Protein binding site prediction using high solvation patterns (working title)
- Sebastian Schneider and Martin Zacharias. Cavity detection using desolvation penalties (working title)

Bibliography

Alessio Amadasi, Francesca Spyrakis, Pietro Cozzini, Donald J Abraham, Glen E Kellogg, and Andrea 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, Apr 2006. doi: 10.1016/j.jmb.2006.01.053. URL http://dx.doi.org/10.1016/j.jmb.2006.01.053.

Alessio Amadasi, J. Andrew Surface, Francesca Spyrakis, Pietro Cozzini, Andrea Mozzarelli, and Glen E Kellogg. Robust classification of "relevant" water molecules in putative protein binding sites. *J Med Chem*, 51(4):1063–1067, Feb 2008. doi: 10.1021/jm701023h. URL http://dx.doi.org/10.1021/jm701023h.

Jianghong An, Maxim Totrov, and Ruben Abagyan. Comprehensive identification of "druggable" protein ligand binding sites. *Genome Inform*, 15(2):31–41, 2004.

Jianghong An, Maxim Totrov, and Ruben Abagyan. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol Cell Proteomics*, 4(6):752–761, Jun 2005. doi: 10.1074/mcp.M400159-MCP200. URL http://dx.doi.org/10.1074/mcp.M400159-MCP200.

Nelly Andrusier, Efrat Mashiach, Ruth Nussinov, and Haim J Wolfson. Principles of flexible protein-protein docking. *Proteins*, 73(2):271–289, Nov 2008. doi: 10.1002/prot.22170. URL http://dx.doi.org/10.1002/prot.22170.

A. Armon, D. Graur, and N. Ben-Tal. Consurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J Mol Biol*, 307(1):447–463, Mar 2001. doi: 10.1006/jmbi.2000.4474. URL http://dx.doi.org/10.1006/jmbi.2000.4474.

R. P. Bahadur and M. Zacharias. The interface of protein-protein complexes: analysis of contacts and prediction of interactions. *Cell Mol Life Sci*, 65(7-8):1059–1072, Apr 2008. doi: 10.1007/s00018-007-7451-x. URL http://dx.doi.org/10.1007/s00018-007-7451-x.

Ranjit Prasad Bahadur, Pinak Chakrabarti, Francis Rodier, and Joël Janin. A dissection of specific and non-specific protein-protein interfaces. *J Mol Biol*, 336(4):943–955, Feb 2004.

I. Bahar, A. R. Atilgan, and B. Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold Des*, 2(3):173–181, 1997.

Ivet Bahar, Chakra Chennubhotla, and Dror Tobi. Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation. *Curr Opin Struct Biol*, 17(6):633–640, Dec 2007. doi: 10.1016/j.sbi.2007.09.011. URL http://dx.doi.org/10.1016/j.sbi.2007.09.011.

Ahmet Bakan and Ivet Bahar. The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. *Proc Natl Acad Sci U S A*, 106(34):14349–14354, Aug 2009. doi: 10.1073/pnas.0904214106. URL http://dx.doi.org/10.1073/pnas.0904214106.

N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci U S A*, 98(18):10037–10041, Aug 2001. doi: 10.1073/pnas.181342398. URL http://dx.doi.org/10.1073/pnas.181342398.

Caterina Barillari, Justine Taylor, Russell Viner, and Jonathan W Essex. Classification of water molecules in protein binding sites. *J Am Chem Soc*, 129(9):2577–2587, Mar 2007. doi: 10.1021/ja066980q. URL http://dx.doi.org/10.1021/ja066980q.

Caterina Barillari, Anna L Duncan, Isaac M Westwood, Julian Blagg, and Rob L M van Montfort. Analysis of water patterns in protein kinase binding sites. *Proteins*, 79(7):2109–2121, Jul 2011. doi: 10.1002/prot.23032. URL http://dx.doi.org/10.1002/prot.23032.

Efrat Ben-Zeev and Miriam Eisenstein. Weighted geometric docking: incorporating external information in the rotation-translation scan. *Proteins*, 52(1):24–27, Jul 2003. doi: 10.1002/prot.10391. URL http://dx.doi.org/10.1002/prot.10391.

Alexander Benedix, Caroline M Becker, Bert L de Groot, Amedeo Caflisch, and Rainer A Böckmann. Predicting free energy changes using structural ensembles. *Nat Methods*, 6(1):3–4, Jan 2009. doi: 10.1038/nmeth0109-3. URL http://dx.doi.org/10.1038/nmeth0109-3.

H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma. The missing term in effective pair potentials. *The Journal of Physical Chemistry*, 91(24):6269–6271, 1987. doi: 10.1021/j100308a038. URL http://pubs.acs.org/doi/abs/10.1021/j100308a038.

H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, and J. Hermans. *Interaction model for water in relation to protein hydration*. B. Pullman, 1981.

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000. doi: 10.1093/nar/28.1.235. URL http://nar.oxfordjournals.org/content/28/1/235.abstract.

Thijs Beuming, Ye Che, Robert Abel, Byungchan Kim, Veerabahu Shanmugasundaram, and Woody Sherman. Thermodynamic analysis of water molecules at the surface of proteins and applications to binding site prediction and characterization. *Proteins: Structure, Function, and Bioinformatics*, pages n/a–n/a, 2011. ISSN 1097-0134. doi: 10.1002/prot.23244. URL http://dx.doi.org/10.1002/prot.23244.

T N Bhat, G A Bentley, G Boulot, M I Greene, D Tello, W Dall'Acqua, H Souchon, F P Schwarz, R A Mariuzza, and R J Poljak. Bound water molecules and conformational stabilization help mediate an antigen-antibody association. *Proceedings of the National Academy of Sciences*, 91(3):1089–1093, 1994. URL http://www.pnas.org/content/91/3/1089.abstract.

A. A. Bogan and K. S. Thorn. Anatomy of hot spots in protein interfaces. *J Mol Biol*, 280(1):1–9, Jul 1998. doi: 10.1006/jmbi.1998.1843. URL http://dx.doi.org/10.1006/jmbi.1998.1843.

Alexandre M J J Bonvin. Flexible protein-protein docking. *Curr Opin Struct Biol*, 16(2):194–200, Apr 2006. doi: 10.1016/j.sbi.2006.02.002. URL http://dx.doi.org/10.1016/j.sbi.2006.02.002.

Andrew J Bordner and Ruben Abagyan. Statistical analysis and prediction of protein-protein interfaces. *Proteins*, 60(3):353–366, Aug 2005. doi: 10.1002/prot.20433. URL http://dx.doi.org/10.1002/prot.20433.

Jenny Bostrom, Shang-Fan Yu, David Kan, Brent A. Appleton, Chingwei V. Lee, Karen Billeci, Wenyan Man, Franklin Peale, Sarajane Ross, Christian Wiesmann, and Germaine Fuh. Variants of the antibody herceptin that interact with her2 and vegf at the antigen binding site. *Science*, 323(5921):1610–1614, 2009. doi: 10.1126/science.1165480. URL http://www.sciencemag.org/content/323/5921/1610.abstract.

James R Bradford and David R Westhead. Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*, 21(8):1487–1494, Apr 2005. doi: 10.1093/bioinformatics/bti242. URL http://dx.doi.org/10.1093/bioinformatics/bti242.

Richard T. Bradshaw, Bhavesh H. Patel, Edward W. Tate, Robin J. Leatherbarrow, and Ian R. Gould. Comparing experimental and computational alanine scanning techniques for probing a prototypical proteinâprotein interaction. *Protein Engineering Design and Selection*, 24(1-2):197–207, 2011. doi: 10.1093/protein/gzq047. URL http://peds.oxfordjournals.org/content/24/1-2/197.abstract.

G. P. Brady and P. F. Stouten. Fast prediction and visualization of protein binding pockets with pass. *J Comput Aided Mol Des*, 14(4):383–401, May 2000.

Ryan Brenke, Dima Kozakov, Gwo-Yu Chuang, Dmitri Beglov, David Hall, Melissa R Landon, Carla Mattos, and Sandor Vajda. Fragment-based identification of druggable 'hot spots' of proteins using fourier domain correlation techniques. *Bioinformatics*, 25(5):621–627, Mar 2009. doi: 10.1093/bioinformatics/btp036. URL http://dx.doi.org/10.1093/bioinformatics/btp036.

Bernard R. Brooks, Robert E. Bruccoleri, Barry D. Olafson, David J. States, S. Swaminathan, and Martin Karplus. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2):187–217, 1983. ISSN 1096-987X. doi: 10.1002/jcc.540040211. URL http://dx.doi.org/10.1002/jcc.540040211.

C. L. Brooks, M. Karplus, and B. M. Pettitt. *Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics*. Wiley, 1990.

Michal Brylinski and Jeffrey Skolnick. A threading-based method (findsite) for ligand-binding site prediction and functional annotation. *Proc Natl Acad Sci U S A*, 105(1):129–134, Jan 2008. doi: 10.1073/pnas.0707684105. URL http://dx.doi.org/10.1073/pnas.0707684105.

Michal Brylinski, Katarzyna Prymula, Wiktor Jurkowski, Marek Kochanczyk, Ewa Stawowczyk, Leszek Konieczny, and Irena Roterman. Prediction of functional sites based on the fuzzy oil drop model. *PLoS Comput Biol*, 3(5):e94, May 2007. doi: 10.1371/journal.pcbi.0030094. URL http://dx.doi.org/10.1371/journal.pcbi.0030094.

Greg Buhrman, Vesna de Serrano, and Carla Mattos. Organic solvents order the dynamic switch ii in ras crystals. *Structure*, 11(7):747–751, Jul 2003.

Nicholas J Burgoyne and Richard M Jackson. Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces. *Bioinformatics*, 22(11):1335–1342, Jun 2006. doi: 10.1093/bioinformatics/btl079. URL http://dx.doi.org/10.1093/bioinformatics/btl079.

D. R. Burton, J. Boyd, A. D. Brampton, S. B. Easterbrook-Smith, E. J. Emanuel, J. Novotny, T. W. Rademacher, M. R. van Schravendijk, M. J. E. Sternberg, and R. A. Dwek. The c1q receptor site on immunoglobulin g. *Nature*, 1980.

Daniel R Caffrey, Shyamal Somaroo, Jason D Hughes, Julian Mintseris, and Enoch S Huang. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci*, 13(1):190–202, Jan 2004. doi: 10.1110/ps.03323604. URL http://dx.doi.org/10.1110/ps.03323604.

Carlos J. Camacho, Zhiping Weng, Sandor Vajda, and Charles DeLisi. Free energy landscapes of encounter complexes in protein-protein association. *Biophysical Journal*, 76(3):1166 – 1178, 1999. ISSN 0006-3495. doi: 10.1016/S0006-3495(99)77281-4. URL http://www.sciencedirect.com/science/article/pii/S0006349599772814.

Stephen J Campbell, Nicola D Gold, Richard M Jackson, and David R Westhead. Ligand binding: functional site location, similarity and docking. *Curr Opin Struct Biol*, 13(3):389–395, Jun 2003.

S. Capaldi, M. Perduca, B. Faggion, M.E. Carrizo, A. Tava, L. Ragona, and H.L. Monaco. Crystal structure of the anticarcinogenic bowman-birk inhibitor from snail medic (medicago scutellata) seeds complexed with bovine trypsin. *Journal of structural biology*, 158(1):71–79, 2007.

Carla and Mattos. Proteinwater interactions in a dynamic world. *Trends in Biochemical Sciences*, 27(4):203 – 208, 2002. ISSN 0968-0004. doi: 10.1016/S0968-0004(02)02067-4. URL http://www.sciencedirect.com/science/article/pii/S0968000402020674.

D.A. Case, T.A. Darden, T.E. Cheatham III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, R.C. Walker, W. Zhang, K.M. Merz, B.P. Roberts, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossvai, K.F. Wong, F. Paesani, J. Vanicek, J. Liu, X. Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D.R. Roe, D.H. Mathews, M.G. Seetin, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, and P.A. Kollman. Amber11, university of california, san francisco. 2010.

Pinak Chakrabarti and Joël Janin. Dissecting protein-protein recognition sites. *Proteins*, 47(3):334–343, May 2002.

D Chandler. Interfaces and the driving force of hydrophobic assembly. *Nature*, 2005.

Sidhartha Chaudhury and Jeffrey J Gray. Conformer selection and induced fit in flexible backbone protein-protein docking using computational and nmr ensembles. *J Mol Biol*, 381(4):1068–1087, Sep 2008. doi: 10.1016/j.jmb.2008.05.042. URL http://dx.doi.org/10.1016/j.jmb.2008.05.042.

Huiling Chen and Huan-Xiang Zhou. Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against nmr data. *Proteins*, 61(1):21–35, Oct 2005. doi: 10.1002/prot.20514. URL http://dx.doi.org/10.1002/prot.20514.

Ke Chen, MarcinÂ J. Mizianty, Jianzhao Gao, and Lukasz Kurgan. A critical comparative assessment of predictions of protein-binding sites for biologically relevant organic compounds. *Structure*, 19(5):613 – 621, 2011. ISSN 0969-2126. doi: 10.1016/j.str.2011.02.015. URL http://www.sciencedirect.com/science/article/pii/S0969212611001079.

Tammy Man-Kuang Cheng, Tom L Blundell, and Juan Fernández-Recio. pydock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. *Proteins*, 68(2):503–515, Aug 2007. doi: 10.1002/prot.21419. URL http://dx.doi.org/10.1002/prot.21419.

Naresh Chennamsetty, Vladimir Voynov, Veysel Kayser, Bernhard Helk, and Bernhardt L Trout. Prediction of protein binding regions. *Proteins*, 79(3):888–897, Mar 2011. doi: 10.1002/prot.22926. URL http://dx.doi.org/10.1002/prot.22926.

T. Clackson and J. A. Wells. A hot spot of binding energy in a hormone-receptor interface. *Science*, 267 (5196):383–386, Jan 1995.

Claudio and Dalvit. Nmr methods in fragment screening: theory and a comparison with other biophysical techniques. *Drug Discovery Today*, 14(21â22):1051 – 1057, 2009. ISSN 1359-6446. doi: 10.1016/j.drudis.2009.07.013. URL http://www.sciencedirect.com/science/article/pii/S1359644609002797.

L. Lo Conte, C. Chothia, and J. Janin. The atomic structure of protein-protein recognition sites. *J Mol Biol*, 285(5):2177–2198, Feb 1999.

Wendy D. Cornell, Piotr Cieplak, Christopher I. Bayly, Ian R. Gould, Kenneth M. Merz, David M. Ferguson, David C. Spellmeyer, Thomas Fox, James W. Caldwell, and Peter A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995. doi: 10.1021/ja00124a002. URL http://pubs.acs.org/doi/abs/10.1021/ja00124a002.

B. C. Cunningham and J. A. Wells. High-resolution epitope mapping of hgh-receptor interactions by alanine-scanning mutagenesis. *Science*, 244(4908):1081–1085, Jun 1989.

Rhiju Das, Ingemar André, Yang Shen, Yibing Wu, Alexander Lemak, Sonal Bansal, Cheryl H Arrowsmith, Thomas Szyperski, and David Baker. Simultaneous prediction of protein folding and docking at high resolution. *Proc Natl Acad Sci U S A*, 106(45):18978–18983, Nov 2009. doi: 10.1073/pnas.0904407106. URL http://dx.doi.org/10.1073/pnas.0904407106.

Andrew M. Davis, Stephen A. St-Gallay, and Gerard J. Kleywegt. Limitations and lessons in the use of x-ray structural information in drug design. *Drug Discovery Today*, 13(19â20):831 – 841, 2008. ISSN 1359-6446. doi: 10.1016/j.drudis.2008.06.006. URL http://www.sciencedirect.com/science/article/pii/S1359644608002183.

B. L. de Groot, D. M. van Aalten, R. M. Scheek, A. Amadei, G. Vriend, and H. J. Berendsen. Prediction of protein conformational freedom from distance constraints. *Proteins*, 29(2):240–251, Oct 1997.

Sjoerd J de Vries and Alexandre M J J Bonvin. How proteins get in touch: interface prediction in the study of biomolecular complexes. *Curr Protein Pept Sci*, 9(4):394–406, Aug 2008.

Sjoerd J de Vries and Alexandre M J J Bonvin. Cport: a consensus interface predictor and its performance in prediction-driven docking with haddock. *PLoS One*, 6(3):e17695, 2011. doi: 10.1371/journal.pone.0017695. URL http://dx.doi.org/10.1371/journal.pone.0017695.

Sjoerd J de Vries, Aalt D J van Dijk, and Alexandre M J J Bonvin. Whiscy: what information does surface conservation yield? application to data-driven docking. *Proteins*, 63(3):479–489, May 2006. doi: 10.1002/prot.20842. URL http://dx.doi.org/10.1002/prot.20842.

Sjoerd J de Vries, Adrien S J Melquiond, Panagiotis L Kastritis, Ezgi Karaca, Annalisa Bordogna, Marc van Dijk, João P G L M Rodrigues, and Alexandre M J J Bonvin. Strengths and weaknesses of data-driven docking in critical assessment of prediction of interactions. *Proteins*, 78(15):3242–3249, Nov 2010. doi: 10.1002/prot.22814. URL http://dx.doi.org/10.1002/prot.22814.

Michelle Dechene, Glenna Wink, Mychal Smith, Paul Swartz, and Carla Mattos. Multiple solvent crystal structures of ribonuclease a: an assessment of the method. *Proteins*, 76(4):861–881, Sep 2009. doi: 10.1002/prot.22393. URL http://dx.doi.org/10.1002/prot.22393.

Sheldon Dennis, Tamas Kortvelyesi, and Sandor Vajda. Computational mapping identifies the binding sites of organic solvents on proteins. *PNAS*, 2002.

Cyril Dominguez, Rolf Boelens, and Alexandre M J J Bonvin. Haddock: a protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc*, 125(7):1731–1737, Feb 2003. doi: 10.1021/ja026939x. URL http://dx.doi.org/10.1021/ja026939x.

Simon Ebbinghaus, Seung Joong Kim, Matthias Heyden, Xin Yu, Udo Heugen, Martin Gruebele, David M Leitner, and Martina Havenith. An extended dynamical hydration shell around proteins. *Proc Natl Acad Sci U S A*, 104(52):20749–20752, Dec 2007. doi: 10.1073/pnas.0709207104. URL http://dx.doi.org/10.1073/pnas.0709207104.

Christian Ehrnthaller, Anita Ignatius, Florian Gebhard, and Markus Huber-Lang. New insights of an old defense system: structure, function, and clinical relevance of the complement system. *Mol Med*, 17(3-4): 317–29, 2011. ISSN 1528-3658. URL http://www.ncbi.nlm.nih.gov/pubmed/21046060.

D Eisenberg, E Schwarz, M Kamaromy, and R Wall. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *Journal of Molecular Biology*, 1984.

Ugur Emekli, Dina Schneidman-Duhovny, Haim J Wolfson, Ruth Nussinov, and Turkan Haliloglu. Hingeprot: automated prediction of hinges in protein structures. *Proteins*, 70(4):1219–1227, Mar 2008. doi: 10.1002/prot.21613. URL http://dx.doi.org/10.1002/prot.21613.

Stefan Engelen, Ladislas A Trojan, Sophie Sacquin-Mora, Richard Lavery, and Alessandra Carbone. Joint evolutionary trees: a large-scale method to predict protein interfaces based on sequence sampling. *PLoS Comput Biol*, 5(1):e1000267, Jan 2009. doi: 10.1371/journal.pcbi.1000267. URL http://dx.doi.org/10.1371/journal.pcbi.1000267.

Andrew C. English, Colin R. Groom, and Roderick E. Hubbard. Experimental and computational mapping of the binding surface of a crystalline protein. *Protein Engineering*, 14(1):47–59, 2001. doi: 10.1093/protein/14.1.47. URL http://peds.oxfordjournals.org/content/14/1/47.abstract.

Iakes Ezkurdia, Lisa Bartoli, Piero Fariselli, Rita Casadio, Alfonso Valencia, and Michael L Tress. Progress and challenges in predicting protein-protein interaction sites. *Brief Bioinform*, 10(3):233–246, May 2009. doi: 10.1093/bib/bbp021. URL http://dx.doi.org/10.1093/bib/bbp021.

Juan Fernández-Recio, Maxim Totrov, and Ruben Abagyan. Icm-disco docking by global energy optimization with fully flexible side-chains. *Proteins*, 52(1):113–117, Jul 2003. doi: 10.1002/prot.10383. URL http://dx.doi.org/10.1002/prot.10383.

Juan Fernández-Recio, Max Totrov, Constantin Skorodumov, and Ruben Abagyan. Optimal docking area: a new method for predicting protein-protein interaction sites. *Proteins*, 58(1):134–143, Jan 2005. doi: 10.1002/prot.20285. URL http://dx.doi.org/10.1002/prot.20285.

Sébastien Fiorucci and Martin Zacharias. Prediction of protein-protein interaction sites using electrostatic desolvation profiles. *Biophys J*, 98(9):1921–1930, May 2010a. doi: 10.1016/j.bpj.2009.12.4332. URL http://dx.doi.org/10.1016/j.bpj.2009.12.4332.

Sébastien Fiorucci and Martin Zacharias. Binding site prediction and improved scoring during flexible protein-protein docking with attract. *Proteins*, 78(15):3131–3139, Nov 2010b. doi: 10.1002/prot.22808. URL http://dx.doi.org/10.1002/prot.22808.

H. Fischer. Einfluß der configuration auf die wirkung der enzyme. *Chemische Berichte*, 1894.

Marshall Fixman. The poissonâboltzmann equation and its application to polyelectrolytes. *J. Chem. Phys*, 1979.

Yoshifumi Fukunishi and Haruki Nakamura. Prediction of ligand-binding sites of proteins by molecular docking calculation for a random ligand library. *Protein Sci*, 20(1):95–106, Jan 2011. doi: 10.1002/pro.540. URL http://dx.doi.org/10.1002/pro.540.

Jonathan C Fuller, Nicholas J Burgoyne, and Richard M Jackson. Predicting druggable binding sites at the protein-protein interface. *Drug Discov Today*, 14(3-4):155–161, Feb 2009. doi: 10.1016/j.drudis.2008.10.009. URL http://dx.doi.org/10.1016/j.drudis.2008.10.009.

Razif R Gabdoulline and Rebecca C Wade. Biomolecular diffusional association. *Curr Opin Struct Biol*, 12 (2):204–213, Apr 2002.

Christine Gaboriaud, Jordi Juanhuix, Arnaud Gruez, Monique Lacroix, Claudine Darnault, David Pignol, Denis Verger, Juan C Fontecilla-Camps, and Gérard J Arlaud. The crystal structure of the globular head of complement protein c1q provides a basis for its versatile recognition properties. *J Biol Chem*, 278(47): 46974–82, November 2003. ISSN 0021-9258. URL http://www.ncbi.nlm.nih.gov/pubmed/12960167.

Christine Gaboriaud, Nicole M Thielens, Lynn A Gregory, Véronique Rossi, Juan C Fontecilla-Camps, and Gérard J Arlaud. Structure and activation of the c1 complex of complement: unraveling the puzzle. *Trends Immunol*, 25(7):368–73, July 2004. ISSN 1471-4906. URL http://www.ncbi.nlm.nih.gov/pubmed/15207504.

Mihaela G Gadjeva, Marieta M Rouseva, Alexandra S Zlatarova, Kenneth B M Reid, Uday Kishore, and Mihaela S Kojouharova. Interaction of human c1q with igg and igm: revisited. *Biochemistry*, 47(49): 13093–102, December 2008. ISSN 1520-4995. URL http://www.ncbi.nlm.nih.gov/pubmed/19006321.

Anne-Claude Gavin, Markus Bösche, Roland Krause, Paola Grandi, Martina Marzioch, Andreas Bauer, Jörg Schultz, Jens M Rick, Anne-Marie Michon, Cristina-Maria Cruciat, Marita Remor, Christian Höfert, Malgorzata Schelder, Miro Brajenovic, Heinz Ruffner, Alejandro Merino, Karin Klein, Manuela Hudak, David Dickson, Tatjana Rudi, Volker Gnau, Angela Bauch, Sonja Bastuck, Bettina Huhse, Christina Leutwein, Marie-Anne Heurtier, Richard R Copley, Angela Edelmann, Erich Querfurth, Vladimir Rybin, Gerard Drewes, Manfred Raida, Tewis Bouwmeester, Peer Bork, Bertrand Seraphin, Bernhard Kuster, Gitte Neubauer, and Giulio Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147, Jan 2002. doi: 10.1038/415141a. URL http://dx.doi.org/10.1038/415141a.

Rohit Ghai, Patrick Waters, Lubka T Roumenina, Mihaela Gadjeva, Mihaela S Kojouharova, Kenneth B M Reid, Robert B Sim, and Uday Kishore. C1q and its growing family. *Immunobiology*, 212(4-5):253–66, 2007. ISSN 0171-2985. URL http://www.ncbi.nlm.nih.gov/pubmed/17544811.

Avijit Ghosh, Chaya S. Rapp, and Richard A. Friesner. Generalized Born Model Based on a Surface Integral Formulation. *The Journal of Physical Chemistry B*, 102(52):10983–10990, December 1998. doi: 10.1021/jp982533o. URL http://dx.doi.org/10.1021/jp982533o.

P. H. Giangrande, E. A. Kimbrel, D. P. Edwards, and D. P. McDonnell. The opposing transcriptional activities of the two isoforms of the human progesterone receptor are due to differential cofactor binding. *Mol Cell Biol*, 20(9):3102–3115, May 2000.

R. E. Gillia and K. R. William. Shading, rare events and rubber bands â a variational verlet algorithm for molecular dynamics. *JCP*, 1992.

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, May 2001.

Fabian Glaser, Tal Pupko, Inbal Paz, Rachel E Bell, Dalit Bechor-Shental, Eric Martz, and Nir Ben-Tal. Consurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, 19(1):163–164, Jan 2003.

Kay-Eberhard Gottschalk, Hani Neuvirth, and Gideon Schreiber. A novel method for scoring of docked protein complexes using predicted protein-protein binding sites. *Protein Eng Des Sel*, 17(2):183–189, Feb 2004. doi: 10.1093/protein/gzh021. URL http://dx.doi.org/10.1093/protein/gzh021.

Jeffrey J Gray, Stewart E Moughon, Tanja Kortemme, Ora Schueler-Furman, Kira M S Misura, Alexandre V Morozov, and David Baker. Protein-protein docking predictions for the capri experiment. *Proteins*, 52 (1):118–122, Jul 2003. doi: 10.1002/prot.10384. URL http://dx.doi.org/10.1002/prot.10384.

Raik Grünberg, Johan Leckner, and Michael Nilges. Complementarity of structure ensembles in protein-protein binding. *Structure*, 12(12):2125–2136, Dec 2004. doi: 10.1016/j.str.2004.09.014. URL http://dx.doi.org/10.1016/j.str.2004.09.014.

Philip J Hajduk and Jonathan Greer. A decade of fragment-based drug design: strategic advances and lessons learned. *Nat Rev Drug Discov*, 6(3):211–219, Mar 2007. doi: 10.1038/nrd2220. URL http://dx.doi.org/10.1038/nrd2220.

Tom Halgren. New method for fast and accurate binding-site identification and analysis. *Chem Biol Drug Des*, 69(2):146–148, Feb 2007. doi: 10.1111/j.1747-0285.2007.00483.x. URL http://dx.doi.org/10.1111/j.1747-0285.2007.00483.x.

Inbal Halperin, Buyong Ma, Haim Wolfson, and Ruth Nussinov. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins*, 47(4):409–443, Jun 2002. doi: 10.1002/prot.10115. URL http://dx.doi.org/10.1002/prot.10115.

Bing Hao, Ning Zheng, Brenda A. Schulman, Geng Wu, Julie J. Miller, Michele Pagano, and Nikola P. Pavletich. Structural basis of the cks1-dependent recognition of p27kip1 by the scfskp2 ubiquitin ligase. *Molecular Cell*, 20(1):9 – 19, 2005. ISSN 1097-2765. doi: 10.1016/j.molcel.2005.09.003. URL http://www.sciencedirect.com/science/article/pii/S1097276505016035.

G. D. Hawkins, C. J. Cramer, and D. G. Truhlar. Pairwise solute descreening of solute charges from a dielectric medium. *Chemical Physics Letters*, 1995.

Richard H Henchman and J. Andrew McCammon. Structural and dynamic properties of water around acetylcholinesterase. *Protein Sci*, 11(9):2080–2090, Sep 2002a. doi: 10.1110/ps.0214002. URL http://dx.doi.org/10.1110/ps.0214002.

Richard H Henchman and J. Andrew McCammon. Extracting hydration sites around proteins from explicit water simulations. *J Comput Chem*, 23(9):861–869, Jul 2002b. doi: 10.1002/jcc.10074. URL http://dx.doi.org/10.1002/jcc.10074.

M. Hendlich, F. Rippmann, and G. Barnickel. Ligsite: automatic and efficient detection of potential small molecule-binding sites in proteins. *J Mol Graph Model*, 15(6):359–63, 389, Dec 1997.

Stefan Henrich, Outi M H Salo-Ahen, Bingding Huang, Friedrich F Rippmann, Gabriele Cruciani, and Rebecca C Wade. Computational approaches to identifying and characterizing protein binding sites for ligand design. *J Mol Recognit*, 23(2):209–219, 2010. doi: 10.1002/jmr.984. URL http://dx.doi.org/10.1002/jmr.984.

B. Hess, H. Bekker, H.J.C. Berendsen, and J.G.E.M. Fraaije. Lincs: A linear constraint solver for molecular simulations. *J. Comp. Chemistry*, 1997.

William C. Ho, Cheng Luo, Kehao Zhao, Xiaomei Chai, Mary X. Fitzgerald, and Ronen Marmorstein. High-resolution structure of the p53 core domain: implications for binding small-molecule stabilizing compounds. *Acta Crystallographica Section D*, 62(12):1484–1493, Dec 2006. doi: 10.1107/S090744490603890X. URL http://dx.doi.org/10.1107/S090744490603890X.

Bingding Huang and Michael Schroeder. Ligsitecsc: predicting ligand binding sites using the connolly surface and degree of conservation. *BMC Struct Biol*, 6:19, 2006. doi: 10.1186/1472-6807-6-19. URL http://dx.doi.org/10.1186/1472-6807-6-19.

Bingding Huang and Michael Schroeder. Using protein binding site prediction to improve protein docking. *Gene*, 422(1-2):14–21, Oct 2008. doi: 10.1016/j.gene.2008.06.014. URL http://dx.doi.org/10.1016/j.gene.2008.06.014.

Philippe Hünenberger, Vincent Krutler, and Wilfired F. van Gusteren. A fast shake algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *Journal of Computational Chemistry*, 2001.

E E Idusogie, L G Presta, H Gazzano-Santoro, K Totpal, P Y Wong, M Ultsch, Y G Meng, and M G Mulkerrin. Mapping of the c1q binding site on rituxan, a chimeric antibody with a human igg1 fc. *J Immunol*, 164(8):4178–84, April 2000. ISSN 0022-1767. URL http://www.ncbi.nlm.nih.gov/pubmed/10754313.

E E Idusogie, P Y Wong, L G Presta, H Gazzano-Santoro, K Totpal, M Ultsch, and M G Mulkerrin. Engineered antibodies with increased activity to recruit complement. *J Immunol*, 166(4):2571–5, February 2001. ISSN 0022-1767. URL http://www.ncbi.nlm.nih.gov/pubmed/11160318.

Wonpil Im, Dmitrii Beglov, and BenoÃ®t Roux. Continuum solvation model: Computation of electrostatic forces from numerical solutions to the poisson-boltzmann equation. *Computer Physics Communications*, 111(1-3):59 – 75, 1998. ISSN 0010-4655. doi: 10.1016/S0010-4655(98)00016-2. URL http://www.sciencedirect.com/science/article/pii/S0010465598000162.

Joël Janin and Bertrand Seraphin. Genome-wide studies of protein-protein interaction. *Curr Opin Struct Biol*, 13(3):383–388, Jun 2003.

JoÃ«l Janin. Wet and dry interfaces: the role of solvent in proteinâprotein and proteinâdna recognition. *Structure*, 7(12):R277 – R279, 1999. ISSN 0969-2126. doi: 10.1016/S0969-2126(00)88333-1. URL http://www.sciencedirect.com/science/article/pii/S0969212600883331.

Abhishek K. Jha, Andres Colubri, Muhammad H. Zaman, Shohei Koide, Tobin R. Sosnick, and Karl F. Freed. Helix, sheet, and polyproline ii frequencies and strong nearest neighbor effects in a restricted coil libraryâ . *Biochemistry*, 44(28):9691–9702, 2005. doi: 10.1021/bi0474822. URL http://pubs.acs.org/doi/abs/10.1021/bi0474822. PMID: 16008354.

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

S. Jones and J. M. Thornton. Prediction of protein-protein interaction sites using patch analysis. *J Mol Biol*, 272(1):133–143, Sep 1997a. doi: 10.1006/jmbi.1997.1233. URL http://dx.doi.org/10.1006/jmbi.1997.1233.

S. Jones and J. M. Thornton. Analysis of protein-protein interaction sites using surface patches. *J Mol Biol*, 272(1):121–132, Sep 1997b. doi: 10.1006/jmbi.1997.1234. URL http://dx.doi.org/10.1006/jmbi.1997.1234.

William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. Comparison of simple potential functions for simulating liquid water. 79(2):926–935, 1983. ISSN 00219606. doi: DOI:10.1063/1.445869. URL http://dx.doi.org/10.1063/1.445869.

William L. Jorgensen, David S. Maxwell, and Julian Tirado-Rives. Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, 118(45):11225–11236, 1996. doi: 10.1021/ja9621760. URL http://pubs.acs.org/doi/abs/10.1021/ja9621760.

Martin Karplus. Molecular dynamics simulations of biomolecules. *Accounts of Chemical Research*, 35(6): 321–323, 2002. doi: 10.1021/ar020082r. URL http://pubs.acs.org/doi/abs/10.1021/ar020082r.

E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A. A. Friesem, C. Aflalo, and I. A. Vakser. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc Natl Acad Sci U S A*, 89(6):2195–2199, Mar 1992.

M Kaul and M Loos. Dissection of c1q capability of interacting with igg. time-dependent formation of a tight and only partly reversible association. *J Biol Chem*, 272(52):33234–44, December 1997. ISSN 0021-9258. URL http://www.ncbi.nlm.nih.gov/pubmed/9407113.

Takeshi Kawabata and Nobuhiro Go. Detection of pockets on protein surfaces using small and large probe spheres to find putative ligand binding sites. *Proteins: Structure, Function, and Bioinformatics*, 68(2): 516–529, 2007. ISSN 1097-0134. doi: 10.1002/prot.21283. URL http://dx.doi.org/10.1002/prot.21283.

Barbara J. Kelly, Branka Mijatov, Cornel Fraefel, Anthony L. Cunningham, and Russell J. Diefenbach. Identification of a single amino acid residue which is critical for the interaction between hsv-1 inner tegument proteins pul36 and pul37. *Virology*, 422(2):308 – 316, 2012. ISSN 0042-6822. doi: 10.1016/j.virol.2011.11.002. URL http://www.sciencedirect.com/science/article/pii/S0042682211005204.

Ozlem Keskin, Buyong Ma, and Ruth Nussinov. Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. *J Mol Biol*, 345(5):1281–1294, Feb 2005. doi: 10.1016/j.jmb.2004.10.077. URL http://dx.doi.org/10.1016/j.jmb.2004.10.077.

Ozlem Keskin, Attila Gursoy, Buyong Ma, and Ruth Nussinov. Principles of protein-protein interactions: what are the preferred ways for proteins to interact? *Chem Rev*, 108(4):1225–1244, Apr 2008. doi: 10.1021/cr040409x. URL http://dx.doi.org/10.1021/cr040409x.

Uday Kishore, Mihaela S Kojouharova, and Kenneth B M Reid. Recent progress in the understanding of the structure-function relationships of the globular head regions of c1q. *Immunobiology*, 205(4-5):355–64, September 2002. ISSN 0171-2985. URL http://www.ncbi.nlm.nih.gov/pubmed/12395999.

Uday Kishore, Rohit Ghai, Trevor J Greenhough, Annette K Shrive, Domenico M Bonifati, Mihaela G Gadjeva, Patrick Waters, Mihaela S Kojouharova, Trinad Chakraborty, and Alok Agrawal. Structural and functional anatomy of the globular domain of complement protein c1q. *Immunol Lett*, 95(2):113–28, September 2004. ISSN 0165-2478. URL http://www.ncbi.nlm.nih.gov/pubmed/15388251.

Isaac Klapper, Ray Hagstrom, Richard Fine, Kim Sharp, and Barry Honig. Focusing of electric fields in the active site of cu-zn superoxide dismutase: Effects of ionic strength and amino-acid modification. *Proteins: Structure, Function, and Bioinformatics*, 1(1):47–59, 1986. ISSN 1097-0134. doi: 10.1002/prot.340010109. URL http://dx.doi.org/10.1002/prot.340010109.

Brian Kobilka and Gebhard F.X. Schertler. New g-protein-coupled receptor crystal structures: insights and limitations. *Trends in Pharmacological Sciences*, 29(2):79 – 83, 2008. ISSN 0165-6147. doi: 10.1016/j.tips.2007.11.009. URL http://www.sciencedirect.com/science/article/pii/S0165614708000084.

Mihaela S Kojouharova, Ivanka G Tsacheva, Magdalena I Tchorbadjieva, Kenneth B M Reid, and Uday Kishore. Localization of ligand-binding sites on human c1q globular head region using recombinant globular head fragments and single-chain antibodies. *Biochim Biophys Acta*, 1652(1):64–74, November 2003. ISSN 0006-3002. URL http://www.ncbi.nlm.nih.gov/pubmed/14580997.

Mihaela S Kojouharova, Mihaela G Gadjeva, Ivanka G Tsacheva, Aleksandra Zlatarova, Liubka T Roumenina, Magdalena I Tchorbadjieva, Boris P Atanasov, Patrick Waters, Britta C Urban, Robert B Sim, Kenneth B M Reid, and Uday Kishore. Mutational analyses of the recombinant globular regions of human c1q a, b, and c chains suggest an essential role for arginine and histidine residues in the c1q-igg interaction. *J Immunol*, 172(7):4351–8, April 2004. ISSN 0022-1767. URL http://www.ncbi.nlm.nih.gov/pubmed/15034050.

Tanja Kortemme and David Baker. A simple physical model for binding energy hot spots in protein-protein complexes. *Proc Natl Acad Sci U S A*, 99(22):14116–14121, Oct 2002. doi: 10.1073/pnas.202485799. URL http://dx.doi.org/10.1073/pnas.202485799.

D. E. Koshland. Application of a theory of enzyme specificity to protein synthesis. *Proc Natl Acad Sci U S A*, 44(2):98–104, Feb 1958.

Noga Kowalsman and Miriam Eisenstein. Combining interface core and whole interface descriptors in postscan processing of protein-protein docking models. *Proteins*, 77(2):297–318, Nov 2009. doi: 10.1002/prot.22436. URL http://dx.doi.org/10.1002/prot.22436.

Dima Kozakov, David R Hall, Gwo-Yu Chuang, Regina Cencic, Ryan Brenke, Laurie E Grove, Dmitri Beglov, Jerry Pelletier, Adrian Whitty, and Sandor Vajda. Structural conservation of druggable hot spots in protein-protein interfaces. *Proc Natl Acad Sci U S A*, 108(33):13528–13533, Aug 2011. doi: 10.1073/pnas.1101835108. URL http://dx.doi.org/10.1073/pnas.1101835108.

Marcin Krol, Raphael A G Chaleil, Alexander L Tournier, and Paul A Bates. Implicit flexibility in protein docking: cross-docking and local refinement. *Proteins*, 69(4):750–757, Dec 2007. doi: 10.1002/prot.21698. URL http://dx.doi.org/10.1002/prot.21698.

Irina Kufareva, Levon Budagyan, Eugene Raush, Maxim Totrov, and Ruben Abagyan. Pier: protein interface recognition for structural proteomics. *Proteins*, 67(2):400–417, May 2007. doi: 10.1002/prot.21233. URL http://dx.doi.org/10.1002/prot.21233.

I. D. Kuntz, K. Chen, K. A. Sharp, and P. A. Kollman. The maximal affinity of ligands. *Proceedings of the National Academy of Sciences*, 96(18):9997–10002, 1999. doi: 10.1073/pnas.96.18.9997. URL http://www.pnas.org/content/96/18/9997.abstract.

David La and Daisuke Kihara. A novel method for proteinâprotein interaction site prediction using phylogenetic substitution models. *Proteins: Structure, Function, and Bioinformatics*, 80(1):126–141, 2012. ISSN 1097-0134. doi: 10.1002/prot.23169. URL http://dx.doi.org/10.1002/prot.23169.

John E. Ladbury. Just add water! the effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. *Chemistry & Biology*, 3(12):973 – 980, 1996. ISSN 1074-5521. doi: 10.1016/S1074-5521(96)90164-7. URL http://www.sciencedirect.com/science/article/pii/S1074552196901647.

R. A. Laskowski. Surfnet: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph*, 13(5):323–30, 307–8, Oct 1995.

R. A. Laskowski, N. M. Luscombe, M. B. Swindells, and J. M. Thornton. Protein clefts in molecular recognition and function. *Protein Sci*, 5(12):2438–2452, Dec 1996. doi: 10.1002/pro.5560051206. URL http://dx.doi.org/10.1002/pro.5560051206.

Alasdair T R Laurie and Richard M Jackson. Q-sitefinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics*, 21(9):1908–1916, May 2005. doi: 10.1093/bioinformatics/bti315. URL http://dx.doi.org/10.1093/bioinformatics/bti315.

Song Hi Lee and Peter J. Rossky. A comparison of the structure and dynamics of liquid water at hydrophobic and hydrophilic surfaces—a molecular dynamics simulation study. *The Journal of Chemical Physics*, 100 (4):3334–3345, 1994. doi: 10.1063/1.466425. URL http://link.aip.org/link/?JCP/100/3334/1.

Simon Leis, Sebastian Schneider, and Martin Zacharias. In silico prediction of binding sites on proteins. *Curr Med Chem*, 17(15):1550–1562, 2010.

Marc F Lensink, Raúl Méndez, and Shoshana J Wodak. Docking and scoring protein complexes: Capri 3rd edition. *Proteins*, 69(4):704–718, Dec 2007. doi: 10.1002/prot.21804. URL http://dx.doi.org/10.1002/prot.21804.

D. G. Levitt and L. J. Banaszak. Pocket: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J Mol Graph*, 10(4):229–234, Dec 1992.

Katrina W Lexa and Heather A Carlson. Full protein flexibility is essential for proper hot-spot mapping. *J Am Chem Soc*, Dec 2010. doi: 10.1021/ja1079332. URL http://dx.doi.org/10.1021/ja1079332.

J. Liang, H. Edelsbrunner, and C. Woodward. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci*, 7(9):1884–1897, Sep 1998. doi: 10.1002/pro.5560070905. URL http://dx.doi.org/10.1002/pro.5560070905.

Shide Liang, Chi Zhang, Song Liu, and Yaoqi Zhou. Protein binding site prediction using an empirical scoring function. *Nucleic Acids Res*, 34(13):3698–3707, 2006. doi: 10.1093/nar/gkl454. URL http://dx.doi.org/10.1093/nar/gkl454.

Shide Liang, Samy O Meroueh, Guangce Wang, Chao Qiu, and Yaoqi Zhou. Consensus scoring for enriching near-native structures from protein-protein docking decoys. *Proteins*, 75(2):397–403, May 2009. doi: 10.1002/prot.22252. URL http://dx.doi.org/10.1002/prot.22252.

O. Lichtarge, H. R. Bourne, and F. E. Cohen. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol*, 257(2):342–358, Mar 1996. doi: 10.1006/jmbi.1996.0167. URL http://dx.doi.org/10.1006/jmbi.1996.0167.

Olivier Lichtarge and Mathew E Sowa. Evolutionary predictions of binding surfaces and interactions. *Curr Opin Struct Biol*, 12(1):21–27, Feb 2002.

Bin Lin and B. Montgomery Pettitt. Note: On the universality of proximal radial distribution functions of proteins. *The Journal of Chemical Physics*, 134(10):106101, 2011. doi: 10.1063/1.3565035. URL http://link.aip.org/link/?JCP/134/106101/1.

Jin Hua Lu, Boon King Teh, Lin da Wang, Yi Nan Wang, Yen Seah Tan, Min Chern Lai, and Kenneth B M Reid. The classical and regulatory functions of c1q in immunity and autoimmunity. *Cell Mol Immunol*, 5(1):9–21, February 2008. ISSN 1672-7681. URL http://www.ncbi.nlm.nih.gov/pubmed/18318990.

James Luccarelli, Julien Michel, Julian Tirado-Rives, and William L Jorgensen. Effects of water placement on predictions of binding affinities for p38α map kinase inhibitors. *J Chem Theory Comput*, 6(12):3850–3856, January 2010. ISSN 1549-9626. URL http://www.ncbi.nlm.nih.gov/pubmed/21278915.

Bhupinder Madan and Kim Sharp. Changes in water structure induced by a hydrophobic solute probed by simulation of the water hydrogen bond angle and radial distribution functions. *Biophysical Chemistry*, 78(1â2):33 – 41, 1999. ISSN 0301-4622. doi: 10.1016/S0301-4622(98)00227-0. URL http://www.sciencedirect.com/science/article/pii/S0301462298002270.

Michael W. Mahoney and William L. Jorgensen. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. 112(20):8910–8922, 2000. ISSN 00219606. doi: DOI:10.1063/1.481505. URL http://dx.doi.org/10.1063/1.481505.

V. A. Makarov, M. Feig, B. K. Andrews, and B. M. Pettitt. Diffusion of solvent around biomolecular solutes: a molecular dynamics simulation study. *Biophys J*, 75(1):150–158, Jul 1998a. doi: 10.1016/S0006-3495(98)77502-2. URL http://dx.doi.org/10.1016/S0006-3495(98)77502-2.

Vladimir A. Makarov, B. Kim Andrews, and B. Montgomery Pettitt. Reconstructing the proteinâwater interface. *Biopolymers*, 45(7):469–478, 1998b. ISSN 1097-0282. doi: 10.1002/(SICI)1097-0282(199806)45:7<469::AID-BIP1>3.0.CO;2-M. URL http://dx.doi.org/10.1002/(SICI)1097-0282(199806)45:7<469::AID-BIP1>3.0.CO;2-M.

Efrat Mashiach, Ruth Nussinov, and Haim J Wolfson. Fiberdock: a web server for flexible induced-fit backbone refinement in molecular docking. *Nucleic Acids Res*, 38(Web Server issue):W457–W461, Jul 2010. doi: 10.1093/nar/gkq373. URL http://dx.doi.org/10.1093/nar/gkq373.

C. Mattos and D. Ringe. Locating and characterizing binding sites on proteins. *Nat Biotechnol*, 14(5): 595–599, May 1996. doi: 10.1038/nbt0596-595. URL http://dx.doi.org/10.1038/nbt0596-595.

C. Mattos and D. Ringe. Proteins in organic solvents. *Curr Opin Struct Biol*, 11(6):761–764, Dec 2001.

Carla Mattos, Cornelia R. Bellamacina, Ezra Peisach, Antonio Pereira, Dennis Vitkup, Gregory A. Petsko, and Dagmar Ringe. Multiple solvent crystal structures: Probing binding sites, plasticity and hydration. *Journal of Molecular Biology*, 357(5):1471 – 1482, 2006. ISSN 0022-2836. doi: 10.1016/j.jmb.2006.01.039. URL http://www.sciencedirect.com/science/article/pii/S0022283606000672.

Andreas May and Martin Zacharias. Accounting for global protein deformability during protein-protein and protein-ligand docking. *Biochim Biophys Acta*, 1754(1-2):225–231, Dec 2005. doi: 10.1016/j.bbapap.2005.07.045. URL http://dx.doi.org/10.1016/j.bbapap.2005.07.045.

Andreas May and Martin Zacharias. Energy minimization in low-frequency normal modes to efficiently allow for global flexibility during systematic protein-protein docking. *Proteins*, 70(3):794–809, Feb 2008. doi: 10.1002/prot.21579. URL http://dx.doi.org/10.1002/prot.21579.

Adrien S.j. Melquiond and Alexandre M.J.J Bonvin. *Protein-Protein Complexes: Analysis, Modeling and Drug Design*, chapter 7: Data-driven Docking: Using External Information to Spark the biomolecular Rendez-vous. Zacharias, Martin, 2010.

Franci Merzel and Jeremy C Smith. High-density hydration layer of lysozymes: molecular dynamics decomposition of solution scattering data. *J Chem Inf Model*, 45(6):1593–1599, 2005. doi: 10.1021/ci0502000. URL http://dx.doi.org/10.1021/ci0502000.

Alexander Metz, Christopher Pfleger, Hannes Kopitz, Stefania Pfeiffer-Marek, Karl-Heinz Baringhaus, and Holger Gohlke. Hot spots and transient pockets: Predicting the determinants of small-molecule binding to a proteinâprotein interface. *Journal of Chemical Information and Modeling*, just accepted(0):null, 2011. doi: 10.1021/ci200322s. URL http://pubs.acs.org/doi/abs/10.1021/ci200322s.

Terje E Michaelsen, John E Thommesen, Oistein Ihle, Tone F Gregers, Randi H Sandin, Ole Henrik Brekke, and Inger Sandlie. A mutant human igg molecule with only one c1q binding site can activate complement and induce lysis of target cells. *Eur J Immunol*, 36(1):129–38, January 2006. ISSN 0014-2980. URL http://www.ncbi.nlm.nih.gov/pubmed/16323243.

Julien Michel, Julian Tirado-Rives, and William L Jorgensen. Energetics of displacing water molecules from protein binding sites: consequences for ligand optimization. *J Am Chem Soc*, 131(42):15403–11, October 2009. ISSN 1520-5126. URL http://www.ncbi.nlm.nih.gov/pubmed/19778066.

I. Mihalek, I. Res, and O. Lichtarge. A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J Mol Biol*, 336(5):1265–1282, Mar 2004. doi: 10.1016/j.jmb.2003.12.078. URL http://dx.doi.org/10.1016/j.jmb.2003.12.078.

D. W. Miller and K. A. Dill. Ligand binding to proteins: the binding landscape model. *Protein Sci*, 6(10): 2166–2179, Oct 1997. doi: 10.1002/pro.5560061011. URL http://dx.doi.org/10.1002/pro.5560061011.

Julian Mintseris, Kevin Wiehe, Brian Pierce, Robert Anderson, Rong Chen, Joël Janin, and Zhiping Weng. Protein-protein docking benchmark 2.0: an update. *Proteins*, 60(2):214–216, Aug 2005. doi: 10.1002/prot.20560. URL http://dx.doi.org/10.1002/prot.20560.

Pierre Monsan and Didier Combes. Effect of water activity on enzyme action and stability. *Annals of the New York Academy of Sciences*, 434(1):048–060, 1984. ISSN 1749-6632. doi: 10.1111/j.1749-6632.1984.tb29799.x. URL http://dx.doi.org/10.1111/j.1749-6632.1984.tb29799.x.

Gregory L Moore, Hsing Chen, Sher Karki, and Greg A Lazar. Engineered fc variant antibodies with enhanced ability to recruit complement and mediate effector functions. *MAbs*, 2(2):181–9, 2010. ISSN 1942-0870. URL http://www.ncbi.nlm.nih.gov/pubmed/20150767.

Irina S Moreira, Pedro A Fernandes, and Maria J Ramos. Hot spots–a review of the protein-protein interface determinant amino-acid residues. *Proteins*, 68(4):803–812, Sep 2007a. doi: 10.1002/prot.21396. URL http://dx.doi.org/10.1002/prot.21396.

Irina S. Moreira, Pedro A. Fernandes, and Maria J. Ramos. Computational alanine scanning mutagenesisâan improved methodological approach. *Journal of Computational Chemistry*, 28(3):644–654, 2007b. ISSN 1096-987X. doi: 10.1002/jcc.20566. URL http://dx.doi.org/10.1002/jcc.20566.

Irina S Moreira, Pedro A Fernandes, and Maria J Ramos. Protein-protein docking dealing with the unknown. *J Comput Chem*, 31(2):317–342, Jan 2010. doi: 10.1002/jcc.21276. URL http://dx.doi.org/10.1002/jcc.21276.

Daniel H Morgan, David M Kristensen, David Mittelman, and Olivier Lichtarge. Et viewer: an application for predicting and visualizing functional sites in protein structures. *Bioinformatics*, 22(16):2049–2050, Aug 2006. doi: 10.1093/bioinformatics/btl285. URL http://dx.doi.org/10.1093/bioinformatics/btl285.

Diana Mustard and David W Ritchie. Docking essential dynamics eigenstructures. *Proteins*, 60(2):269–274, Aug 2005. doi: 10.1002/prot.20569. URL http://dx.doi.org/10.1002/prot.20569.

H. G. Nagendra, N. Sukumar, and M. Vijayan. Role of water in plasticity, stability, and action of proteins: the crystal structures of lysozyme at very low levels of hydration. *Proteins*, 32(2):229–240, Aug 1998.

Murad Nayal and Barry Honig. On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins*, 63(4):892–906, Jun 2006. doi: 10.1002/prot.20897. URL http://dx.doi.org/10.1002/prot.20897.

Hani Neuvirth, Ran Raz, and Gideon Schreiber. Promate: a structure based prediction program to identify the location of protein-protein binding sites. *J Mol Biol*, 338(1):181–199, Apr 2004. doi: 10.1016/j.jmb.2004.02.040. URL http://dx.doi.org/10.1016/j.jmb.2004.02.040.

R. Norel, D. Fischer, H. J. Wolfson, and R. Nussinov. Molecular surface recognition by a computer vision-based technique. *Protein Eng*, 7(1):39–46, Jan 1994.

Vaheh Oganesyan, Changshou Gao, Lena Shirinian, Herren Wu, and William F Dall'Acqua. Structural characterization of a human fc fragment engineered for lack of effector functions. *Acta Crystallogr D Biol Crystallogr*, 64(Pt 6):700–4, June 2008. ISSN 0907-4449. URL http://www.ncbi.nlm.nih.gov/pubmed/18560159.

Alexey Onufriev. *Continuum Electrostatics Solvent Modeling with the Generalized Born Model*, pages 127–165. Wiley-VCH Verlag GmbH & Co. KGaA, 2010. ISBN 9783527629251. doi: 10.1002/9783527629251.ch6. URL http://dx.doi.org/10.1002/9783527629251.ch6.

Alexey Onufriev, Donald Bashford, and David A. Case. Modification of the generalized born model suitable for macromolecules. *The Journal of Physical Chemistry B*, 104(15):3712–3720, 2000. doi: 10.1021/jp994072s. URL http://pubs.acs.org/doi/abs/10.1021/jp994072s.

Alexey Onufriev, David A Case, and Donald Bashford. Effective born radii in the generalized born approximation: the importance of being perfect. *J Comput Chem*, 23(14):1297–1304, Nov 2002. doi: 10.1002/jcc.10126. URL http://dx.doi.org/10.1002/jcc.10126.

C Oubridge, N Ito, P R Evans, C H Teo, and K Nagai. Crystal structure at 1.92 a resolution of the rna-binding domain of the u1a spliceosomal protein complexed with an rna hairpin. *Nature*, 372(6505):432–8, December 1994. ISSN 0028-0836. URL http://www.ncbi.nlm.nih.gov/pubmed/7984237.

Samir Kumar Pal and Ahmed Zewail. Dynamics of water in biological recognition. *Chem. Rev.*, 2004.

Sheldon Park and Jeffery G. Saven. Statistical and molecular dynamics studies of buried waters in globular proteins. *Proteins: Structure, Function, and Bioinformatics*, 60(3):450–463, 2005. ISSN 1097-0134. doi: 10.1002/prot.20511. URL http://dx.doi.org/10.1002/prot.20511.

Brian Pierce and Zhiping Weng. Zrank: reranking protein docking predictions with an optimized energy function. *Proteins*, 67(4):1078–1086, Jun 2007. doi: 10.1002/prot.21373. URL http://dx.doi.org/10.1002/prot.21373.

Brian Pierce and Zhiping Weng. A combination of rescoring and refinement significantly improves protein docking performance. *Proteins*, 72(1):270–279, Jul 2008. doi: 10.1002/prot.21920. URL http://dx.doi.org/10.1002/prot.21920.

Francesco Pizzitutti, Massimo Marchi, Fabio Sterpone, and Peter J Rossky. How protein surfaces induce anomalous dynamics of hydration water. *J Phys Chem B*, 111(26):7584–7590, Jul 2007. doi: 10.1021/jp0717185. URL http://dx.doi.org/10.1021/jp0717185.

Carles Pons, Solène Grosdidier, Albert Solernou, Laura Pérez-Cano, and Juan Fernández-Recio. Present and future challenges and limitations in protein-protein docking. *Proteins*, 78(1):95–108, Jan 2010. doi: 10.1002/prot.22564. URL http://dx.doi.org/10.1002/prot.22564.

Aleksey Porollo and JarosÅaw Meller. Prediction-based fingerprints of protein-protein interactions. *Proteins*, 66(3):630–645, Feb 2007. doi: 10.1002/prot.21248. URL http://dx.doi.org/10.1002/prot.21248.

Leonard G Presta. Molecular engineering and design of therapeutic antibodies. *Curr Opin Immunol*, 20(4): 460–70, August 2008. ISSN 0952-7915. URL http://www.ncbi.nlm.nih.gov/pubmed/18656541.

Sanbo Qin and Huan-Xiang Zhou. meta-ppisp: a meta web server for protein-protein interaction site prediction. *Bioinformatics*, 23(24):3386–3387, Dec 2007. doi: 10.1093/bioinformatics/btm434. URL http://dx.doi.org/10.1093/bioinformatics/btm434.

Tanya M Raschke. Water structure and interactions with protein surfaces. *Curr Opin Struct Biol*, 16(2):152–159, Apr 2006. doi: 10.1016/j.sbi.2006.03.002. URL http://dx.doi.org/10.1016/j.sbi.2006.03.002.

D. Reichmann, O. Rahat, S. Albeck, R. Meged, O. Dym, and G. Schreiber. The modular architecture of protein-protein binding interfaces. *Proc Natl Acad Sci U S A*, 102(1):57–62, Jan 2005. doi: 10.1073/pnas.0407280102. URL http://dx.doi.org/10.1073/pnas.0407280102.

Dana Reichmann, Ofer Rahat, Mati Cohen, Hani Neuvirth, and Gideon Schreiber. The molecular architecture of protein-protein binding sites. *Curr Opin Struct Biol*, 17(1):67–76, Feb 2007. doi: 10.1016/j.sbi.2007.01.004. URL http://dx.doi.org/10.1016/j.sbi.2007.01.004.

Dana Reichmann, Yael Phillip, Asaf Carmi, and Gideon Schreiber. On the contribution of water-mediated interactions to protein-complex stability. *Biochemistry*, 47(3):1051–1060, Jan 2008. doi: 10.1021/bi7019639. URL http://dx.doi.org/10.1021/bi7019639.

Shlomi Reuveni, Rony Granek, and Joseph Klafter. Proteins: coexistence of stability and flexibility. *Phys Rev Lett*, 100(20):208101, May 2008.

David W Ritchie, Dima Kozakov, and Sandor Vajda. Accelerating and focusing protein-protein docking correlations using multi-dimensional rotational fft generating functions. *Bioinformatics*, 24(17):1865–1873, Sep 2008. doi: 10.1093/bioinformatics/btn334. URL http://dx.doi.org/10.1093/bioinformatics/btn334.

Francis Rodier, Ranjit Prasad Bahadur, Pinak Chakrabarti, and Joël Janin. Hydration of protein-protein interfaces. *Proteins*, 60(1):36–45, Jul 2005. doi: 10.1002/prot.20478. URL http://dx.doi.org/10.1002/prot.20478.

Alexey N. Romanov, Sergey N. Jabin, Yaroslav B. Martynov, Alexey V. Sulimov, Fedor V. Grigoriev, and Vladimir B. Sulimov. Surface generalized born method:â a simple, fast, and precise implicit solvent model beyond the coulomb approximation. *The Journal of Physical Chemistry A*, 108(43):9323–9327, 2004. doi: 10.1021/jp046721s. URL http://pubs.acs.org/doi/abs/10.1021/jp046721s.

Lubka T Roumenina, Marieta M Ruseva, Alexandra Zlatarova, Rohit Ghai, Martin Kolev, Neli Olova, Mihaela Gadjeva, Alok Agrawal, Barbara Bottazzi, Alberto Mantovani, Kenneth B M Reid, Uday Kishore, and Mihaela S Kojouharova. Interaction of c1q with igg1, c-reactive protein and pentraxin 3: mutational studies using recombinant globular head modules of human c1q a, b, and c chains. *Biochemistry*, 45(13): 4093–104, April 2006. ISSN 0006-2960. URL http://www.ncbi.nlm.nih.gov/pubmed/16566583.

Jean-François Rual, Kavitha Venkatesan, Tong Hao, Tomoko Hirozane-Kishikawa, Amélie Dricot, Ning Li, Gabriel F Berriz, Francis D Gibbons, Matija Dreze, Nono Ayivi-Guedehoussou, Niels Klitgord, Christophe Simon, Mike Boxem, Stuart Milstein, Jennifer Rosenberg, Debra S Goldberg, Lan V Zhang, Sharyl L Wong, Giovanni Franklin, Siming Li, Joanna S Albala, Janghoo Lim, Carlene Fraughton, Estelle Llamosas, Sebiha Cevik, Camille Bex, Philippe Lamesch, Robert S Sikorski, Jean Vandenhaute, Huda Y Zoghbi, Alex Smolyar, Stephanie Bosak, Reynaldo Sequerra, Lynn Doucette-Stamm, Michael E Cusick, David E Hill, Frederick P Roth, and Marc Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, Oct 2005. doi: 10.1038/nature04209. URL http://dx.doi.org/10.1038/nature04209.

Anatoly M Ruvinsky, Tatsiana Kirys, Alexander V Tuzikov, and Ilya A Vakser. Side-chain conformational changes upon protein-protein association. *J Mol Biol*, 408(2):356–365, Apr 2011. doi: 10.1016/j.jmb.2011.02.030. URL http://dx.doi.org/10.1016/j.jmb.2011.02.030.

S. Ryu, S. Zhou, A. G. Ladurner, and R. Tjian. The transcriptional cofactor complex crsp is required for activity of the enhancer-binding protein sp1. *Nature*, 397(6718):446–450, Feb 1999. doi: 10.1038/17141. URL http://dx.doi.org/10.1038/17141.

Sergey Samsonov, Joan Teyra, and M. Teresa Pisabarro. A molecular dynamics approach to study the importance of solvent in protein interactions. *Proteins*, 73(2):515–525, Nov 2008. doi: 10.1002/prot.22076. URL http://dx.doi.org/10.1002/prot.22076.

J Vidya Sarma and Peter A Ward. The complement system. *Cell Tissue Res*, 343(1):227–35, January 2011. ISSN 1432-0878. URL http://www.ncbi.nlm.nih.gov/pubmed/20838815.

L. F. Scatena, M. G. Brown, and G. L. Richmond. Water at hydrophobic surfaces: Weak hydrogen bonding and strong orientation effects. *Science*, 292(5518):908–912, 2001. doi: 10.1126/science.1059514. URL http://www.sciencemag.org/content/292/5518/908.abstract.

J. A. Schellman. A simple model for solvation in mixed solvents. applications to the stabilization and destabilization of macromolecular structures. *Biophys Chem*, 37(1-3):121–140, Aug 1990.

John A Schellman. Protein stability in mixed solvents: a balance of contact interaction and excluded volume. *Biophys J*, 85(1):108–125, Jul 2003. doi: 10.1016/S0006-3495(03)74459-2. URL http://dx.doi.org/10.1016/S0006-3495(03)74459-2.

Peter Schmidtke, Axel Bidon-Chanal, F. Javier Luque, and Xavier Barril. Mdpocket : Open source cavity detection and characterization on molecular dynamics trajectories. *Bioinformatics*, 2011a. doi: 10.1093/bioinformatics/btr550. URL http://bioinformatics.oxfordjournals.org/content/early/2011/10/03/bioinformatics.btr550.abstract.

Peter Schmidtke, F. Javier Luque, James B. Murray, and Xavier Barril. Shielded hydrogen bonds as structural determinants of binding kinetics: Application in drug design. *Journal of the American Chemical Society*, 133(46):18903–18910, 2011b. doi: 10.1021/ja207494u. URL http://pubs.acs.org/doi/abs/10.1021/ja207494u.

Sebastian Schneider and Martin Zacharias. Flexible protein-protein docking. In Xuhua Xia, editor, *Selected Works in Bioinformatics*, pages 161–176. InTech, 2011.

Sebastian Schneider and Martin Zacharias. Scoring optimisation of unbound protein-protein docking including protein binding site predictions. *Journal of Molecular Recognition*, 25(1):15–23, 2012a. ISSN 1099-1352. doi: 10.1002/jmr.1165. URL http://dx.doi.org/10.1002/jmr.1165.

Sebastian Schneider and Martin Zacharias. Atomic resolution model of the antibody fc interaction with the complement c1q component. *Molecular Immunology*, 51(1):66 – 72, 2012b. ISSN 0161-5890. doi: 10.1016/j.molimm.2012.02.111. URL http://www.sciencedirect.com/science/article/pii/S0161589012001356. <ce:title>7th International EMBO Workshop on Antigen Presentation and Processing</ce:title>.

G. Schreiber and A. R. Fersht. Rapid, electrostatically assisted association of proteins. *Nat Struct Biol*, 3 (5):427–431, May 1996.

G. Schreiber, G. Haran, and H-X. Zhou. Fundamental aspects of protein-protein association kinetics. *Chem Rev*, 109(3):839–860, Mar 2009. doi: 10.1021/cr800373w. URL http://dx.doi.org/10.1021/cr800373w.

Claudia N. Schutz and Arieh Warshel. What are the dielectric âconstantsâ of proteins and how to validate electrostatic models? *Proteins: Structure, Function, and Bioinformatics*, 44(4):400–417, 2001. ISSN 1097-0134. doi: 10.1002/prot.1106. URL http://dx.doi.org/10.1002/prot.1106.

Jesus Seco, F. Javier Luque, and Xavier Barril. Binding site detection and druggability index from first principles. *J Med Chem*, 52(8):2363–2371, Apr 2009. doi: 10.1021/jm801385d. URL http://dx.doi.org/10.1021/jm801385d.

David E. Shaw, Martin M. Deneroff, Ron O. Dror, Jeffrey S. Kuskin, Richard H. Larson, John K. Salmon, Cliff Young, Brannon Batson, Kevin J. Bowers, Jack C. Chao, Michael P. Eastwood, Joseph Gagliardo, J. P. Grossman, C. Richard Ho, Douglas J. Ierardi, István Kolossváry, John L. Klepeis, Timothy Layman, Christine McLeavey, Mark A. Moraes, Rolf Mueller, Edward C. Priest, Yibing Shan, Jochen Spengler, Michael Theobald, Brian Towles, and Stanley C. Wang. Anton, a special-purpose machine for molecular dynamics simulation. *SIGARCH Comput. Archit. News*, 35:1–12, June 2007. ISSN 0163-5964. doi: http://doi.acm.org/10.1145/1273440.1250664. URL http://doi.acm.org/10.1145/1273440.1250664.

David E. Shaw, Ron O. Dror, John K. Salmon, J. P. Grossman, Kenneth M. Mackenzie, Joseph A. Bank, Cliff Young, Martin M. Deneroff, Brannon Batson, Kevin J. Bowers, Edmond Chow, Michael P. Eastwood, Douglas J. Ierardi, John L. Klepeis, Jeffrey S. Kuskin, Richard H. Larson, Kresten Lindorff-Larsen, Paul Maragakis, Mark A. Moraes, Stefano Piana, Yibing Shan, and Brian Towles. Millisecond-scale molecular dynamics simulations on anton. In *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, SC '09, pages 65:1–65:11, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-744-8. doi: http://doi.acm.org/10.1145/1654059.1654126. URL http://doi.acm.org/10.1145/1654059.1654126.

Felix B Sheinerman and Barry Honig. On the role of electrostatic interactions in the design of protein-protein interfaces. *J Mol Biol*, 318(1):161–177, Apr 2002. doi: 10.1016/S0022-2836(02)00030-X. URL http://dx.doi.org/10.1016/S0022-2836(02)00030-X.

Felix B Sheinerman, Raquel Norel, and Barry Honig. Electrostatic aspects of proteinåprotein interactions. *Current Opinion in Structural Biology*, 10(2):153 – 159, 2000. ISSN 0959-440X. doi: 10.1016/S0959-440X(00)00065-8. URL http://www.sciencedirect.com/science/article/pii/S0959440X00000658.

Diwakar Shukla, Chetan Shinde, and Bernhardt L Trout. Molecular computations of preferential interaction coefficients of proteins. *J Phys Chem B*, 113(37):12546–12554, Sep 2009. doi: 10.1021/jp810949t. URL http://dx.doi.org/10.1021/jp810949t.

Ivan L Shulgin and Eli Ruckenstein. A protein molecule in an aqueous mixed solvent: fluctuation theory outlook. *J Chem Phys*, 123(5):054909, Aug 2005. doi: 10.1063/1.2011388. URL http://dx.doi.org/10.1063/1.2011388.

Nikolai Smolin, Alla Oleinikova, Ivan Brovchenko, Alfons Geiger, and Roland Winter. Properties of spanning water networks at protein surfaces. *The Journal of Physical Chemistry B*, 109(21):10995–11005, 2005. doi: 10.1021/jp050153e. URL http://pubs.acs.org/doi/abs/10.1021/jp050153e. PMID: 16852340.

Eduardo M. Sproviero, Jose A. Gascon, James P. McEvoy, Gary W. Brudvig, and Victor S. Batista. Quantum mechanics/molecular mechanics study of the catalytic cycle of water splitting in photosystem ii. *Journal of the American Chemical Society*, 130(11):3428–3442, 2008. doi: 10.1021/ja076130q. URL http://pubs.acs.org/doi/abs/10.1021/ja076130q.

Jayashree Srinivasan, Megan W. Trevathan, Paul Beroza, and David A. Case. Application of a pairwise generalized born model to proteins and nucleic acids: inclusion of salt effects. *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)*, 101:426–434, 1999. ISSN 1432-881X. URL http://dx.doi.org/10.1007/s002140050460. 10.1007/s002140050460.

P. J. Steinbach and B. R. Brooks. Protein hydration elucidated by molecular dynamics simulation. *Proc Natl Acad Sci U S A*, 90(19):9135–9139, Oct 1993.

W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrikson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *Journal of the American Chemical Society*, 1990.

ShihChe Sue, Harold C. Jarrell, JeanRobert Brisson, and Wenguey Wu. Dynamic characterization of the water binding loop in the p-type cardiotoxin:â implication for the role of the bound water moleculeå . *Biochemistry*, 40(43):12782–12794, 2001. doi: 10.1021/bi010848f. URL http://pubs.acs.org/doi/abs/10.1021/bi010848f. PMID: 11669614.

D. I. Svergun, S. Richard, M. H. Koch, Z. Sayers, S. Kuprin, and G. Zaccai. Protein hydration in solution: experimental observation by x-ray and neutron scattering. *Proc Natl Acad Sci U S A*, 95(5):2267–2272, Mar 1998.

Y. H. Tan, C. H.and Tan and R. Luo. Implicit nonpolar solvent models. *J. Phys. Chem. B*, 2007.

C Tanford. The hydrophobic effect and the organization of living matter. *Science*, 200(4345):1012–1018, 1978. doi: 10.1126/science.653353. URL http://www.sciencemag.org/content/200/4345/1012.abstract.

Charles Tanford. Interfacial free energy and the hydrophobic effect. *Proceedings of the National Academy of Sciences*, 76(9):4175–4176, 1979. URL http://www.pnas.org/content/76/9/4175.abstract.

M. Tarek and D. J. Tobias. Role of protein-water hydrogen bond dynamics in the protein dynamical transition. *Phys Rev Lett*, 88(13):138101, Apr 2002.

J E Thommesen, T E Michaelsen, Láset GA, I Sandlie, and O H Brekke. Lysine 322 in the human igg3 c(h)2 domain is crucial for antibody dependent complement activation. *Mol Immunol*, 37(16):995–1004, November 2000. ISSN 0161-5890. URL http://www.ncbi.nlm.nih.gov/pubmed/11395138.

K. S. Thorn and A. A. Bogan. Asedb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, 17(3):284–285, Mar 2001.

S. N. Timasheff and H. Inoue. Preferential binding of solvent components to proteins in mixed water–organic solvent systems. *Biochemistry*, 7(7):2501–2513, Jul 1968.

Serge N Timasheff. Protein hydration, thermodynamic binding, and preferential hydration. *Biochemistry*, 41(46):13473–13482, Nov 2002.

Ilario G. Tironi, Rene Sperb, Paul E. Smith, and Wilfired F. van Gunsteren. A generalized reaction field method for molecular dynamics simulations. *J. Chem. Phys.*, 1995.

Dror Tobi and Ivet Bahar. Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. *Proc Natl Acad Sci U S A*, 102(52):18908–18913, Dec 2005. doi: 10.1073/pnas.0507603102. URL http://dx.doi.org/10.1073/pnas.0507603102.

Maxim Totrov and Ruben Abagyan. Flexible ligand docking to multiple receptor conformations: a practical alternative. *Curr Opin Struct Biol*, 18(2):178–184, Apr 2008. doi: 10.1016/j.sbi.2008.01.004. URL http://dx.doi.org/10.1016/j.sbi.2008.01.004.

Leendert A Trouw and Mohamed R Daha. Role of complement in innate immunity and host defense. *Immunol Lett*, 138(1):35–7, July 2011. ISSN 1879-0542. URL http://www.ncbi.nlm.nih.gov/pubmed/21333684.

C. J. Tsai, S. L. Lin, H. J. Wolfson, and R. Nussinov. Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci*, 6(1):53–64, Jan 1997. doi: 10.1002/pro.5560060106. URL http://dx.doi.org/10.1002/pro.5560060106.

C. J. Tsai, S. Kumar, B. Ma, and R. Nussinov. Folding funnels, binding funnels, and protein function. *Protein Sci*, 8(6):1181–1190, Jun 1999. doi: 10.1110/ps.8.6.1181. URL http://dx.doi.org/10.1110/ps.8.6.1181.

Vincent Vagenende, Miranda G S Yap, and Bernhardt L Trout. Molecular anatomy of preferential interaction coefficients by elucidating protein solvation in mixed solvents: methodology and application for lysozyme in aqueous glycerol. *J Phys Chem B*, 113(34):11743–11753, Aug 2009. doi: 10.1021/jp903413v. URL http://dx.doi.org/10.1021/jp903413v.

Sandor Vajda and Frank Guarnieri. Characterization of protein-ligand interaction sites using experimental and computational methods. *Curr Opin Drug Discov Devel*, 9(3):354–362, May 2006.

Sandor Vajda and Dima Kozakov. Convergence and combination of methods in protein-protein docking. *Curr Opin Struct Biol*, 19(2):164–170, Apr 2009. doi: 10.1016/j.sbi.2009.02.008. URL http://dx.doi.org/10.1016/j.sbi.2009.02.008.

W.F. van Gunsteren and H.J.C. Berendsen. A leap-frog algorithm for stochastic dynamics. *Mol.Sim.*, 1988.

Wilfred F. van Gunsteren, Xavier Daura, and Alan E. Mark. *GROMOS Force Field*. John Wiley & Sons, Ltd, 2002. ISBN 9780470845011. doi: 10.1002/0470845015.cga011. URL http://dx.doi.org/10.1002/0470845015.cga011.

C. M. Venkatachalam, X. Jiang, T. Oldfield, and M. Waldman. Ligandfit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *J Mol Graph Model*, 21(4):289–307, Jan 2003.

Jouko J Virtanen, Lee Makowski, Tobin R Sosnick, and Karl F Freed. Modeling the hydration layer around proteins: Hypred. *Biophys J*, 99(5):1611–1619, Sep 2010. doi: 10.1016/j.bpj.2010.06.027. URL http://dx.doi.org/10.1016/j.bpj.2010.06.027.

Andrea Volkamer, Axel Griewel, Thomas Grombacher, and Matthias Rarey. Analyzing the topology of active sites: on the prediction of pockets and subpockets. *J Chem Inf Model*, 50(11):2041–2052, Nov 2010. doi: 10.1021/ci100241y. URL http://dx.doi.org/10.1021/ci100241y.

Rebecca C. Wade and Peter J. Goodford. Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 2. ligand probe groups with the ability to form more than two hydrogen bonds. *Journal of Medicinal Chemistry*, 36(1):148–156, 1993. doi: 10.1021/jm00053a019. URL http://pubs.acs.org/doi/abs/10.1021/jm00053a019.

Chu Wang, Philip Bradley, and David Baker. Protein-protein docking with backbone flexibility. *J Mol Biol*, 373(2):503–519, Oct 2007. doi: 10.1016/j.jmb.2007.07.050. URL http://dx.doi.org/10.1016/j.jmb.2007.07.050.

J. Warwicker and H.C. Watson. Calculation of the electric potential in the active site cleft due to alpha-helix dipoles. *Journal of Molecular Biology*, 157(4):671 – 679, 1982. ISSN 0022-2836. doi: 10.1016/0022-2836(82)90505-8. URL http://www.sciencedirect.com/science/article/pii/0022283682905058.

Mark Nicholas Wass, Gloria Fuentes, Carles Pons, Florencio Pazos, and Alfonso Valencia. Towards the prediction of protein interaction partners using physical docking. *Mol Syst Biol*, 7:469, Feb 2011. doi: 10.1038/msb.2011.3. URL http://dx.doi.org/10.1038/msb.2011.3.

Scott J. Weiner, Peter A. Kollman, David A. Case, U. Chandra Singh, Caterina Ghio, Guliano Alagona, Salvatore Profeta, and Paul Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society*, 106(3):765–784, 1984. doi: 10.1021/ja00315a051. URL http://pubs.acs.org/doi/abs/10.1021/ja00315a051.

Martin Weisel, Ewgenij Proschak, and Gisbert Schneider. Pocketpicker: analysis of ligand binding-sites with shape descriptors. *Chem Cent J*, 1:7, 2007. doi: 10.1186/1752-153X-1-7. URL http://dx.doi.org/10.1186/1752-153X-1-7.

Martin Weisel, Ewgenij Proschak, Jan M. Kriegl, and Gisbert Schneider. Form follows function: Shape analysis of protein cavities for receptor-based drug design. *PROTEOMICS*, 9(2):451–459, 2009. ISSN 1615-9861. doi: 10.1002/pmic.200800092. URL http://dx.doi.org/10.1002/pmic.200800092.

J. A. Wells. Additivity of mutational effects in proteins. *Biochemistry*, 29(37):8509–8517, Sep 1990.

J. A. Wells. Systematic mutational analyses of protein-protein interfaces. *Methods Enzymol*, 202:390–411, 1991.

James A Wells and Christopher L McClendon. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature*, 450(7172):1001–1009, Dec 2007. doi: 10.1038/nature06526. URL http://dx.doi.org/10.1038/nature06526.

Jerome Wielens, Stephen J. Headey, John J. Deadman, David I. Rhodes, Michael W. Parker, David K. Chalmers, and Martin J. Scanlon. Fragment-based design of ligands targeting a novel site on the integrase enzyme of human immunodeficiency virusâ1. *ChemMedChem*, 6(2):258–261, 2011. ISSN 1860-7187. doi: 10.1002/cmdc.201000483. URL http://dx.doi.org/10.1002/cmdc.201000483.

Angela D. Williams, Shankaramma Shivaprasad, and Ronald Wetzel. Alanine scanning mutagenesis of aβ(1-40) amyloid fibril stability. *Journal of Molecular Biology*, 357(4):1283 – 1294, 2006. ISSN 0022-2836. doi: 10.1016/j.jmb.2006.01.041. URL http://www.sciencedirect.com/science/article/pii/S0022283606000696.

K. Wüthrich, G. Otting, and E. Liepinsh. Protein hydration in aqueous solution. *Faraday Discuss*, (93): 35–45, 1992.

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

Jian Yu, Yong Zhou, Isao Tanaka, and Min Yao. Roll: a new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. *Bioinformatics*, 26(1):46–52, Jan 2010. doi: 10.1093/bioinformatics/btp599. URL http://dx.doi.org/10.1093/bioinformatics/btp599.

M. Zacharias and H. Sklenar. Harmonic modes as variables to approximately account for receptor flexibility in ligand–receptor docking simulations: Application to a dna minor groove ligand complex. *Journal of Comput Chemistry*, 1999.

Martin Zacharias. Protein-protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Sci*, 12(6):1271–1282, Jun 2003. doi: 10.1110/ps.0239303. URL http://dx.doi.org/10.1110/ps.0239303.

Martin Zacharias. Accounting for conformational changes during protein-protein docking. *Curr Opin Struct Biol*, 20(2):180–186, Apr 2010a. doi: 10.1016/j.sbi.2010.02.001. URL http://dx.doi.org/10.1016/j.sbi.2010.02.001.

Martin Zacharias. *Protein-Protein Complexes: Analysis, Modeling and Drug Design*, chapter Scoring and refinement of predicted protein-protein complexes. Zacharias, Martin, 2010b.

Chi Zhang, Song Liu, and Yaoqi Zhou. Docking prediction using biological information, zdock sampling technique, and clustering guided by the dfire statistical energy function. *Proteins*, 60(2):314–318, Aug 2005. doi: 10.1002/prot.20576. URL http://dx.doi.org/10.1002/prot.20576.

Luyuan Zhang, Yi Yang, Ya-Ting Kao, Lijuan Wang, and Dongping Zhong. Protein hydration dynamics and molecular mechanism of coupled water-protein fluctuations. *J Am Chem Soc*, 131(30):10677–10691, Aug 2009a. doi: 10.1021/ja902918p. URL http://dx.doi.org/10.1021/ja902918p.

Qing Zhang, Michel Sanner, and Arthur J Olson. Shape complementarity of protein-protein complexes at multiple resolutions. *Proteins*, 75(2):453–467, May 2009b. doi: 10.1002/prot.22256. URL http://dx.doi.org/10.1002/prot.22256.

Huan-Xiang Zhou and Sanbo Qin. Interaction-site prediction for protein complexes: a critical assessment. *Bioinformatics*, 23(17):2203–2209, Sep 2007. doi: 10.1093/bioinformatics/btm323. URL http://dx.doi.org/10.1093/bioinformatics/btm323.

Jan Zielkiewicz. Structural properties of water: comparison of the spc, spce, tip4p, and tip5p models of water. *J Chem Phys*, 123(10):104501, Sep 2005. doi: 10.1063/1.2018637. URL http://dx.doi.org/10.1063/1.2018637.

Alexandra S Zlatarova, Marieta Rouseva, Lubka T Roumenina, Mihaela Gadjeva, Martin Kolev, Ivan Dobrev, Neli Olova, Rohit Ghai, Jens Chr Jensenius, Kenneth B M Reid, Uday Kishore, and Mihaela S Kojouharova. Existence of different but overlapping igg- and igm-binding sites on the globular domain of human c1q. *Biochemistry*, 45(33):9979–88, August 2006. ISSN 0006-2960. URL http://www.ncbi.nlm.nih.gov/pubmed/16906756.

Danksagung

Ich hoffe, dass ich auch abseits einer schriftlichen Danksagung in der Lage bin, meinen Dank den Leuten gegenüber auszudrücken, die an meinem Leben partizipieren und essentiell für mich sind. Da meine Promotion nicht nur mich, sondern alle Menschen in meinem Umfeld direkt oder indirekt beeinflusst hat, soll dennoch eine Danksagung folgen:

Ich danke Natalie und meiner Familie für die Unterstützung und den Rückhalt, Herrn Zacharias für das Vertrauen in mich und die Betreuung meiner Arbeit und meinen Kollegen für interessante Gespräche bezüglich und abseits wissenschaftlicher Themen.

Ich danke all meinen Freunden, die noch mit mir sprechen, trotz meiner Versicherung, dass ich mich alsbald melde, sobald ich meine Doktor Arbeit beendet habe.
Zuletzt möchte ich denjenigen danken, denen ich nicht mehr danken kann.