In Silico Prediction of Binding Sites on Proteins

Simon Leis, Sebastian Schneider and Martin Zacharias*

Physik-Department T38, Technische Universität München, James-Franck-Strasse, D-85748 Garching, Germany

Abstract: The majority of biological processes involve the association of proteins or binding of other ligands to proteins. The accurate prediction of putative binding sites on the protein surface can be very helpful for rational drug design on target proteins of medical relevance, for predicting the geometry of protein-protein as well as protein-ligand complexes and for evaluating the tendency of proteins to aggregate or oligomerize. A variety of computational methods to rapidly predict protein-protein binding interfaces or binding sites for small drug-like molecules have been developed in recent years. The principles of methods available for protein interface and pocket detection are summarized, including approaches based on sequence conservation, as well as geometric and physicochemical surface properties. The performance of several Web-accessible methods for ligand binding site prediction has been compared using protein structures in bound and unbound conformation and homology modeled proteins. All methods tested gave very promising predictions even on unbound and homology modeled protein structures, thus indicating that current methods are robust in relation to modest conformational changes associated with the ligand binding process.

INTRODUCTION

Biomolecules and many other organic ligands can bind to proteins with high affinity at specific sites on the protein surface. The question of what distinguishes such recognition sites from other surface regions of proteins has been the subject of intense experimental and theoretical research [1, 2]. In recent years, the possibility to predict putative binding regions on the surface of protein molecules has become increasingly important. Together with the rapidly growing structural knowledge of proteins of biological and medical importance, such prediction methods become more applicable and can be helpful for rational drug design and to elucidate the function of a protein molecule. Both these applications, function prediction as well as rational drug design, require a reliable method for identifying and characterizing the ligand-binding sites of a protein.

Knowing the location of the functional sites (e.g., substrate or ligand-binding sites of enzymes or receptor proteins) on the protein surfaces prior to experiment, makes it possible to design inhibitors or antagonists and to introduce targeted mutations aimed at improving the protein function. It is also possible to apply these methods to assist in modeling the three-dimensional (3D)-structure of protein-ligand complexes [3-5].

The availability of 3D structures of many proteins in complex with proteins or other types of ligands (lipids, nucleic acids or drug-like molecules) allows the systematic comparison of protein surfaces involved in interactions [2, 6-19]. Comparative studies of the amino acid distribution and physicochemical features of protein-protein interfaces [6-13] and proteins in complex with small organic drug-like ligands [14-19] made it possible to characterize recognition sites. Furthermore, often interface residues around binding sites are evolutionary more conserved than other surface regions. A variety of computational methods have been developed that try to integrate this information for predicting putative binding sites in proteins.

The realistic prediction of putative ligand or protein binding sites has not only important implications for rational drug design but could also have an impact on a better understanding of protein-protein interaction networks. The possibility to identify and to characterize putative protein binding sites on proteins can help to elucidate the number and kind of protein interaction partners. *In silico* methods to predict protein-protein interaction sites can also be used to predict the propensity of proteins to aggregate [20, 21] or to bind non-specifically to many different partners [22]. Recent approaches to predict not only binding sites on proteins but also which partner protein may bind could potentially be useful to predict protein interaction networks [23].

In this review we first give an overview of the geometric and physicochemical properties of protein-protein interfaces and of protein binding sites for small drug-like ligands (in the following: protein-ligand complexes) based on the analysis of known 3D structures. Recent approaches for predicting putative protein-protein interfaces and binding sites for drug-like ligands will be discussed in the second part, followed by an analysis of the robustness of ligand binding site prediction, with respect to conformational changes or inaccuracies in the protein structure. Finally, challenging future issues will be discussed.

COMPARISON OF PROTEIN-PROTEIN AND PROTEIN-LIGAND INTERACTION REGIONS

Complexes of proteins are non-covalent protein assemblies that fold separately and associate under certain physiological conditions. Examples of protein-protein complexes are antigen-antibody, enzyme-inhibitor, and many signal transduction and cell cycle protein complexes [24]. The majority of known protein-protein complex structures have been determined by X-ray crystallography which requires stable complex structures that can form well ordered crystals. Based on known complex structures the geometric and physicochemical properties of protein-protein interfaces have been characterized in detail [2, 6-12, 24]. In the following an overview of the main results will be given. It is important to indicate that the analysis of interface properties of protein-protein binding sites is restricted to sufficiently stable protein-protein complexes (that can form well ordered crystals).

*Address correspondence to this author at the Physik-Department T38, Technische Universität München, James-Franck-Strasse, D-85748 Garching, Germany; Tel: 0049-89-28912335;
E-mail: martin.zacharias@ph.tum.de

0929-8673/10 $55.00+.00

Rules derived for these complexes may differ from interfaces formed during transient interactions with a short lifetime. Most protein-protein complexes bury a surface area in the range of 1200-2000 Å^2 which is much larger than the buried surface area upon binding small drug-like molecules of a few hundred Å^2 depending on the size of the ligand [14]. The comparison of known protein-protein complex structures indicates that protein-protein interfaces are in many cases overall flat in shape with the exception of several enzyme-inhibitor complexes [10-13] where the inhibitor site often forms a convex surface fitting to the concave shape of the enzyme active site. This contrasts to binding sites for enzyme substrates or other small organic ligands that are usually very non-planar allowing contacts to the ligand from many different sides of the binding pocket [14-20].

Protein interface regions clearly differ on average from the rest of the protein surface in terms of physicochemical properties and geometric characteristics [9-13]. However, the interactions between proteins are very diverse. It is therefore not possible to distinguish a binding site from the rest of the protein surface based on a single surface attribute [13]. Interface residues in protein-protein complexes can be divided into two distinct regions, the ‘core’ and the ‘rim’ region, based on the solvent accessibility in the complex [11, 12]. The ‘core’ region contains residues that have at least one fully buried interface atom (i.e. zero accessibility after complex formation) and usually contain mostly non-polar residues surrounded by the more polar ‘rim’ region, which contains residues that are at least partially solvent exposed even in the complex. The composition of amino acid residues at specific protein-protein interfaces differs from the rest of the protein surface. Interface regions are enriched in aliphatic (Leu, Val, Ile, Met) and aromatic (His, Phe, Tyr, Trp) residues, and depleted in charged residues (Asp, Glu, Lys) with the exception of arginine [7-13]. The higher abundance of Arg at interfaces compared to Lys has been attributed to formation of cation-π-interactions [10] and the greater capacity of the guanidinium group in Arg to form hydrogen bonds (compared to Lys) [12, 24]. The role of arginine-arginine pairing and its contribution to protein-protein interactions was recently investigated by Vondrášek and coworkers employing computational approaches [25].

One way to characterize the relative contributions of interface residues to the binding free energy, is to determine the change in affinity upon mutation of interface residues to alanine. Substitution of residues by alanine (alanine-scanning mutagenesis) corresponds (except for glycine) to the removal of side chain atoms from the interface and its effect on binding strength [26-30]. Interestingly, for most protein-protein complexes analysed by alanine scanning mutagenesis only a fraction of substitutions showed a substantial effect on binding affinity [26, 27]. This finding has led to the concept of “hot spots” on protein surfaces that are responsible for most of the interaction between proteins [27, 30] and methods for *in silico* alanine-scanning have been developed [28,31-33]. Several methods to predict protein-protein interaction sites aim at identifying such “hot spots” on protein surfaces [reviewed in 30].

It is important however to keep in mind that the binding affinity between two proteins is determined by interacting pairs of residues or even higher order motifs and not only by individual amino acids on just one partner. Hence, a given contacting pair (e.g. of two polar or charged residues) at an interface may overall contribute little to binding, for example, because the desolvation of the two polar residues upon binding offsets the interaction energy between the residues. Nevertheless, substitution of one of the polar or charged residues by alanine may result in a significant drop of binding affinity (because alanine cannot form polar contacts), which may lead to the erroneous conclusion that the region is a hot spot. The substitution of a residue with zero contribution to binding energy can still result in a large drop of binding affinity if it creates an unfavorable contact with another residue.

Similar to protein-protein interaction sites, high affinity binding cavities for small drug-like ligands are often less polar (low desolvation penalty) or more hydrophobic compared to the rest of the protein surface [14-20]. However, due to the smaller size of organic drug-like molecules compared to proteins, the buried surface-area upon small molecule protein-ligand interaction is generally smaller than in the case of protein-protein interactions. In order to achieve strong interactions through a sufficiently large number of favorable protein-ligand contacts, high-affinity binding sites are usually strongly concave pockets or cavities on the surface of proteins or sometimes partially buried [14].

Algorithms for predicting protein-protein interfaces are, in many aspects, similar to methods for predicting binding regions for small drug-like molecules. However, there are also some important differences due to the distinct general architecture of these types of binding sites [18, 34].

APPROACHES TO PREDICT PROTEIN-PROTEIN INTERFACES

Protein-protein interaction sites or interfaces can be defined as those protein surface residues or atoms that become buried upon complex formation. It is possible to identify interface residues (atoms) by calculating solvent accessibility in the presence and absence of the binding partner and define a threshold for counting a change in accessibility as being part of the interface. Alternatively (and more common), a distance criterion for intermolecular contacts in the complex is used to define all residues that are at the interface of the complex. The purpose of protein-protein interface prediction methods is to predict residues or atoms that belong to a putative protein binding interface.

In order to recognize putative protein-protein interaction regions one can distinguish between methods that are based on the physicochemical properties of protein surfaces and approaches that are based on the evolutionary conservation of exposed surface residues [reviewed in 35-39]. It is possible to define a third category not considered here based on sequence homology of a protein-protein complex to a complex of known structure. Under the assumption that the binding geometry in the unknown complex is similar to the known complex, it is possible to define a possible binding region. Although the majority of interface prediction methods require the 3D structure of the protein as input, there are also attempts that aim at predicting interfaces solely on the basis of the protein sequence [40-43]. In addition, recent approaches aim at predicting not only putative interaction regions of a protein, but also what kind of protein might be the binding partner [23].

METHODS BASED ON SURFACE RESIDUE CONSERVATION

An amino acid residue in a family of proteins can be evolutionary conserved because the residue is important for the folding of the protein. This concerns mostly residues located in the interior of the folded protein (buried residues) responsible for the hydrophobic core or tight packing of the protein. In addition, residues not involved in the correct folding of the protein can be conserved for functional reasons [35]. Possible conserved functions include the binding of a protein or other ligands at a distinct site. Conservation of residues in a set of related proteins can be derived from a multiple alignment of corresponding sequences [44-46]. Several protein interface prediction methods are based on residue conservation as the main input information. Examples are the ConSurf [46, 47], Rate4Site [48], SiteFinder3D [49] and the evolutionary trace (ET)-Viewer [45]. The evolutionary trace method defines an evolutionary tree that partitions the protein family into an increasing number of subgroups and ends with a subgroup for each protein. Based on the branches of the tree, it can define evolutionary conserved residues (trace residues) of functional importance [35, 44, 45]. From the clustering of trace residues at the protein surface, it is possible to identify putative interface residues (or residues of other functional importance).

It has, however, been demonstrated that conserved patches of surface residues alone may not be sufficient to clearly discriminate between protein binding interfaces and the rest of the protein surface [50]. Therefore, most prediction algorithms based on residue conservation combine the conservation information with geometric or physicochemical data on the protein surface. Examples of this class are the WHISCY [51] and the JET (Joint evolutionary tree) method [52].

PHYSICOCHEMICAL BASED METHODS AND COMBINED METHODS

Based on the analysis of known protein-protein complexes several surface properties have been used to identify putative binding interfaces. These properties include the chemical composition or type of amino acid, the shape of the surface, the overall hydrophobicity, the electrostatic field and the solvation characteristics of surface regions. In addition, crystallographic B-factors and detection of putative binding hot-spots have been used to predict protein-protein interaction sites [reviewed in 37, 38]. It has also been noted that the secondary structure at protein-protein interfaces has a preference for β-strands (over α-helices) and may contain loops that are longer than the typical loop length in proteins [53]. Residues that are part of interfaces tend to have a solvent accessibility that is larger than the average accessibility of the amino acid type on protein surfaces [54-56]. It also appears that interface residues may adopt a more limited set of side chain rotamer conformations in order to minimize the entropic cost of freezing the side chain in one conformation upon complex formation. This property has also been included in identifying putative protein binding sites [57].

However, no single physicochemical or geometric surface property has so far been identified in known protein-protein complexes, that unambiguously distinguishes protein-protein interface regions from other surface regions [36]. Therefore, most approaches use a combination of surface properties and also often include conservation of surface residues to predict protein interface regions. The predictive power of different surface properties to detect putative interaction sites has been investigated by Burgoyne and Jackson [34]. It was found that desolvation properties and residue conservation have the strongest predictive power. Typically, the desolvation properties of a putative protein binding region are calculated from the loss of solvent accessible surface that becomes buried upon complex formation. The desolvation penalty is calculated by assigning each surface element with an atom-type specific surface tension. Fernandez-Recio *et al.* [58] have used such an approach to successfully identify putative protein interfaces on a set of protein structures. Although the surface area based approach provides a rapid estimate, the calculation neglects the influence of the neighbourhood on the solvation of a residue. In polar solvents like water, solute-solvent and solvent-solvent, electrostatic interactions are predominant. The perturbation of the electrostatic field in the vicinity of an amino acid upon removal of water molecules is crucial for the desolvation process. Local effects that reduce solvation penalties for example due to the “neutralisation” of a charged residue due to a nearby opposite charge or due to long range electrostatic interactions are omitted by surface-area-based solvation calculations. A conceptually different approach to estimate the penalty to desolvate a protein surface region has recently been introduced, which is based on the desolvation properties of a probe placed at the protein surface and calculating the change in electrostatic energy by solving the finite-difference Poisson-Boltzmann equation [59]. This approach showed promising results on many different types of interacting proteins.

An overview of Web-accessible methods for protein-protein interface prediction is given in Table **1** together with a brief explanation of the underlying principle of the approach. Various available protein-protein interface prediction approaches have recently been reviewed and the performance on several benchmark sets were collected and compared [37, 38]. Several of the most recent methods perform on average similarly although the performance is case dependent and may also depend on the type of protein-protein complex. In most cases antigen-antibody complexes have been excluded from the analysis due to the larger variance of antibody-antigen interfaces compared to other types of complexes. According to a recent comparison by deVries and Bonvin [37] the best available methods can achieve 35-40 % specificity at a sensitivity of 30% or 25-30% at a sensitivity of 50%. Specificity is defined as the ratio of correctly predicted interface residues (true positives) relative to the sum of true positives and false positives. Sensitivity is the ratio between true positives and the sum of true positives and false negative predicted residues.

Although some methods are available which are only based on the protein sequence to predict interface residues

Table 1. ***In Silico*** **Methods for Predicting Putative Protein-Protein Interfaces**

PPI-Server & Web-site	Method
HotPatch [60] http://hotpatch.mbi.ucla.edu/	Identification of patches of residues on protein surface corresponding to functional sites
ISIS [40] NN	Prediction of interface residues based on sequence alone (from multiple alignments)
PIER [61] NN	Identification based on physicochemical and statistical surface properties of atomic groups at protein surface
PINUP [57] http://sparks.informatics.iupui.edu/PINUP/	Based on physicochemical properties and conservation including residue-energy score and accessible-surface-area of surface residues
consPPISP [55] http://pipe.scs.fsu.edu/ppisp.html	Sequence conservation (position-specific sequence profiles), solvent accessibilities of residues and neighbors
metaPPISP [36] http://pipe.scs.fsu.edu/meta-ppisp.html	Combined server input from cons-PPISP, ProMate, PINUP, trained by linear regression
ProMate [53] http://bioinfo.weizmann.ac.il/promate/promate.html	Physicochemical properties & sequence conservation averaged over circular region on protein surface
ProteMot [62] http://protemot.csie.ntu.edu.tw/step1.cgi	Based on matching surface structures with template motifs extracted from known complexes
InterProSurf [63] http://curie.utmb.edu/	Clustering of putative interface residues based on solvent accessible surface area in the isolated subunits and propensity scale for interface residues
SHARP^2 [64] http://www.bioinformatics.sussex.ac.uk/SHARP2/sharp2.html	Patch analysis of surface residues with high probability to be part of interface (obtained from known complexes)
ET-Viewer [65] http://mammoth.bcm.tmc.edu/traceview/	Evolutionary trace server to identify putative functional regions based on sequence conservation
SPPIDER [56] http://sppider.cchmc.org/	Physicochemical properties integrating enhanced relative solvent accessibility (RSA) based on actual vs. predicted solvent accessibility of surface residues
PPI-PRED [66] http://bioinformatics.leeds.ac.uk/ppi-pred	Protein interface prediction based on surface shape and electrostatics
WHISCY [51] http://www.nmr.chem.uu.nl/Software/whiscy/startpage.htm	Uses surface sequence conservation supplemented with physicochemical surface features
JET [52] http://www.ihes.fr/~carbone/data.htm	Combination of evolutionary tree analysis and physicochemical surface properties

[40-43] the majority of methods requires the 3D-structure of the protein. An important issue, that has not yet been systematically investigated, is how much the predicted interface depends on the exact structure of the protein. Some surface properties are sensitive with respect to the backbone and side chain conformations of surface residues and the prediction accuracy may depend on whether a bound, unbound or homology modeled structure has been employed for the application. It is expected that methods based on sequence conservation are not sensitive to the protein conformation but several physicochemical properties will depend significantly on the conformation of the protein. Protein-protein interaction site prediction methods have been extensively investigated and compared by de Vries and Bonvin [37]. In the following we, therefore, focus on the methods that are available for predicting small drug-like ligand binding sites on proteins.

METHODS FOR PREDICTION OF PUTATIVE SMALL MOLECULE BINDING SITES

Similar surface properties and sequence conservation as used to predict protein-protein interfaces can also be used to identify putative interaction sites for small drug-like ligands. However, the predictive power of each property may differ from predictions of protein-protein interfaces due to the difference in architecture of high affinity binding sites for organic ligands compared to protein-protein interfaces (see previous paragraphs). Burgoyne and Jackson [34] compared the predictive power of different surface properties and found that it is in general easier to identify putative protein-ligand interfaces compared to protein-protein interfaces. Binding cleft detection and desolvation properties, as well as sequence conservation and to some degree electrostatic potential, have been identified as the strongest signals for predicting protein-ligand interfaces [34].

Since ligand binding sites involve in most cases the presence of a concave binding cleft on the protein surface (in contrast to the more flat protein-protein interfaces) the detection of binding pockets or protein cavities deserves special attention. Presumably, the better performance of binding site prediction for small drug-like ligands compared to protein binding site prediction is due to the importance of a concave binding site in the latter case. Apparently, such concave regions are less frequently found on protein surfaces than flat or slightly curved surfaces typical for protein-protein interfaces. Several algorithms based on different detection principles have been designed in recent years. Only the principles of the most common methods will be explained here, since pocket detection methods and explanations of algorithmic design have been reviewed in detail in [67]. In the following, we summarize the basic algorithmic ideas of pocket detection of the most common available methods. An overview on available (mostly Web-accessible) ligand binding site prediction methods including the respective web-links is given in Table **2**.

PRINCIPLES OF POCKET DETECTION

Various algorithms to identify surface clefts in proteins have been reviewed and explained in detail by Laurie &

Table 2. Protein-Ligand Binding Site Prediction Methods

Method	Description
GRID^{C} [70]	Protein-probe energies computed by Lennard-Jones, electrostatic and hydrogen bonding potentials are mapped onto a grid around the protein
Pocket^{C} [71]	A 3Å probe scans the protein along a Cartesian grid for line segments not overlapping with protein but surrounded by overlapping segments.
Delaney [72]	Expansion and contraction of spherical surface spherical probes is used to detect pockets where probe particles concentrate
Del Carpio [73]	Closest distances between the protein's centre of gravity and protein surface points are used to identify pockets.
VOIDOO^{C} [74]	Cavities are detected by stepwise increase of Van-der-Waals radii of all protein atoms. After a floodfill algorithm, sealed off localizations can be identified as cavities
SurfNet^{C} [75]	Spheres between two atoms containing no other atoms are created and scanned for the cluster of spheres with the largest volume
APROPOS^{C} [76]	Protein pockets are determined employing an alpha-shape algorithm that allows for a complete global envelope of the protein
LIGSITE^{C} [77]	On a regular grid around the protein, lines are drawn from each grid point along the x/y/z-axis as well as the cubic diagonals. Segments of lines that are enclosed by protein from both sides are considered as cavities.
Superstar [78]	Creates propensity maps of basic molecular probes along the protein surface.
PASS^{C} [68]	The algorithm repeats filtering and expanding a set of initial probe spheres on the protein surface to eventually find "active site points"
ConSurf^{W} [79]	Identifying functional sites on proteins by determining the conservation of sequence homologues.
CASTp^{W} [80]	Uses alpha shape theory and triangulation methods to predict pockets.
LigandFit^{C} [81]	Identifies possible binding sites using a flood-filling-algorithm and docks ligands using a Monte Carlo conformational search
Q-SiteFinder^{W} [82]	Energetically based method: clusters of protein surface regions that show favorable Van-der-Waals interactions with a methyl-group are collected and ranked
DrugSite [83]	Predicts binding sites on the basis of Van-der-Waals potential grid point maps
MEDock^{W} [84]	Evolutionary algorithm utilizing the maximum entropy (ME) property of the Gaussian probability distribution
LIGSITEcsc^{W} [85]	In extension to the traditional LigSite method, the Connolly surface area is calculated and grid points are scanned for surface-solvent-surface events. Additionally, the top three predicted pockets are re-ranked according to sequence conservation.
Screen/Mark-Us^{W} [86]	Cavities are geometrically determined *via* the difference between the molecular surface and the probe-specified molecular envelope and statistically analysis.
Pocket-Picker^{C} [87]	A rectangular grid is used to segregate relevant points along the protein surface which are then clustered and ranked according to shape descriptors.
Fuzzy-Oil-Drop-Model^{W} [88]	Analyzes the protein for regions with high hydrophobic deficiency, i.e. the difference between observed and idealized hydrophobicity distribution declared by the 'Fuzzy Oil Drop Model'
SiteMap [89]	Sets of relevant points are identified by geometric and energetic means and analyzed for hydrophobicity and other physicochemical properties
FINDSITE [90]	The method uses protein threading to identify ligand bound templates which are then superimposed and analyzed for similarities in the ligand binding sites

^{W}web server available.

^{C}source code/program available.

Webserver: *Consurf*: http://consurf.tau.ac.il/ - *CASTp*: http://sts-fw.bioengr.uic.edu/castp/calculation.php - *Q-SiteFinder*: http://www.modelling.leeds.ac.uk/qsitefinder/ - *MEDock*: http://medock.csie.ntu.edu.tw/ - *Protemot*: http://protemot.csbb.ntu.edu.tw/ - *LIGSITE*^{csc}: http://gopubmed2.biotec.tu-dresden.de/cgi-bin/index.php - *SCREEN/Mark-Us*: http://interface.bioc.columbia.edu/screen/ - *Fuzzy-Oil-Drop-Model*: http://www.bioinformatics.cm-uj.krakow.pl/activesite/

Jackson [67]. One can distinguish between geometry-based and energy-based detection methods. The latter methods define favorable cleft regions based on energetic evaluations, the former based on sterical considerations. Many methods employ a regular 3D grid and move probes along grid lines to define accessible and inaccessible or energetically favorable and unfavorable positions. Alternatively, probes placed on the solvent accessible surface of the protein can be used in combination with a variety of algorithms to define pocket regions. For example, the PASS program [68] filters out highly accessible surface probes and creates additional layers of surface probes on top of surface probes located in clefts. The procedure is repeated until all clefts are filled with probes. In addition, structural motifs typical for binding pockets have also been used to define binding sites [69]. The principles of the most common ligand binding site prediction methods are briefly explained in Table **2**. For a more detailed explanation the reader is referred to the original literature and for a comparison of the pocket detection methods to the review by Laurie & Jackson [67]. In the following, five of the most common Web-accessible programs are considered and subsequently used to compare the performance on several proteins in bound, unbound and in the form of modeled structures. The LigSite method is based on a regular 3D grid placed around the protein [77]. A probe is moved along the x, y and z directions and the cube diagonals of the grid. A grid point that is counted as part of a pocket is assigned if the grid line contains points before and after the point that overlapped with the protein. A Web-accessible extension of the method (LIGSITE^{csc}) includes the degree of surface conservation and has been shown to improve the performance [85]. The Q-SiteFinder [82] is an energy-based binding site predictor that clusters grid points of favorable (van der Waals) interactions with the protein to define a putative binding site. The CASTp algorithm uses an entirely different principle to detect binding pockets [80]. In the CASTp method a Delaunay triangulation of the protein is performed (means the entire protein shape is approximated by triangles). A pocket can be detected, based on the direction of norm vectors associated with triangles for a set of neighboring triangles. A Web-accessible server for this method is available (Table **2**). The Mark-Us method is another binding site prediction tool based on the SCREEN algorithm [86], which is based on a large set of physicochemical, structural, and geometric descriptors extracted from known complexes. Finally, the Web-accessible Fuzzy-Oil-Drop (FOD) method [88] identifies primarily hydrophobic patches on protein surfaces to assign putative binding regions.

ROBUSTNESS OF LIGAND BINDING SITE PREDICTION WITH RESPECT TO PROTEIN CONFORMATIONAL CHANGES

Proteins can undergo conformational changes upon ligand binding that may influence the steric accessibility of a binding cleft and can interfere with the ability of an algorithm to identify a potential binding site. In an effort to detect binding pockets for inhibitors of protein-protein interactions, Eyrisch and Helms [91, 92] recently applied the PASS program [68] for pocket detection to three unbound and bound protein conformations. The inhibitors target the interaction of the proteins interleukin-2 (IL-2) and IL-2Ra, the interaction of MDM2 and p53 and the pair BCL-X$_L$ and BH3, respectively. The native inhibitor pockets on BCL-X$_L$ and IL-2 proteins in the unbound form could not be identified using PASS. However, open pockets among them also the native pockets were found transiently during molecular dynamics simulations starting from the unbound proteins [91]. This result emphasizes the possible influence of protein dynamics on formation of binding pockets. However, the same proteins were later analysed using the Q-SiteFinder and it was able to predict for IL2-2 and for BCL-X$_L$ binding pockets in the unbound conformation near the native pocket [93].

For many proteins of biological and pharmaceutical importance, no 3D structure is available but very frequently a structure of a protein with similar sequence can be used to generate a homology model of the target protein. Depending on the degree of target-template similarity such homology modeled structures frequently include structural inaccuracies that may interfere with the prediction of putative ligand binding sites. It is of importance to check the performance of prediction methods under realistic conditions where only the unbound structure or a model structure of the target protein is available. This corresponds to an often realistic scenario of a rational drug design project where for a given protein target of interest only a sequence but no 3D structure is available.

In order to obtain an impression on the performance of several of the most recent Web-accessible binding site prediction methods we have compared the application to several protein structures in bound and unbound conformation or even generated homology modeled variants for some of the proteins. The pairs of bound and unbound structures show varying degrees of structural similarity and are listed in Table **3**. The proteins correspond to typical targets for rational drug design. Most ligand binding site prediction methods have been tested on bound and unbound protein structures described in the original publications but the test set and conditions may vary for each case. A direct comparison of available methods applied to the same targets can be useful to obtain a hint of the performance of each method and may indicate overall trends. It should be emphasized that it is not the purpose of the review to provide a comprehensive benchmark test or to provide a quantitative evaluation of the prediction results. The selected target structures do not represent unsolved problems of drug-design but well-known examples to give interested researchers an overview of the available methods and their performances by comparing the predictions with the known binding sites.

The ConSurf-method provides a map of the sequence conservation of residues extracted from a multiple alignment of proteins homologous to the target protein [79]. Although the predicted regions of high sequence conservation frequently overlapped with the binding sites for ligands on the target proteins (Fig. **1**), the conserved surface regions in the example cases extend often beyond the ligand binding site or include parts of the protein surface that are far off the binding site. Hence, the specificity of sequence conservation alone may in general not be sufficient to exactly locate the

Table 3. Protein Test Structures

pdb entry	molecule	state	ligand	Rmsd (Å)[b]
2ANO	E.coli dihydrofolate reductase	bound	Inh. MS-SH08-17	0
5DFR	E.coli dihydrofolate reductase	apo	---	0.7
5DFR_2KGK	DHFR based on 2KGK structure	homology	---	1.6
5DFR_3IA5	DHFR based on 3IA5 structure	homology	---	1.3
3BNZ	Thymidylate synthase	bound	8A inhibitor	0
1NJB	Thymidylate synthase	apo	---	0.8
1EB2	Trypsin inhibitor complex	bound	BPO	0
1BTY	Trypsin inhibitor complex	apo	benzamidine	0.3
1BTY_1GVL	Trypsin based on 1GVL structure	homology	---	0.8
1BTY_1L2E	Trypsin based on 1L2E structure	homology	---	0.9
1FKS	FK506 binding protein	apo	---	1.3
1FKS_2VCD	1FKS based on 2VCD structure	homology	---	2.3
1FKS_2KE0	1FKS based on 2KE0 structure	homology	---	2.6
1D6O	FK506 binding protein	apo	---	0.3
1D7H	FK506 binding protein	bound	DMSO	0
1APB	Arabinose binding protein	bound	arabinose	0
1ANF	maltose-binding protein (MPB)	bound	maltose	0
1OMP	maltose-binding protein (MPB)	apo	---	3.8
1OMP_2FNC	MPB based on 2FNC structure	homology	---	2.4
1OMP_2GHA	MPB based on 2GHA structure	homology	---	2.7
1R2D/1Y2D	BCL-X_L bound vs. unbound	---	---	3.7
1M47/1PY2	IL-2 bound vs. unbound	---	---	2.9
1T4E/1Z1M	MDM2 bound vs. unbound	---	---	2.2

[a]Homology models were generated using the Modeller program with default settings [95] based on a template structure (indicated as second pdb-code in the name given in the first column) with a sequence identity of 30% to 50%.
[b]main chain Rmsd relative to bound structure.

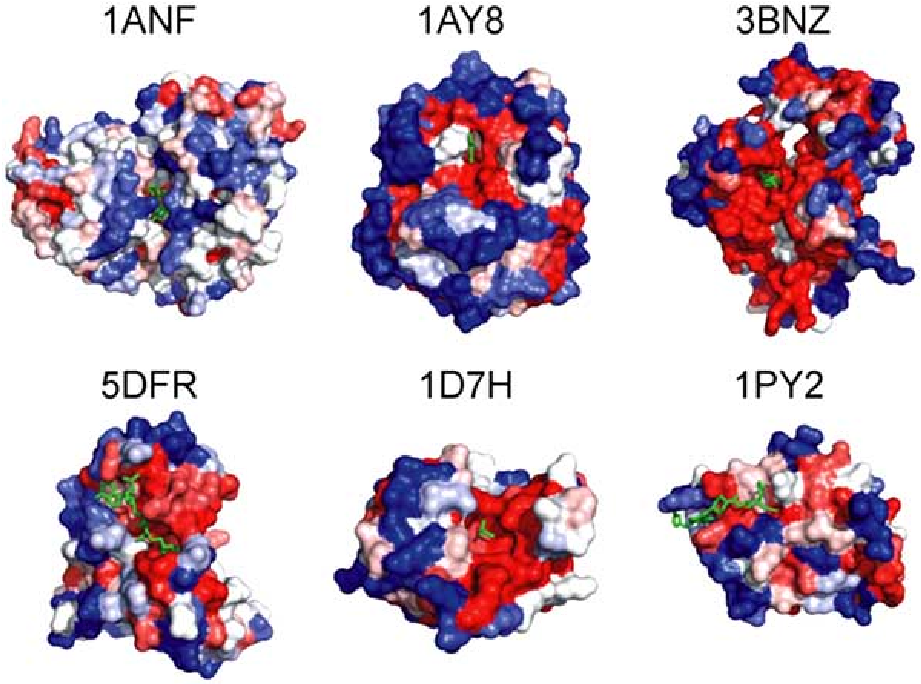

Fig. (1). Result of ConSurf-Server [79] application to six ligand binding protein structures (labeled with pdb-code). The sequence conservation of protein surface residues is color-coded with increasing conservation from blue to red. Bound ligands are indicated as stick models (green).

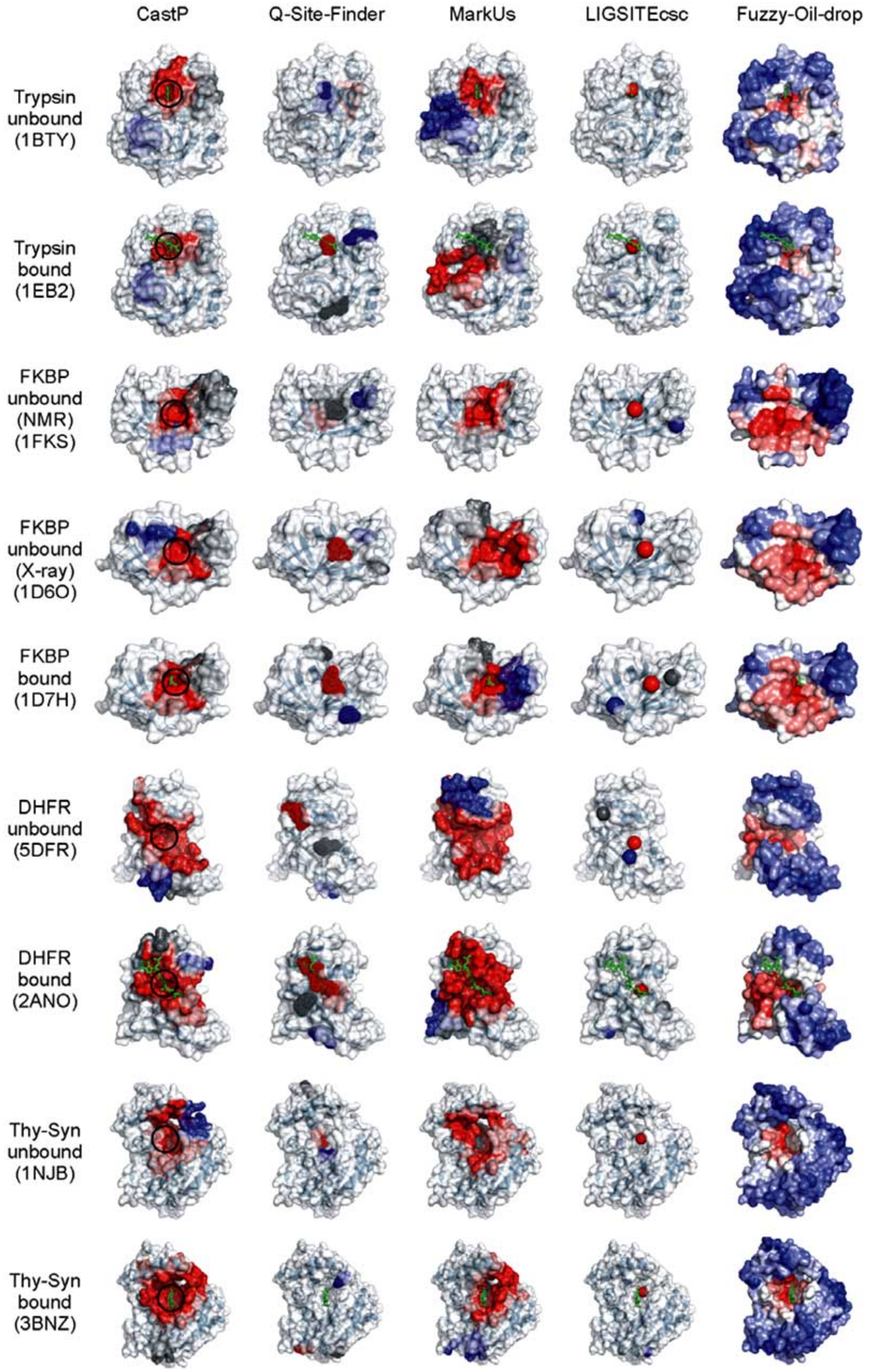

Fig. (2). Results of five ligand binding site prediction servers on four target proteins in bound and unbound conformations. The predicted binding regions are either shown as colored molecular surfaces (CASTp, SCREEN and FOD) with increasing probability from blue to red for a ligand binding site or as colored probes (Q-SiteFinder, LIGSITE^{csc}). Up to three predicted binding sites are shown (red highest score followed by grey and blue). The location of the binding site is encircled black at the most left column. Ligand molecules in the bound structures are shown as stick models (green).

putative ligand binding site (compare ligand position and red colored protein surfaces in Fig. **1**). It should be emphasized that conserved regions not overlapping with the known ligand binding site can be of other functional importance (for example a binding site for another protein). In addition to ConSurf, the programs CASTp [80], Q-SiteFinder [82], LIGSITE^{csc} [85], Mark-Us [86] and Fuzzy-Oil-Drop FOD, [88] were applied on the test proteins (see above and Table **2**). The Web-accessible methods were employed using default parameters. The results were evaluated qualitatively by visual inspection and distance calculations between predicted site and ligand atoms in the native complex.

In the case of using bound structures and for the four protein cases illustrated in Fig. (**2**), all tested methods performed very well in identifying the native binding pocket as the top ranking or one of the top ranking solutions. The most likely site predicted by LIGSITE^{csc} or Q-SiteFinder was close or overlapped with atoms of the ligand in all cases. Only in the case of Thymidylate synthase (Thy_Syn), Q-SiteFinder scored a position close to the binding site at rank 3. For the CASTp, Mark-Us, and FOD methods that encode likely binding sites as B-factors in the pdb-file, the predicted binding regions overlapped or completely included the known binding region in all cases. However, sometimes the predicted top ranked regions significantly extended beyond the known binding region lowering the specificity.

Despite the success of the tested approaches to predict binding regions that overlap with the known binding site, it is important to consider differences in the prediction results that depend on the topology of the binding region. The LIGSITE^{csc} server located the binding region well in all cases. However, it uses a one-sphere prediction which in case of elongated ligands does not give information about a possible orientation of a bound ligand. The Q-SiteFinder returns a large set of spheres which gives a better representation of the complete binding site. In contrast to the sphere-based LIGSITE^{csc} prediction, the Fuzz-Oil-Drop-Model assigns a hydrophobicity distribution to the protein surface. It achieves a high sensitivity (overlap with binding site) but the large size of the predicted surface patch may lower the specificity of the prediction and makes it more difficult to define the exact binding site. Examples 1FKS, 1D6O, and 2ANO in Fig. (**2**) illustrate this problem where protein regions quite distant from the known ligand binding site are included in the prediction.

Interestingly, the overall performance was only slightly worse in the case of employing unbound structures as target proteins. This was even the case for DHFR and Thy_Syn for which the conformational difference between bound and unbound structures near the binding site is more significant compared to Trypsin or FKBP. For example, LIGSITE^{csc} predicted in every unbound structure a pocket that formed at least part of the binding site for the full ligand (Fig. **2**).

For the homology modeling we selected templates with a sequence identity between 30-50% with respect to the target protein (Table **3**). This degree of sequence similarity is typically considered as yielding reliable models with an overall realistic structure [95]. Homology modeling was performed with the Modeller program ([94, 95]; see legend of Table **3**). Two models were generated for four proteins based in each case on two different template proteins. No information on the known structure of the target proteins was included during the comparative modeling step. The Rmsd (main chain) between modeled structures and the corresponding native structures was case depended (1-3 Å, Table **3**). Remarkably, for several of the homology modeled protein structures (e.g. one Trypsin model, one DHFR model, both FKBP models and both MBP -maltose-binding protein- models), the prediction methods performed qualitatively almost as well as for the native protein cases (shown for Q-Site-Finder and LIGSITE^{csc} in Fig. **3**). In case of the FKBP protein models, the Rmsd with respect to the bound native conformation was > 2 Å (Table **3**) but still preserved a detectable pocket. Similar for MBP, the Rmsd of the models exceeded 2 Å but this concerned mostly the global arrangement of the two domains that encompass the binding site. The modeled MBP structures contained an even more open binding cleft (similar to the unbound conformation) that is detected by the prediction programs. Interestingly, for one trypsin model and one DHFR model the prediction of a binding cleft was less precise (Fig. **3**) although the conformational difference of the models from the corresponding bound structure was < 2 Å. This indicates that overall deviation of a model from the native structure is not necessarily a good measure for the usefulness of a model to identify putative binding sites. The degree of change and the type of conformational change near the binding cleft is decisive. Even large changes near the binding cleft (in case of MBP) may result in a detectable pocket (for example a more open pocket) but even small changes that result in closure of the pocket can interfere with the ability to detect the pocket.

Finally, we also tested the prediction methods on the three protein-protein inhibitor cases mentioned above that have been studied previously [91-93]. In these cases the deviation between proteins in unbound and inhibitor bound conformation exceeded the Rmsd for the above discussed cases (Table **3**). For the bound forms, Q-SiteFinder and LIGSITE^{csc} detected pockets that are at least part of the native pockets in all cases (Fig. **4**). For the unbound BCL-X_L, LIGSITE^{csc} was able to predict two pockets very close to the native binding pocket (rank 2 and 3) and one (rank 3) pocket close to the native inhibitor in the case of IL-2. Q-SiteFinder was successful in the case of IL-2 with a detected pocket close to the prediction in the bound form as top ranking prediction and one other predicted pocket close to a second binding regime of the inhibitor as also described by Fuller *et al.* [93]. None of the three top ranked predicted pockets overlapped with the native binding sites in the other two cases (using standard parameters). It was, however, possible to hit regions close to the native binding site if predicted binding sites of lower rank were included (not shown). Fuller *et al.* [93] also indicated recognisable ligand binding pockets using Q-SiteFinder applied to the unbound state of BCL-X_L considering, however, a larger number of predicted putative pocket sites.

The three examples indicate the limits of current pocket detection if the Rmsd (main chain) between bound and unbound structures reaches 2.5 Å or may even exceed 3 Å (BCL-X_L case, Table **3**) and if binding pockets are largely closed in the unbound state. Here, methods that allow for conformational changes, like molecular dynamics simula-

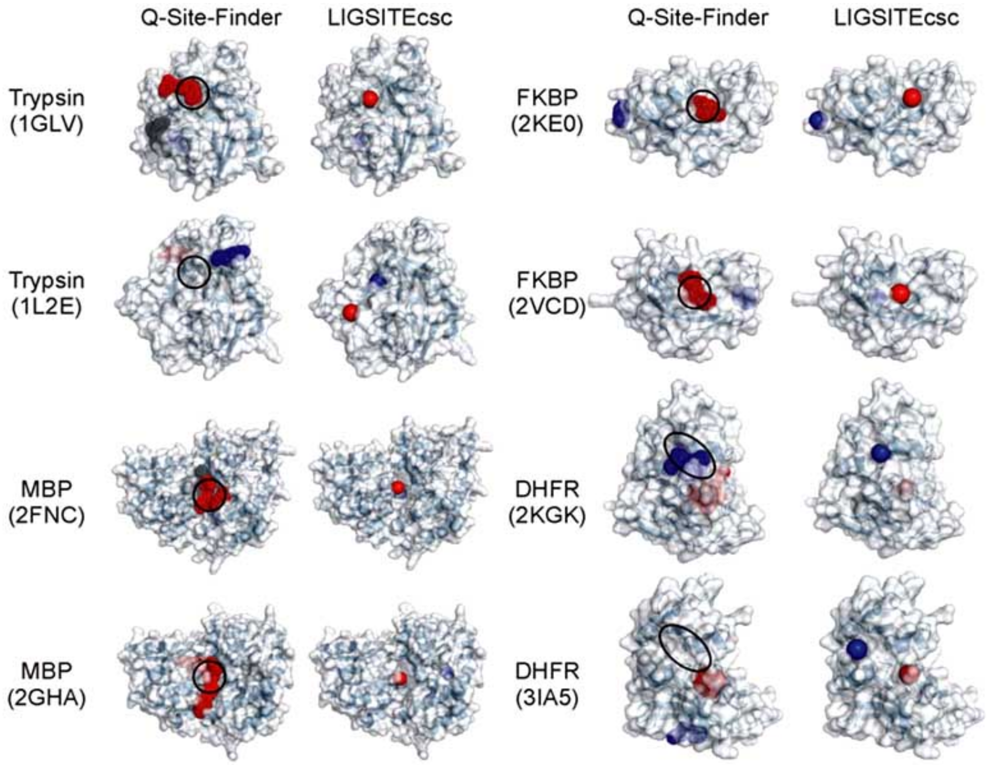

Fig. (3). Performance of Q-SiteFinder and LIGSITE^{csc} on homology modeled protein structures. The name of target protein and the pdb-entry of the template used for homology modeling are indicated. Coloring and labeling scheme is the same as in Fig. (**2**).

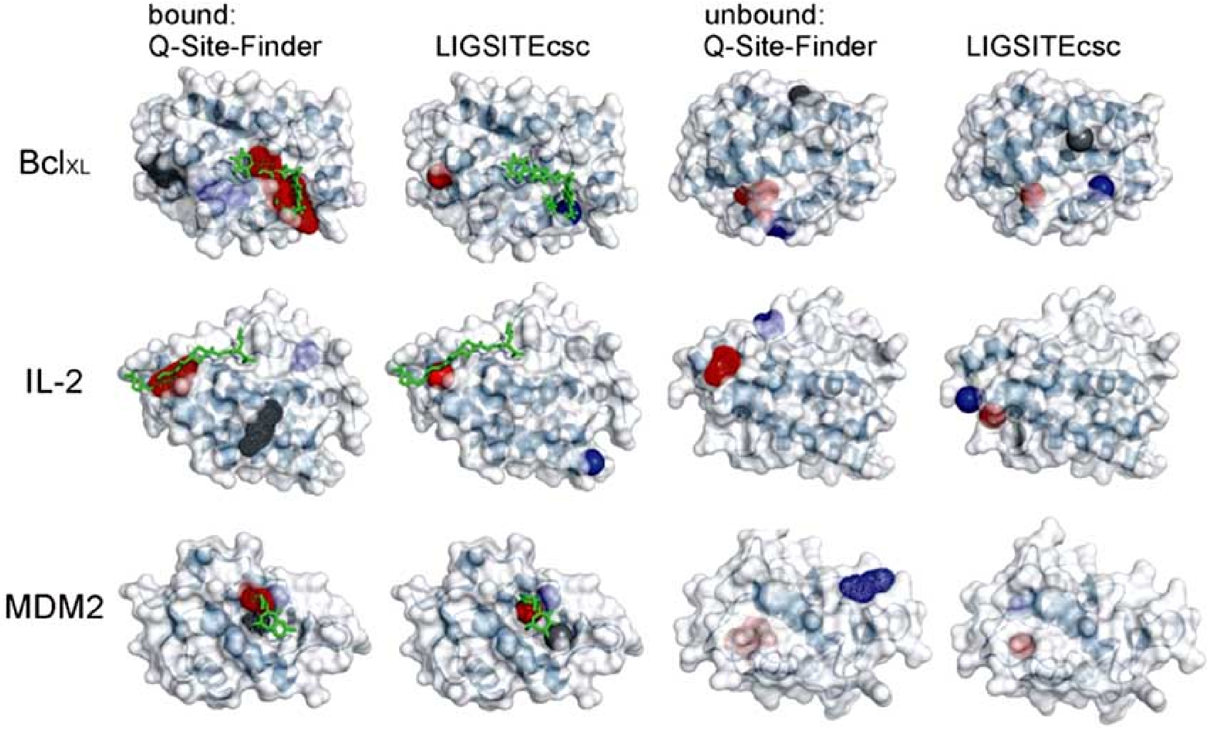

Fig. (4). Application of Q-SiteFinder and LIGSITE^{csc} to detect inhibitor binding sites near protein-protein interaction sites for three proteins in bound and unbound conformations. Coloring and labeling scheme is the same as in Figs. (**2** and **3**).

tions, could become useful to identify transient pockets [91] albeit at much larger computational costs compared to current prediction methods that require seconds or minutes to perform a prediction.

CONCLUSIONS AND OUTLOOK

The availability of an increasing number of protein-protein and protein-ligand complexes has resulted in an improved understanding of the properties of binding sites. In recent years, this knowledge has been used to design many computational prediction tools to identify putative ligand and protein binding sites on proteins. There are significant differences in the architecture of protein-protein interfaces and high affinity sites for binding small drug-like ligands. The latter require, in the majority of cases, a strongly concave binding pocket to maximize the number of close contacts in order to achieve strong interaction. In the case of proteins, the much larger buried interface area allows a wider distribution of interactions and the exclusion of water from a larger interface area may also contribute to the enhanced interaction at hot spot residues located at the interface. It might be especially useful to focus the design of drugs to interfere with protein-protein interactions to those proteins with defined clefts at the protein-protein interface. Further developments in the area of binding site prediction could also aim at predicting not only where a ligand could potentially bind but also which type of ligand might be suitable for a given binding pocket.

It is expected that conformational differences between bound and unbound proteins affect the ability of prediction methods to locate potential protein or ligand binding sites. Our test on a limited set of proteins in unbound and bound conformations indicates that several of the available Web-accessible methods tolerate a certain degree of conformational difference. Encouragingly, for deviations of proteins in bound vs. unbound structure of up to 1.3 Å of the backbone (~2 Å for heavy atoms) most tested programs identified the native ligand binding site as top ranking or among the top ranking predicted pockets. Even for larger deviations or for some of the homology modeled structures with main chain Rmsd up to ~2.5 Å a pocket close to the known site could be identified as a potential ligand binding position. It needs to be emphasized that the number of evaluated cases in the present review does not represent a comprehensive test set. However, it may form a starting point for more systematic and exhaustive studies including more methods and employing larger sets of proteins including homology modeled structures with varying degrees of structural accuracy. The results also indicate that if ligand binding involves structural changes for pocket opening beyond an Rmsd of ~2 Å, new methods may be required that allow for conformational adjustment during the pocket detection phase.

ACKNOWLEDGEMENTS

Support by a grant (Za153/5-3) from the DFG (Deutsche Forschungsgemeinschaft) to M. Z. is gratefully acknowledged.

REFERENCES

[1] Wodak, S. J.; Janin, J. Protein modules and protein-protein interaction. Introduction. *Adv. Protein Chem.*, **2002**, *61*, 9-73.

[2] Reichmann, D.; Rahat, O.; Cohen, M.; Neuvirth, H.; Schreiber, G. The molecular architecture of protein-protein binding sites. *Curr. Opin. Struct. Biol.*, **2007**, *17*, 67-76.

[3] Bonvin A. M. J. J. Flexible protein-protein docking. *Curr. Opin. Struct. Biol.*, **2006**, *16*, 194-200.

[4] May, A.; Sieker F.; Zacharias M. How to efficiently include receptor flexibility during computational docking. *Curr. Comput. Aided Drug Des.*, **2008**, *4*, 143-153.

[5] An, J.; Totrov, M.; Abagyan, R. Comprehensive identification of "druggable" protein ligand binding sites. *Genome Inform.*, **2004**, *15*, 31-41.

[6] Janin, J.; Chothia, C. The structure of protein-protein recognition sites. *J. Biol. Chem.*, **1990**, *265*, 16027-16030.

[7] Jones, S.; Thornton, J. M. Principles of protein-protein interactions. *PNAS*, **1996**, *93*, 13-20.

[8] Tsai, C. J.; Lin, S. L.; Wolfson, H. J.; Nussinov, R. Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci.*, **1997**, *6*, 53-64.

[9] Lo Conte, L.; Chothia, C.; Janin J. The atomic structure of protein-protein recognition sites. *J. Mol. Biol.*, **1999**, *285*, 2177-2198.

[10] Glaser, F.; Steinberg, D.M.; Vakser, I. A.; Ben-Tal, N. Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins*, **2001**, *43*, 89-102.

[11] Janin, J.; Seraphin, B. Genome-wide studies of protein-protein interaction. *Curr. Opin. Struct. Biol.*, **2003**, *13*, 383-388.

[12] Bahadur, R. P.; Chakrabarti, P.; Rodier, F.; Janin J. A dissection of specific and non-specific protein-protein interfaces. *J. Mol. Biol.*, **2004**, *336*, 943-955.

[13] Bahadur, R. P.; Zacharias M. The interface of protein-protein complexes: analysis of contacts and prediction of interactions. *Cell. Mol. Life Sci.*, **2008**, *65*, 1059-1072.

[14] Liang, J.; Edelsbrunner, H.; Woodward, C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.*, **1998**, *7*, 1884-1897.

[15] Laskowski, R. A.; Luscombe, N. M.; Swindells, M. B.; Thornton, J. M. Protein clefts in molecular recognition and function. *Protein Sci.*, **1996**, *5*, 2438-2452.

[16] Mattos, C.; Ringe, D. Locating and Characterizing Binding Sites on Proteins. *Nat. Biotechnol.*, **1996**, *14*, 595-599.

[17] Miller, D. W.; Dill, K. A. Ligand binding to proteins: the binding landscape model. *Protein Sci.*, **1997**, *6*, 2166-2179.

[18] Campbell, S. J.; Gold, N. D.; Jackson, R. M.; Westhead, D. R. Ligand binding: functional site location, similarity and docking. *Curr. Opin. Struct. Biol.*, **2003**, *13*, 389-395.

[19] Vajda, S.; Guarnieri, F. Characterization of protein-ligand interaction sites using experimental and computational methods. *Curr. Opin. Drug Discov. Devel.*, **2006**, *9*, 354-362.

[20] Chiti, F.; Stefani, M.; Taddei, N.; Ramponi, G.; Dobson, C. M. Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature*, **2003**, *424*, 805-808.

[21] Pechmann, S.; Levy, E. D.; Tartaglia, G. G.; Vendruscolo, M. Physicochemical principles that regulate the competition between functional and dysfunctional association of proteins. *Proc. Natl. Acad. Sci. USA*, **2009**, *106*, 10159-10164.

[22] Fernandez-Escamilla, A. M.; Rousseau, F.; Schymkowitz, J.; Serrano, L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.*, **2004**, *22*, 1302-1306.

[23] Chung, J.-L.; Wang, W.; Bourne, P. E. High-throughput identification of Interacting Protein-Protein Binding Sites. *BMC Bioinf.*, **2007**, *8*, 223.

[24] Chakrabarti, P.; Janin J. Dissecting protein-protein interfaces in homodimeric proteins. *Proteins*, **2002**, *47*, 334-343.

[25] Vondrášek, J.; Mason, P. E.; Heyda, J.; Collins, K. D.; Jungwirth P. The Molecular Origin of Like-Charge Arginine−Arginine Pairing in Water. *J. Phys. Chem. B*, **2009**, *113*, 9041-9045.

[26] Clackson, T.; Wells, J. A. A hot spot of binding energy in a hormone-receptor interface. *Science*, **1995**, *267*, 383-386.

[27] Thorn, K. S.; Bogan, A. A. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, **2001**, *17*, 284-285.

[28] Kortemme, T.; Baker, D. A simple physical model for binding energy hot spots in protein-protein complexes. *PNAS*, **2002**, *99*, 14116-14121.

[29] Bogan, A. A.; Thorn, K. S**.** Anatomy of hot spots in protein interfaces. *J. Mol. Biol.*, **1998**, *280*, 1-9.

[30] Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Hot spots - A review of the protein-protein interface determinant amino-acid residues. *Proteins*, **2007**, *68*, 803-812.

[31] Darnell, S. J.; Page, D.; Mitchell, J. C. Automated decision-tree approach to predicting protein-protein interaction hot spots. *Proteins*, **2007**, *68*, 813-823.

[32] Schymkowitz, J.; Borg J.; Stricher, F.; Nys, R.; Rousseau, F.; Serrano, L The FoldX web server: an online force field. *Nucleic Acids Res.,* **2005**, *33*, W382-W388.

[33] Benedix, A.; Becker, C. M.; de Groot, B. L.; Caflisch, A.; Böckmann, R. A. Predicting Free Energy Changes Using Structural Ensembles. *Nat. Methods,* **2009**, *6*, 3-4.

[34] Burgoyne, N. J.; Jackson, R. M. Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces. *Bioinformatics,* **2006**, *22*, 1335-1342.

[35] Lichtarge, O.; Sowa, M. E. Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol.,* **2002**, *12*, 21-27.

[36] Qin, S.; Zhou H.-X. meta-PPISP: a meta web server for protein-protein interaction site prediction. *Bioinformatics,* **2007**, *23*, 2203-2209.

[37] de Vries, S. J.; Bonvin, A. M. J. J. How Proteins Get in Touch: Interface Prediction in the Study of Biomolecular Complexes. *Curr. Protein Pept. Sci.,* **2008**, *9*, 394-406.

[38] Ezkurdia, I.; Bartoli, L.; Fariselli, P.; Casadio, R.; Valencia, A.; Tress, M. L. Progress and challenges in predicting protein-protein interaction sites. *Brief. Bioinformatics,* **2009**, *10*, 233-246.

[39] Bordner, A. J.; Abagyan, R. Statistical analysis and prediction of protein-protein interfaces. *Proteins,* **2005**, *60*, 353-366.

[40] Bock, J. R.; Gough, D. A. Predicting protein-protein interactions from primary structure. *Bioinformatics,* **2001**, *17*, 455-460.

[41] Friedrich, T.; Pils, B.; Dandekar, T.; Schultz, J.; Müller, T. Modelling interaction sites in protein domains with interaction profile hidden Markov models. *Bioinformatics,* **2006**, *21*, 2851-2857.

[42] Ofran, Y.; Rost, B. Protein-protein interaction hotspots carved into sequences. *Bioinformatics,* **2007**, *23*, 13-16.

[43] Valdar, W. S. Scoring residue conservation. *Proteins,* **2002**, *48*, 227-241

[44] Lichtarge, O.; Bourne, H. R.; Cohen, F. E. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.,* **1996**, *257*, 342-358.

[45] Mihalek, I.; Res, I.; Lichtarge, O. A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J. Mol. Biol.,* **2004**, *336*, 1265-1282.

[46] Armon, A.; Graur, D. T.; Ben-Tal, N. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.,* **2001**, *307*, 447-463.

[47] Landau, M.; Mayrose, I.; Rosenberg, Y.; Glaser, F.; Martz, E.; Pupko, T.; Ben-Tal, N. ConSurf 2005: The projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.,* **2005**, *33*, W299-W302.

[48] Pupko, T.; Bell, R. E.; Mayrose I.; Glaser, F.; Ben-Tal, N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics,* **2002**, *18*, S17-S77.

[49] Innis, C. A. siteFiNDER|3D: a web-based tool for predicting the location of functional sites in proteins. *Nucleic Acids Res.,* **2007**, *35*, 489-494.

[50] Caffrey, D. R.; Somaroo, S.; Hughes, J. D.; Mintseris, J.; Huang, E. S. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci.,* **2004**, *13*, 190-202.

[51] de Vries S. J.; van Dijk, A. D. J.; Bonvin, A. M. J. J. WHISCY: What information does surface conservation yield? Application to data-driven docking. *Proteins,* **2006**, *63*, 479-489.

[52] Engelen, S.; Trojan, L. A.; Sacquin-Mora, S.; Lavery, R.; Carbone, A. Joint Evolutionary Trees: detection and analysis of protein interfaces. *PLoS Comput. Biol.,* **2009**, *5*, e1000267.

[53] Neuvirth, H.; Raz, R.; Schreiber, G. ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J. Mol. Biol.,* **2004**, *338*, 181-99.

[54] Jones, S.; Thornton, J. M. Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol.,* **1997**, *272*, 121-132.

[55] Chen, H.; Zhou H.-X. Prediction of interface residues in protein-protein complexes by a consensus neural network method: Test against NMR data. *Proteins,* **2005**, *61*, 21-35.

[56] Porollo, A.; Meller, J. Prediction-based fingerprints of protein-protein interactions. *Proteins,* **2007**, *66*, 630-645.

[57] Liang, S.; Zhang, C.; Liu, S.; Zhou, Y. Protein binding site prediction using an empirical scoring function. *Nucleic Acids Res.,* **2006**, *34*, 3698-3707.

[58] Fernández-Recio, J.; Totrov, M.; Skorodumov, C.; Abagyan, R. Optimal Docking Area: a new method for predicting protein-protein interaction sites. *Proteins,* **2005**, *58*, 134-143.

[59] Fiorucci, S.; Zacharias, M. Prediction of protein-protein interaction sites using electrostatic desolvation profiles. *Biophys. J.,* **2010**, (in press).

[60] Pettit, F. K.; Bare, E.; Tsai, A.; Bowie, J. U. HotPatch: a statistical approach to finding biologically relevant features on protein surfaces. *J. Mol. Biol.,* **2007**, *369*, 863-879.

[61] Kufareva, I.; Budagyan, L.; Raush, E.; Totrov, M.; Abagyan, R. PIER: protein interface recognition for structural proteomics. *Proteins,* **2007**, *66*, 353-366.

[62] Chang, D. T.-H.; Weng, Y.-Z.; Lin, J.-H.; Hwang, M.-J.; Oyang, Y.-J. Protemot: prediction of protein binding sites with automatically extracted geometrical templates. *Nucleic Acids Res.,* **2006**, *34*, W303-309.

[63] Negi, S. S.; Schein, C. H.; Oezguen, N.; Power, T. D.; Braun, W. InterProSurf: a web server for predicting interacting sites on protein surfaces. *Bioinformatics,* **2007**, *23*, 3397-3399.

[64] Murakami,Y.; Jones, S. SHARP^2: protein-protein interaction predictions using patch analysis. *Bioinformatics,* **2006**, *22*, 1794-1795.

[65] Morgan, D. H.; Kristensen, D. M.; Mittleman, D.; Lichtarge, O. ET Viewer: An Application for Predicting and Visualizing Functional Sites in Protein Structures. *Bioinformatics,* **2006,** *22*, 2049-2050.

[66] Bradford, J. R.; Westhead, D. R. Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics,* **2005**, *21*, 1487-1494.

[67] Laurie, A. T. R.; Jackson, R. M. Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening. *Curr. Protein Pept. Sci.,* **2006**, *7*, 395-406.

[68] Brady G. P.; Stouten P. F. Fast prediction and visualization of protein binding pockets with PASS. *J. Comput. Aided Mol. Des.,* **2000**, *14*, 383-401.

[69] Konc, J.; Janezic, D. Protein-protein binding-sites prediction by protein surface structure conservation. J. *Chem. Inform. Mod.,* **2007**, *47*, 940-944.

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

[71] Levitt, D. G.; Banaszak L. J. POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J. Mol. Graph.,* **1992**, *10*, 229-234.

[72] Delaney, J. S. Finding and filling protein cavities using cellular logic operations. *J. Mol. Graph.,* **1992** *10*, 174-177.

[73] Del Carpio C. A.; Takahashi Y.; Sasaki S. A new approach to the automatic identification of candidates for ligand receptor sites in proteins: (I). Search for pocket regions. *J. Mol. Graph.,* **1993**, *11*, 23-29.

[74] Kleywegt, G. J.; Jones, T. A. Efficient Rebuilding of Protein Structures. *Acta Crystallogr. Sect. D: Biol. Crystallogr.,* **1994**, *50*, 178-185.

[75] Laskowski, R. A. SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.,* **1995**, *13*, 323-330.

[76] Peters, K. P.; Fauck, J.; Frömmel, C. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J. Mol. Biol.,* **1996**, *256*, 201-213.

[77] Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: automatic and efficient detection of potential small-molecule binding sites in proteins. *J. Mol. Graph. Model.,* **1997**, *15*, 359-363.

[78] Verdonk, M. L.; Cole, J. C.; Taylor, R. SuperStar: a knowledge-based approach for identifying interaction sites in proteins. *J. Mol. Biol.,* **1999**, *289*, 1093-1108.

[79] Glaser, F.; Pupko, T.; Paz, I.; Bell, R. E.; Bechor-Shental, D.; Martz, E.; Ben-Tal, N. ConSurf: identification of functional regions in proteins by surface- mapping of phylogenetic information. *Bioinformatics,* **2003**, *19*, 163-164.

[80] Dundas, J.; Ouyang, Z.; Tseng, J.; Binkowski, A.; Turpaz, Y.; Liang, J. CASTp: computed atlas of surface topography of proteins

with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res.*, **2006**, *34*, W116-W118.

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

[82] Laurie, A. T. R.; Jackson, R. M. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics*, **2005**, *21*, 1908-1916.

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

[84] Chang, D. T.-H.; Oyang, Y.-J.; Lin, H.-H. MEDock: a web server for efficient prediction of ligand binding sites based on a novel optimization algorithm. *Nucleic Acids Res.,* **2005**, *33*, W233-W238.

[85] Huang, B.; Schroeder, M. LIGSITE^{csc}: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol.,* **2006**, *6*, 19.

[86] Nayal, M.; Honig, B. On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins,* **2006**, *63*, 892-906.

[87] Weisel, M.; Proschak, E.; Schneider, G. PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chem. Cent. J.,* **2007**, *1*, 7.

[88] Byliński, M.; Prymula, K.; Jurkowski, W.; Kochańczyk, M.; Stawowczyk, E.; Konieczny, L.; Roterman, I. Prediction of functional sites based on the fuzzy oil drop model. *PLoS Comput. Biol.,* **2007**, *3*, e94.

[89] Halgren, T. New Method for Fast and Accurate Binding-site Identification and Analysis. *Chem. Biol. Drug Des.*, **2007**, *69*, 146-148.

[90] Byliński, M.; Skolnick, J. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *PNAS,* **2008**, *105*, 129-134.

[91] Eyrisch, S.; Helms, V. Transient Pockets on Protein Surfaces Involved in Protein-Protein Interaction. *J. Med. Chem.*, **2007**, *50*, 2518-2525.

[92] Eyrisch, S.; Helms, V. What induces pocket openings on protein surface patches involved in protein-protein interactions? *J. Comput. Aided Mol. Des*., **2009**, *23*, 73-86.

[93] Fuller, J. C.; Burgoyne, N. J.; Jackson, R. M. Predicting druggable binding sites at the protein-protein interface. *Drug Discov. Today,* **2009**, *14*, 155-161.

[94] Sali, A.; Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.,* **1993**, *234*, 779-815.

[95] Eswar, N.; Eramian, D.; Webb, B.; Shen, M.-Y.; Sali, A. Protein structure modelling using MODELLER. *Methods Mol. Biol.*, **2008**, *426*, 145-159.

Received: November 29, 2009 Revised: February 12, 2010 Accepted: February 14, 2010