Special Issue Article

Journal of Molecular Recognition

Received: 13 May 2009, Revised: 13 July 2009, Accepted: 15 July 2009, Published online in Wiley InterScience: 2009

(www.interscience.wiley.com) DOI:10.1002/jmr.984

Computational approaches to identifying and characterizing protein binding sites for ligand design

Stefan Henrich[a], Outi M. H. Salo-Ahen[a], Bingding Huang[a], Friedrich F. Rippmann[b], Gabriele Cruciani[c,d] and Rebecca C. Wade[a]*

Given the three-dimensional structure of a protein, how can one find the sites where other molecules might bind to it? Do these sites have the properties necessary for high affinity binding? Is this protein a suitable target for drug design? Here, we discuss recent developments in computational methods to address these and related questions. Geometric methods to identify pockets on protein surfaces have been developed over many years but, with new algorithms, their performance is still improving. Simulation methods show promise in accounting for protein conformational variability to identify transient pockets but lack the ease of use of many of the (rigid) shape-based tools. Sequence and structure comparison approaches are benefiting from the constantly increasing size of sequence and structure databases. Energetic methods can aid identification and characterization of binding pockets, and have undergone recent improvements in the treatment of solvation and hydrophobicity. The "druggability" of a binding site is still difficult to predict with an automated procedure. The methodologies available for this purpose range from simple shape and hydrophobicity scores to computationally demanding free energy simulations. Copyright © 2009 John Wiley & Sons, Ltd.

Keywords: ligand binding site; protein pocket; drug design; drug target; druggability

INTRODUCTION

Proteins bind to many types of molecules using a wide variety of binding sites. They have binding sites used by natural ligands, e.g., enzyme active sites and allosteric regulatory sites, as well as "novel" binding sites at which artificial or non-natural ligands, such as drugs, bind. Proteins are often bound to cofactors or are post-translationally modified and these non-protein components can have an important influence on the protein binding sites. The properties of the environment, such as pH and ionic strength, as well as the presence of ordered water molecules, also influence protein binding properties. The complexity of protein–ligand interactions makes the full characterization of binding sites on proteins by computational means a demanding problem. It is also an important problem because computational methods to identify and characterize binding sites are needed, not only to understand molecular interactions in natural and disease states, but also to exploit information on protein structures for the design of compounds with application in the pharmaceutical and biotechnological domains.

Emil Fischer's observation more than one hundred years ago that binding of a substrate to an enzyme is like the insertion of a key into a lock (Fischer, 1894) provides the basis of the majority of methods used to identify protein binding sites today. Shape complementarity between the ligand and the protein is an important determinant of binding and small molecules usually bind in concave pockets on protein surfaces. In the next section, we discuss computational methods to analyze protein shape to detect such pockets.

Proteins are dynamic and their shape is constantly changing. So the rigid lock and key model has been "softened" over time resulting in a range of models that account for protein conformational variation including the zipper model, the hand-in-glove induced fit model, and the conformational selection model (reviewed in (Motiejunas and Wade, 2007; Feldman-Salit and Wade, 2008)). The motions involved in ligand binding range from small side-chain adjustments to large domain motions and their complexity poses a challenge to the identification of transient pockets which may only be present in some protein conformations but which may be particularly important for gaining binding specificity in drug design. These issues are discussed below.

Protein–ligand binding relies not only on shape complementarity but also on physicochemical complementarity. The balance of the van der Waals, electrostatic, hydrogen-bonding, hydrophobic, and solvent interactions must result in energetically favored binding (for review, see (Motiejunas and Wade, 2007; Feldman-Salit and Wade, 2008)). Binding is determined by the

* *Correspondence to: R. C. Wade, EML Research gGmbH, Schloss-Wolfsbrunnenweg 33, 69118 Heidelberg, Germany. E-mail: rebecca.wade@eml-r.villa-bosch.de*

a *S. Henrich, O. M. H. Salo-Ahen, B. Huang, R. C. Wade Molecular and Cellular Modeling Group, EML Research, Schloss-Wolfsbrunnenweg 33, 69118 Heidelberg, Germany*

b *F. F. Rippmann Merck KGaA, Frankfurter Str. 250, D 64293 Darmstadt, Germany*

c *G. Cruciani Molecular Discovery, Via Stoppani, 38, 06087-Ponte San Giovanni-PG, Italy*

d *G. Cruciani Laboratory for Chemometrics and Chemoinformatics, Department of Chemistry, University of Perugia, Via Elce di Sotto 10, 06123 Perugia, Italy*

sum of many contributions, many of which are large opposing energy terms. Furthermore, the binding free energy is often a result of enthalpy–entropy compensation. Computational characterization of the physicochemical properties of protein binding sites makes use of knowledge-based, statistical approaches or energetic models (see below). The difficulty is to develop procedures that are generally applicable across all protein binding sites because protein binding sites vary in the relative importance of the different interactions contributing to binding.

In drug design projects, one usually aims to discover ligands that bind with high affinity and specificity to a given protein target. It is clear that some proteins are easier targets for drug design than others: they have a better "druggability." It would be very useful to be able to computationally estimate a target's druggability before embarking on experimental work in the drug discovery process. The development of tools for this purpose is currently an area of very active research. Protein-structure based methods are, therefore, discussed in the last section of this review.

METHODS TO IDENTIFY PROTEIN BINDING SITES

Ligand-binding site identification methods using static 3D protein structures: geometric approaches

The binding sites for small molecules are usually pockets or crevices on the protein surface or cavities in the protein. Therefore, the identification of pockets and cavities is often the starting point for protein function annotation, protein–ligand docking, and protein structure-based drug design. In recent decades, many computational methods have been developed to detect protein pockets (for a recent review, see (Laurie and Jackson, 2005)). Methods using the geometric characteristics of protein shape include POCKET (Levitt and Banaszak, 1992), LIGSITE (Hendlich *et al*., 1997), LIGSITE^{cs} (Huang and Schroeder, 2006). SURFNET (Laskowski, 1995), CAST (Liang *et al*., 1998), PASS (Brady and Stouten, 2000), PocketPicker (Weisel *et al*., 2007), and Fpocket (Le Guilloux *et al*., 2009). None of these require any knowledge of the ligands.

One of the first geometric methods, POCKET, introduced the idea of protein–solvent–protein events as the key concept for the identification of binding pockets (see Figure 1a). The protein is mapped onto a 3D grid. A grid point is assigned as part of the protein if it is within 3 Å of an atom center; otherwise it is assigned as solvent. Next, the x, y, and z-axes are scanned for pockets which are characterized as a sequence of grid points, starting and ending with the label "protein," and having solvent grid points in between. These sequences are called protein–solvent–protein events. Only grid points that exceed a defined threshold of the number of protein–solvent–protein events are retained for the final pocket prediction. Since the definition of a pocket in POCKET is dependent on the angle of rotation of the protein relative to the axes, LIGSITE extends POCKET by scanning along the four cubic diagonals, in addition to the x, y, and z directions (Hendlich *et al*., 1997). Huang and Schroeder (Huang and Schroeder, 2006) made two extensions to LIGSITE. The first extension is LIGSITE^{cs}, in which surface–solvent–surface events are monitored using the

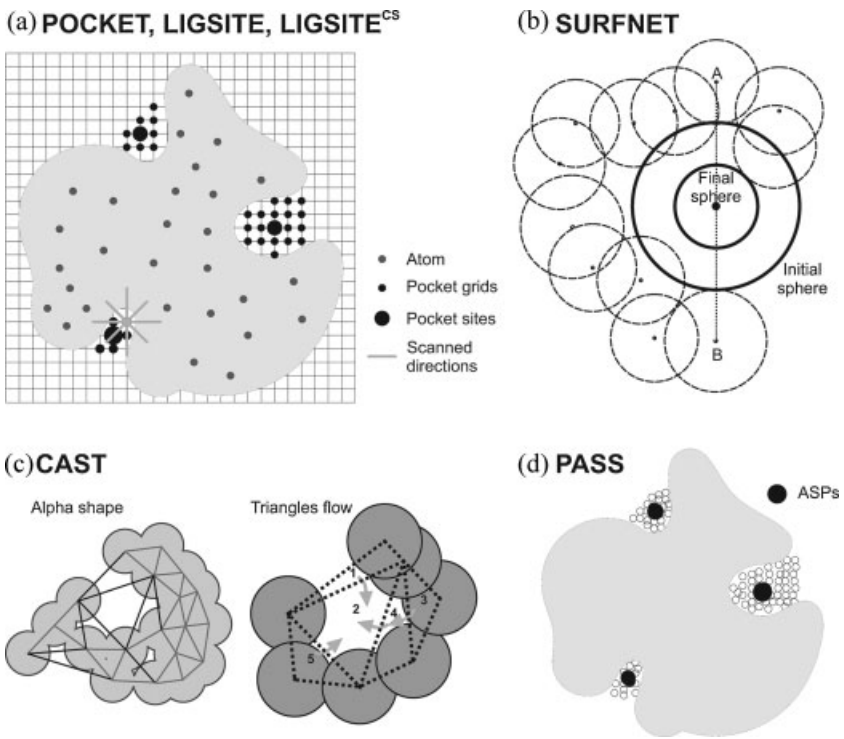

Figure 1. Illustration of different pocket identification methods using protein geometry only, taken from (Huang and Schroeder, 2006) (a) In POCKET, LIGSITE, and LIGSITE^{cs}, the methods scan the grid points outside the protein for protein-solvent-protein and surface-solvent-surface events, respectively. (b) SURFNET places a sphere, which must not contain any atoms, between two atoms. The clustered spheres with maximal volume define the largest pocket. (c) CAST triangulates the surface atoms and clusters triangles by merging small triangles to neighboring large triangles. (d) In PASS, the protein is coated with probe spheres and the probes with many atom contacts are selected. Then coating is repeated until no new probes are kept. The pockets, or active site points, are the probes with the largest number of atom contacts. See text for details.

protein's Connolly surface (Connolly, 1983a), rather than monitoring protein–solvent–protein events. The second extension is LIGSITE^{csc} (LIGSITE^{cs} + Conservation), in which the top three pockets identified by LIGSITE^{cs} are re-ranked by the degree of conservation of the surface residues lining the pockets. This re-ranking improved the probability of the top ranked pocket corresponding to the ligand binding pocket from 67% to 75% for 210 protein structures taken from crystal structures of protein–ligand complexes. PocketPicker (Weisel *et al*., 2007) is another extension of LIGSITE which uses a finer scanning approach to calculate the buriedness-index of grid points (see below).

An alternative approach is taken in SURFNET (Laskowski, 1995), in which a sphere is placed between all pairs of protein atoms so that these two atoms are on opposite sides on the sphere's surface (see Figure 1b). If the sphere contains any other atoms, it is reduced in size until it contains no other atoms. Only spheres with a radius of 1–4 Å are kept. The result of this procedure is a number of separate groups of interpenetrating spheres, both inside the protein and on its surface, which correspond to the pocket sites. SURFNET was tested on 67 enzyme–ligand structures and a ligand was found to be bound in the largest pockets identified in 83% of the cases (Laskowski *et al*., 1996). In a method recently reported by Kawabata and Go (Kawabata and Go, 2007), small and large spheres are used to define the pocket size and depth, respectively. Their PHECOM program was shown to perform better than SURFNET and PASS in detecting binding pockets, although it requires more computing time.

CAST, which is also available as the web server CASTp ((Dundas *et al*., 2006), http://sts.bioengr.uic.edu/castp/), computes a triangulation (see Figure 1c) of the protein's surface atoms using α shapes (Edelsbrunner *et al*., 1995). Then triangles are grouped by letting small triangles flow toward neighboring larger triangles, which act as sinks. The pocket is then defined as a collection of empty triangles. CAST was tested on 51 of the 67 enzyme–ligand complexes from the SURFNET dataset and achieved a success rate of 74%. Instead of using α shapes, the recently published open source program Fpocket (Le Guilloux *et al*., 2009) is based on α spheres and Voronoi tessellation. PASS (Brady and Stouten, 2000) uses probe spheres to fill cavities layer by layer (see Figure 1d). First, an initial coating of the protein with probe spheres is calculated. Each probe has a burial count, which counts the number of atoms within an 8 Å distance. Only probes with a count above a threshold are retained. This procedure is iterated until a layer produces no more new buried probe spheres. Then these probes are clustered to identify a small number of active site points (ASP), which correspond to pocket sites.

Huang and Schroeder systematically compared the performance of CAST, LIGSITE, LIGSITE^{cs}, LIGSITE^{csc}, PASS, and SURFNET on datasets of 48 proteins with unbound and bound structures and of 210 non-redundant proteins with bound structures using the same evaluation criteria (Huang and Schroeder, 2006). Their results showed that, considering the top 3 predictions, these geometric methods achieved success rates of 71–77% for the 48 unbound structures, and 80–87% for the 210 bound structures. By re-ranking the top 3 pocket sites from LIGSITE^{cs} using sequence conservation, LIGSITE^{csc} achieved success rates of 71 and 75% for the 48 unbound and 210 bound structures respectively, for the highest ranked prediction.

The approach of Bock and colleagues is based on a representation of the protein surface as a collection of semi-local shape descriptors: spin-images. Spin-images provide a high-dimensional description of the appearance of a 3D-object in a local reference system (Bock *et al*., 2007). This method performed comparably to SURFNET-ConSurf (Glaser *et al*., 2006) which combines a geometric method with knowledge of sequence. In addition, the technique of Bock and coworkers allows a comparison of protein surface cavities, which can be useful for identifying different binding sites that bind similar ligands.

The MetaPocket metaserver (Huang, 2009) (http://metapocket.eml.org/) combines several of these geometry-based methods by clustering and reranking the top 3 predicted pocket sites from LIGSITE^{cs}, PASS, Q-SiteFinder (Laurie and Jackson, 2005), and SURFNET. MetaPocket thereby improves the success rate from 70 to 75% for the highest ranked prediction for dataset of 210 bound structures. Combining geometric approaches with methods based on physicochemical properties and/or sequence conservation can improve the success rate of binding site prediction (see next section).

Ligand-binding site identification methods using static 3D protein structures: energetic approaches

Some energetic methods simply use van der Waals energies to describe protein shape and give similar results to the geometric methods. Other energetic methods go beyond a shape description to treat other physicochemical terms relevant for ligand binding by considering the interaction with different chemical fragments or molecules as well as solvation effects e.g., GRID (Goodford, 1985), the "sticky spot" method (Ruppert *et al*., 1997) and CS-Map (Landon *et al*., 2007).

In the first category, Q-SiteFinder (Laurie and Jackson, 2005) coats the protein surface with a layer of methyl (-CH_3) probes to calculate van der Waals interaction energies between the protein and probes. Probes with favorable interaction energies are retained and clusters of these probes are ranked based on the number of probes in a cluster. The largest or the energetically most favorable cluster is then ranked as first and considered as a potential ligand-binding site. Morita *et al.* refined Q-SiteFinder to achieve a success rate of 80% for 35 bound structures by using a better probe distribution technique and more comprehensive force field parameters to calculate interaction energies (Morita *et al*., 2008).

In the second category, the GRID algorithm calculates the energy of interaction between a molecular probe and the target at positions on a three-dimensional grid superimposed on the target. The interaction energy is computed as the sum of Lennard-Jones, Coulombic and directional hydrogen-bond energy terms (Boobbyer *et al*., 1989; Wade and Goodford, 1993; Wade *et al*., 1993). The molecular probes can represent diverse chemical fragments, and are free to rotate to optimize the interaction with the target, which itself can include partial flexibility in response to the interaction with the probe. GRID does not identify binding pockets *per se*, but interaction sites in the target of interest, which have been used directly in drug design (von Itzstein *et al*., 1993). Recently, a pocket finding algorithm called FLAPsite has been developed (Cruciani G., Baroni M., Birkso Larsen S., Cross S., unpublished data, available from www.moldiscovery.com), which uses GRID molecular interaction fields (MIFs) to bias a geometric approach toward hydrophobic regions on the protein. A similar algorithm is used to that described for PocketPicker, one of the best performing geometric algorithms (Weisel *et al*., 2007), in which grid points in close proximity to the protein surface are selected and buriedness indices calculated using 30-directional scanning around probes placed at each grid point. In FLAPsite, the buriedness indices are

modified using the GRID "DRY" probe energy potential at the corresponding grid points. The DRY probe is used to identify hydrophobic regions. Neighboring grid points with similar indices above a certain threshold are combined into clusters, and then a morphological refinement technique is applied. First, erosion is applied to the clusters, which has the effect of shrinking all clusters and removing all but the largest. Then a corresponding dilation step is applied to grow the remaining clusters to their original size; this dilation is limited by the protein surface. Using the bound protein structures from 1300 protein–ligand complexes in the PDBBind database (Wang *et al*., 2004), FLAPsite identified a pocket where a ligand bound as top-ranked in 94% of the cases, demonstrating its utility in pocket finding.

An alternative approach is computational solvent mapping (CS-Map) which simulates the experimental Multiple Solvent Crystal Structures (MSCS) method introduced by Ringe and coworkers in 1996 (Mattos and Ringe, 1996). This approach uses small organic molecules as probes and its scoring potential includes a solvation term. Measures to avoid irrelevant local energy minima (false positives) are staged sampling of energy terms (van der Waals interactions are introduced only after the probes have assembled in favorable regions as regards electrostatic and desolvation terms), clustering of the docked probe positions, and the detection of consensus sites where different bound probes overlap (Vajda and Guarnieri, 2006). Another energy-based method that aims at finding such consensus sites for organic solvent molecules employs variable chemical potential grand canonical Monte Carlo simulations (Guarnieri and Mezei, 1996; Clark *et al*., 2006). In these computationally intensive simulations, a relative chemical potential difference is set between a protein in a simulation cell and a bath of organic solvents. The high affinity sites on the protein for a particular small organic compound are identified at the end of a simulation.

Ligand-binding site identification methods using structure and sequence comparison

Many computational methods have been developed to compare protein ligand-binding sites. For example, Cavbase is a method for describing and comparing protein binding pockets on the basis of their geometrical and physicochemical properties (Kuhn *et al*., 2006; Kuhn *et al*., 2007). It has been applied to the functional classification of the binding pockets of the family of protein kinases based on the binding motif in the active sites. Najmanovich *et al.* developed IsoCleft (Najmanovich *et al*., 2008), a graph-matching-based method for the detection of local 3D atomic similarities. The method detects nearly-optimal approximate solutions for the graph-matching problem thus making it possible to compare large sets of atoms such as those obtained from naive geometric definitions of the binding site and discriminate those proteins that bind similar ligands based on local 3D atomic similarities.

Proteins with distant evolutionary relationships may have local sequence similarities on their surfaces leading to similar pocket sites and binding of the same or similar small molecules. Different knowledge-based algorithms, such as Rate4Site (Pupko *et al*., 2002), ConSurf (Glaser *et al*., 2003), and FINDSITE (Brylinski and Skolnick, 2008), recognize sequence and/or structural similarity with proteins of known function to identify binding sites. FINDSITE combines evolution and structure-based approaches in that it uses threading to identify binding site conservation across groups of weakly homologous template structures. Importantly, the method performs comparably when either high-resolution crystal structures or approximate protein models are used.

Ligand-binding site identification methods considering the dynamics of protein structures

Since proteins are flexible by nature, it is not always sufficient to use just one static structure to predict binding sites. A good example of the adaptive nature of a protein–protein interface is given by the cytokine IL-2 (Arkin *et al*., 2003; Hyde *et al*., 2003). Some interface grooves and clefts were shown to form only upon complexation with the binding partner. Therefore, whilst the flexibility and plasticity of a protein can provide novel binding sites for small molecules, they also pose a challenge for structure-based pocket detection.

For some proteins, multiple experimental conformations are available from NMR or X-ray crystallography. However, computational methods that take into account the protein flexibility are of particular importance for proteins with only one (or a few) experimental conformation(s) available. Ensembles of representative protein conformations for binding-site detection can be obtained from classic molecular dynamics (MD) simulations (e.g., (Lin *et al*., 2002; Wong *et al*., 2005; Landon *et al*., 2008)). Schames *et al.* (Schames *et al*., 2004) successfully detected a new binding trench in HIV-integrase from snapshots of a 2-ns MD trajectory. They showed that ligands designed to bind to both the active site and this proximal trench displayed greater selective affinity when able to take advantage of the trench. This led to the discovery and development of the first HIV-integrase inhibitor, raltegravir, by Merck (Summa *et al*., 2008). Furthermore, Frembgen-Kesner and Elcock (Frembgen-Kesner and Elcock, 2006) demonstrated the utility of MD in predicting novel, previously unknown binding sites by MD simulations of the unliganded p38 MAP kinase which reproduced the experimentally detected cryptic drug binding site.

Computationally less costly approaches for the generation of dynamic ensembles of protein conformations have also been developed. The constrained geometric simulation methods ROCK (Lei *et al*., 2004; Zavodszky *et al*., 2004) and FRODA (Wells *et al*., 2005) are examples. For instance, Gohlke succeeded in generating a conformer similar to that found in the bound state of IL-2 by simulating the unbound IL-2 with an enhanced version of FRODA (Gonzalez-Ruiz and Gohlke, 2006). Normal mode analysis (NMA) is a powerful tool that can predict large-amplitude motions of proteins (Brooks and Karplus, 1983; Ma, 2005; Hayward and de Groot, 2008). Usually only the one or two lowest frequency modes are able to describe the majority of the observed protein movements, e.g., conformational changes upon ligand binding (Tama and Sanejouand, 2001; Krebs *et al*., 2002). Protein conformers along the normal mode trajectories can be used for binding site detection or docking (e.g., (Cavasotto *et al*., 2005; Ericksen *et al*., 2009)). In addition, programs such as CONCOORD (de Groot *et al*., 1997) and its extension tCONCOORD (Seeliger *et al*., 2007) generate protein conformers within distance bounds that are calculated on the basis of the interatomic interactions of the initial structure. Recently, Ho and Agard (Ho and Agard, 2009) introduced a computationally efficient MD perturbation method, RIP (Rotamerically Induced Perturbation), which can probe large conformational changes of proteins by applying local perturbations to individual side-chains. Other such enhanced MD methods can also provide representative ensembles of protein conformations.

Recently, Eyrisch and Helms (Eyrisch and Helms, 2007) developed a protocol for identifying transient interface pockets at protein–protein interfaces by applying the PASS algorithm to MD snapshots. They showed that several distinct transient pockets which were large and deep enough to accommodate small molecule inhibitors opened at the interfacial sides of three proteins during the simulations starting with their unbound structures. Many pockets appeared multiple times. Their extended PASS algorithm for analysis of ensembles of protein conformations also calculates pocket properties (volume, polarity, and depth) (see EPOS^{BP}, http://gepard.bioinformatik.uni-saarland.de/software/epos-bp). In a recent comparative study, Eyrisch and Helms (Eyrisch and Helms, 2009) showed that backbone rearrangements together with the side-chain dynamics are important for the opening of transient interface pockets. Additionally, using a more hydrophobic solvent (in this case, methanol) in the MD simulations seems to facilitate the formation of transient binding pockets on protein surfaces. Their results also demonstrate that CONCOORD and NMA methods alone cannot produce as many and as large pockets as tCONCOORD and MD. González-Ruiz and Gohlke (Gonzalez-Ruiz and Gohlke, 2006) suggest that hybrid MD/NMA techniques such as described by Zhang and coworkers (Zhang *et al*., 2003) may be valuable, since the large-scale conformational changes can be amplified by NMA, whilst MD accounts for more localized motions. In summary, methods that can efficiently generate a representative ensemble of protein conformations, possibly using a hydrophobic solvent environment, and detect transient pockets, are useful for structure-based ligand design projects.

Finally, as an allosteric protein binding site is linked to a functional site through a connected network of residues, a potential way for predicting the opening of such sites in proteins could be detecting cooperative coupling of different protein regions. Such approaches include the sequence-based method, statistical coupling analysis (SCA) (Lockless and Ranganathan, 1999) and the COREX algorithm (Hilser and Freire, 1996) that relies on the structure of a protein. Furthermore, conformational strain in the unbound structure can also be an indicator of an allosteric site (Horn and Shoichet, 2004).

APPROACHES TO CHARACTERIZE PROTEIN BINDING SITES

Analysis of the physicochemical properties of binding sites

The basis of molecular recognition is shape and physicochemical complementarity between the receptor and the ligand. Therefore, analysis and characterization of a (putative) binding site is important to aid the design of high affinity ligands. We here consider some of the important properties of binding sites.

Geometry

Some of the geometric approaches for pocket detection can also be used to quantitatively describe the shape of a binding site and to calculate the solvent accessible or molecular surface area (Lee and Richards, 1971; Kuntz *et al*., 1982; Connolly, 1983b; Liang *et al*., 1998). Whereas the program SCREEN (Nayal and Honig, 2006) can evaluate shallow pockets by calculating the difference between the molecular surface and a low-resolution envelope surface (e.g., derived from using a probe of 5 Å radius) for various cavity properties, the program CAST (Liang *et al*., 1998) uses the discrete flow algorithm (Figure 1) and restricts the definition of pockets to those empty concavities having an opening mouth to the bulk solvent. Pure geometric approaches are relatively straightforward and give a first impression of the pockets (e.g., size, shape, surface, and atoms lining the concavity).

Amino acid residue composition

Ligand binding sites are often conserved within protein families and a sequence comparison of different proteins and their pockets can single out residues that are conserved because of their functional importance (Pupko *et al*., 2002). Bartlett and co-workers (Bartlett *et al*., 2002) found catalytic residues occur with a higher frequency as charged residues than as polar or hydrophobic residues. Soga *et al.* analyzed a set of complexes of proteins with drug-like molecules and found different normalized occurrences of amino acids in the ligand binding sites compared to elsewhere on the protein surface, e.g., the occurrence of tryptophan was much higher at ligand binding sites (Soga *et al*., 2007). They constructed an amino acid composition index from their training set and used it to predict binding sites for drug-like molecules and also extended this approach to account for the concurrence of pairs of amino acid residues in a cavity (Soga *et al*., 2008). Carlson *et al.* recently compared the binding sites of high and low affinity ligands to enzymes and non-enzymes (Carlson *et al*., 2008). They found that differences in the amino acid compositions of the pockets seemed to result in higher ligand efficiencies (binding affinity/ligand atom) in the non-enzymes, supporting their suitability as drug targets. In particular, enzyme sites with high-affinity ligands had a high occurrence of glycine residues while non-enzyme sites with high-affinity ligands had a high occurrence of leucine, followed by tyrosine. These data suggest that different approaches may be necessary to characterize the druggability of enzyme and non-enzyme binding sites (see below).

Freely available tools to analyze amino acid residue composition include ConSurf (Glaser *et al*., 2003; Landau *et al*., 2005) (http://consurf.tau.ac.il/) and InterPare (http://interpare.kobic.re.kr/). Whereas, the former can aid the identification of functionally important residues and shows their amino acid conservation to known homologue proteins, the latter gives information about the amino acid composition according to its location (surface, interface, and interior).

Solvation

Proteins are embedded in solvent, mostly water molecules, and some of the molecules are bound in an ordered way to the protein surface. The binding affinity of the surrounding water molecules is not uniform and they have different exchange rates. Based on a study of crystal structures of one protein, they could be classified into three groups: water molecules that are observed in all of the structures, in only one of the structures, and that cannot be identified individually (Ringe, 1995). The binding of a ligand usually requires the displacement of water molecules, which affects the thermodynamics of binding and can be entropically favored or disfavored. These effects can be described by a desolvation energy. Sometimes ligand binding also requires water molecules to mediate complex formation and occupy an interfacial position. Therefore, an important component of the characterization of the binding properties of a protein binding pocket is the identification of the sites at which ordered water

molecules bind. A variety of methods have been developed to predict water binding sites on proteins. One strategy is the identification of conserved sites in series of structures of the same protein which can be done with the WatCH software (Sanschagrin and Kuhn, 1998). Another is to compute the interaction energy between a water probe and the protein as done in the GRID method (Wade and Goodford, 1993). Alternatively, knowledge-based methods, e.g., Waterbase (Günther, 2003), or empirical force fields can be used, e.g., in Fold-X (Schymkowitz *et al*., 2005).

Hydrophobicity

The hydrophobicity of a molecule can be described by its partition coefficient logP, which is given by the ratio of its solubility between octanol and water (Fujita *et al*., 1964). The calculation of the hydrophobic interaction for amino acids based on solution measurements or empirical calculations is complicated and different methods lead to different proposed scales of hydrophobicity (Heiden *et al*., 1993). The strength of the hydrophobic interactions of a protein is influenced by the shape of the pocket and the exposed surface area of residues (Rose *et al*., 1985). There are several approaches for calculating the hydrophobicity of a protein surface or binding pocket. These are either based on the interactions of its three dimensional structure, e.g., HINT (Kellogg *et al*., 1991), MLP (Heiden *et al*., 1993), GRID (using the DRY probe which is akin to an "inverted water molecule") or SuperStar (Verdonk *et al*., 1999), or on approaches using characterization of the surface area by residue (Eisenberg *et al*., 1982) or atomic (Eisenberg *et al*., 1989) hydrophobicities. The latter can be computed and displayed, along with electrostatic potentials, for known interfaces using Molsurfer (Gabdoulline *et al*., 2003). To consider the non-additive effects of hydrophobicity due to the shape and extent of non-polar regions of a pocket, Kelly and Mancera (Kelly and Mancera, 2005) constructed a dot surface surrounding the binding site in which the hydrophobicity value assigned to each dot is dependent not only on the atom type of the closest atom but also on the position, exposure and type of the closest atom to the dots within a threshold distance. The hydrophobicity value is subsequently projected back onto the protein atoms and can be displayed by coloring the atoms according to hydrophobicity.

Electrostatics

The electrostatic potential of a protein can be computed by solving the Poisson-Boltzmann equation using programs such as GRASP (Nicholls *et al*., 1991), UHBD (Madura *et al*., 1995), DelPhi (Rocchia *et al*., 2001) and APBS (Baker *et al*., 2001). For proteins, this requires the assignment of the positions of hydrogen atoms which are usually not resolved in protein crystal structures and which is dependent on the assignment of pK_a values to the titratable residues. This may be done by optimization of the hydrogen-bond network e.g., using WHATIF (Vriend, 1990), or by carrying out pK_a calculations using empirical methods, such as PROPKA (Bas *et al*., 2008) or physics-based methods (for review see (Nielsen, 2009))

Chemical fragment interactions

Experimentally, the binding sites of organic solvent molecules (representing different chemical functionalities that could be used in drug design) to the protein surface can be determined by NMR (Shuker *et al*., 1996) or by screening many crystal structures in which the protein is crystallized in different organic solvents (Ringe, 1995; Mattos and Ringe, 1996; Mattos *et al*., 2006). Most of these probes cluster in known binding sites, and it is assumed their binding locations correspond to the subsites of the binding site with the most complementarity (see Figure 2). Both methods are time consuming and expensive and have limits in

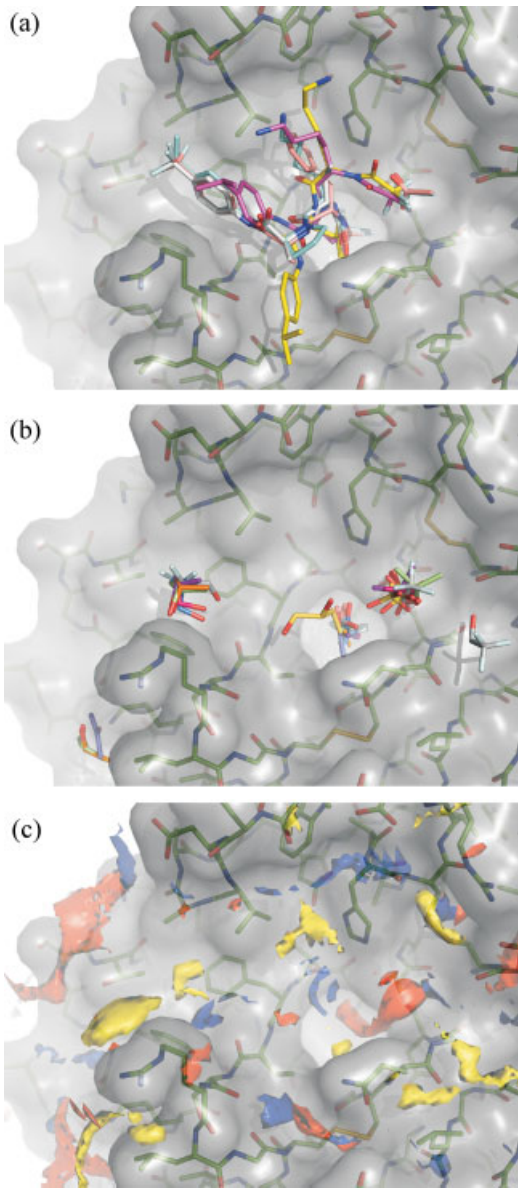

Figure 2. Characterization of the binding properties of porcine elastase by experimental and computational techniques. The binding site of elastase (gray surface, PDB code:1b0e) is shown with (a) five inhibitors superimposed that have different crystallographically determined binding modes (cyan—1ela, pink—1elb, yellow—1elc, light red—1eld, white—1ele) (Mattos *et al*., 1994); (b) bound organic solvent molecules determined crystallographically (blue—acetone + sulphate (2fo9), orange—2-propanol (2foa), green—2-propanol (2fob), sulphate—(2foc), pink—ethanol (2fod), yellow—hexane + sulphate (2foe), 2-propanol (2fof), white—trifluoroethanol (2fog), cyan—trifluroethanol (2foh)) (Mattos *et al*., 2006); and (c) isoenergy contours computed with the GRID programme for various probes (Probes: yellow—DRY (−0.5 kcal/mol), blue—N1 (hydrogen-bond donor, −6.5 kcal/mol), red—O (hydrogen-bond acceptor, −5 kcal/mol)). Note how different ligands exploit different parts of the binding site (a) and how these are detected to varying extents by the small organic molecules (b) and by the GRID probes (c). See text.

applicability. There are a number of computational approaches that are analogous in spirit such as the GRID and CS-Map methods referred to above. Some GRID maps are shown in Figure 2 where they can be compared to the binding modes of inhibitors of elastase and the binding positions of small organic solvent molecules. It can be seen that different ligands exploit different parts of the binding site (Figure 2a) and that these are detected to varying extents by the small organic molecules (Figure 2b) and by the GRID probes (Figure 2c). For example, the blue contours toward the top of figure 2c show a favorable region for a nitrogen probe and this is occupied by a nitrogen in one of the inhibitors (Figure 2a).

MCSS (multiple copy simultaneous search) (Miranker and Karplus, 1991; Stultz and Karplus, 1999) and FTMAP (Brenke *et al*., 2009)) determine consensus binding sites for organic solvent molecules using rigid body docking followed by a clustering and ranking of the docked probes. Seco *et al.* (Seco *et al*., 2009) on the other hand have recently developed a procedure to use molecular dynamics simulations of a constrained protein in a mixture of isopropyl alcohol and water molecules to compute interaction free energies between the protein and the organic molecules. From this, they identify binding sites and estimate their corresponding maximal affinity of a drug-like molecule, thereby obtaining a druggability index from first principles.

Comparison of protein binding sites

Many of the methods to identify binding sites can also be used to compare binding sites. In addition, methods have been developed specifically to compare binding sites. For example, SitesBase (Gold and Jackson, 2006) (http://www.modelling.leeds.ac.uk/sb/) is a database with which a comparison of ligand binding sites with scoring of atomic similarity based on atomic coordinates can be performed via a webserver.

Another approach to comparing binding sites is to compare their MIFs such as the electrostatic potential or the probe interaction grids computed with the GRID method (for review see (Wade, 2005)). A comparison of the MIFs across large protein families can be done using protein interaction property similarity analysis (PIPSA) (Blomberg *et al*., 1999; Wade *et al*., 2001). In this method, the MIFs of the superimposed proteins are compared and quantified by similarity indices. Recently, the webPIPSA webserver (Richter *et al*., 2008) (http://pipsa.eml.org) was released to provide an automated workflow for the modeling of homologous proteins from different species and the computation and comparison of their electrostatic potentials. Such an approach can aid the identification of target sites for the design of selective ligands (Henrich *et al*., 2008). GRID maps can also be subjected to principal component analysis as in the GRID/PCA (Cruciani and Goodford, 1994) and GRID/CPCA (Kastenholz *et al*., 2000) approaches in order to identify regions of homologous protein binding sites important for binding selectivity.

The FLAP algorithm (Baroni *et al*., 2007) provides a means to analyze GRID MIFs to compare two binding sites that may be in structurally unrelated proteins. The MIFs of several GRID probes are computed for each target. These MIFs are then condensed into discrete pharmacophoric points representing favorable and unfavorable interactions using a weighted energy-based and space coverage function. Using these discrete points, all four-point quadruplets are generated, and the resulting pharmacophore quadruplet fingerprint describes the target of interest. In addition to the fingerprints, the GRID MIFs are retained. The targets are then aligned by matching quadruplets in Cartesian space and a field similarity computed using the pre-calculated MIFs. Hence, the fingerprints are used to find matching pharmacophoric regions and the entire fields are used to score the match. The FLAP method can be used for protein structure-based and ligand-based virtual screening as well as protein comparison (which can be limited to the pockets identified with the FLAPsite algorithm described above). Figure 3 provides an illustration of how pocket similarity can be identified between the estrogen receptor and the sarcoplasmic reticulum Ca^{2+} ion channel ATPase protein (SERCA). The side-effects of selective estrogen receptor modulators have been proposed to be due to inhibition of the sarcoplasmic reticulum Ca^{2+} ion channel ATPase protein (Xie *et al*., 2007). Once the cavities have been identified on each protein, FLAP is used to align the pockets using the GRID MIFs, and the best quadruplet match can be visualized. The identified pockets can be clustered using the FLAP MIF-based similarity, enabling more informed target selection and design for selectivity. Further classification using the individual MIF similarities should assist the prediction of the druggability of each pocket.

Detecting druggable ligand-binding sites

In the pharmaceutical industry, computational methods for the assessment of protein druggability are applied in several ways, ranging from target identification and target prioritization to the identification of unwanted targets (i.e., targets that could cause side effects) and to drug repurposing (i.e., the use of an exiting

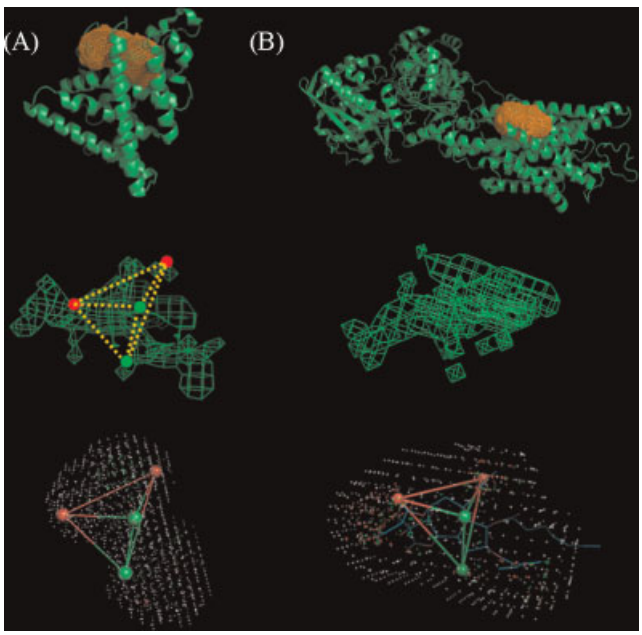

Figure 3. Comparison of two binding sites on structurally unrelated proteins (see text for details): FLAPsite is applied to (A) the estrogen receptor and (B) the SERCA Ca^{2+} ATPase to detect the pockets (top). FLAP condenses the GRID molecular interaction fields (MIFs) (represented by isoenergy contours for the DRY probe in green) into discrete pharmacophore points, and all possible quadruplets are generated (illustrated by a single quadruplet) (middle). After aligning the proteins using quadruplet matches, the GRID MIF similarities are calculated; here the best alignment is shown (bottom). See text.

drug for a new indication, via action on a so far unrecognized target). In particular, in a disease context, there are usually several potential targets. Due to the high costs and high failure rates, the pharmaceutical industry has a strong need to prioritize the targets before working on them experimentally.

The term "protein druggability" is open to different interpretations and definitions. Egner and Hillig (Egner and Hillig, 2008) defined it as the "likelihood of finding a selective, low-molecular weight molecule that binds with high affinity to the target." This can be considered as a useful, minimal definition to which other criteria such as the dependence on cellular and tissue localization of a protein could be added.

It is known that the binding of low-molecular weight molecules to proteins is facilitated by the presence of a complementary concave pocket on the protein surface. Thus, in an early approach to structure-based target identification, Brady and Stouten (Brady and Stouten, 2000) developed their PASS algorithm to identify and assess cavities across the whole Protein Data Bank, with the aim of identifying all druggable targets. PASS is freely available and easy to use, but a practical limitation is that the scripts it generates in order to work with various visualization programs are no longer fully consistent with current versions of these programs. A variety of newer software, most of which is relatively easy to use, permits cavities to be detected with higher accuracy (see above). However, these methods do not take protein flexibility into account, and so it is not surprising that the assessment of even just two independent structures of the same protein can lead to considerably diverging results. Furthermore, several examples have shown that protein flexibility can open up extensions to existing or entirely new binding sites for drugs (see above). Currently, although simulation techniques have been used to identify such transient pockets, there are no established, straightforward and user-friendly procedures available to identify these cases that could be of utmost relevance for drug discovery.

Going beyond pocket identification and shape analysis, several methods have been developed specifically to evaluate the druggability of binding sites. Weisel *et al.* combined calculated druggability-descriptors with the PocketPicker algorithm to estimate how small drug-like molecules would bind to the identified pockets (Weisel *et al*., 2009). Another geometry-based pocket detection method uses a machine learning technique, Random Forests, to distinguish druggable binding sites from non-druggable ones as judged by a computed pocket property profile based on over 400 attributes (Nayal and Honig, 2006). Other programs that assess the binding site druggability are, for example, DrugSite (An *et al*., 2005) which uses a physical potential for *de novo* identification of binding sites, and SiteMap (Halgren, 2009) which characterizes binding sites in a manner resembling the GRID algorithm. Overall, these approaches show that pocket shape and hydrophobicity are the main protein structure-based properties determining druggability (Egner and Hillig, 2008). This has been formulated in a druggability score based on per cent hydrophobic solvent accessible surface area and a scaling factor for ligand efficiency that is dependent on the curvature of the site (Coleman *et al*., 2006; Cheng *et al*., 2007).

An alternative approach to assess protein druggability is to use ligand docking techniques. Zhong and MacKerell applied their docking-based method, which uses the "binding response" of a test set of ligands as a descriptor of druggability, to select putative protein–protein interface binding sites for virtual screening experiments (Zhong and MacKerell, 2007). Li *et al.* (Li *et al*., 2006) developed the web-based tool Target Fishing Dock (TarFisDock (http://www.dddc.ac.cn/tarfisdock/)) for docking a given ligand into the proteins in a database of potential protein targets. Chen *et al.* (Chen and Zhi, 2001) used "inverse" docking and scoring to screen the whole Protein Data Bank for novel targets for known compounds. Their method, INVDOCK, is also applicable to the identification of "unwanted"- or "off"-targets. Such approaches will need to be coupled with knowledgebases on side-effects and toxicity (e.g., (Campillos *et al*., 2008)) so that novel therapeutically promising compound-target interactions can be distinguished for repurposing.

Theoretical methods to assess protein druggability are also being developed in industry. Egner and colleagues at Schering AG, Berlin, now Bayer Schering Pharma, initiated and supported the development of a database named StruDLE by Inpharmatica, now BioFocus, to support an automated, knowledge-based approach for druggability scoring taking geometrical and physicochemical properties into account (personal communication). The StruDLE database was transferred to the European Bioinformatics Institute (EBI) by Biofocus and is planned to be made available on the EBI website in 2009. Machine-learning techniques provide three different scoring schemes to distinguish druggable from non-druggable targets: "tractability," "druggability," and an "Ensemble scoring," a combined score from a variety of support vector machines.

Clearly, further work is necessary to develop tools that improve and integrate the computation of all the relevant properties for determining druggability. Currently, the expert human usually outperforms the computational tools in assessing protein druggability.

SUMMARY

In this review, we have aimed at covering the key concepts for identifying and characterizing protein binding pockets computationally and discussing some of the recent advances as well as limitations of the methods. The review is not comprehensive and we have undoubtedly failed to mention many useful techniques and software tools for studying protein binding sites. Nevertheless, the review shows that there are many powerful computational tools available to the scientist interested in identifying binding pockets and designing ligands for protein targets. Some of these, such as the tools for identifying and comparing binding sites on the basis of geometry, are quick and easy to use. Others, such as the energy- and simulation-based approaches, require more user expertise and time. The performance of the methods is continuing to improve and there is a need for further advances to overcome limitations, such as in the treatment of conformational variability and the protonation states of titratable groups. Apart from the development of the binding site identification and characterization methods themselves, an important task, which is being actively pursued, is the integration of these methods with other methods for high-throughput functional annotation of proteins, drug target identification and drug discovery.

Acknowledgements

We gratefully acknowledge the support of the Klaus Tschira Foundation, European Union (FP 6 STREP project LIGHTS), and Biotechnology Cluster Rhein Neckar (Project INE-TP03).

REFERENCES

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

Arkin MR, Randal M, DeLano WL, Hyde J, Luong TN, Oslob JD, Raphael DR, Taylor L, Wang J, McDowell RS, Wells JA, Braisted AC. 2003. Binding of small molecules to an adaptive protein–protein interface. *Proc. Natl Acad. Sci. USA* **100**: 1603–1608.

Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA. 2001. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl Acad. Sci. USA* **98**: 10037–10041.

Baroni M, Cruciani G, Sciabola S, Perruccio F, Mason JS. 2007. A common reference framework for analyzing/comparing proteins and ligands. fingerprints for ligands and proteins (flap): theory and application. *J. Chem. Inf. Model.* **47**: 279–294.

Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. 2002. Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.* **324**: 105–121.

Bas DC, Rogers DM, Jensen JH. 2008. Very fast prediction and rationalization of pK_a values for protein–ligand complexes. *Proteins: Structure, Function, and Bioinformatics* **73**: 765–783.

Blomberg N, Gabdoulline RR, Nilges M, Wade RC. 1999. Classification of protein sequences by homology modeling and quantitative analysis of electrostatic similarity. *Proteins* **37**: 379–387.

Bock ME, Garutti C, Guerra C. 2007. Effective labeling of molecular surface points for cavity detection and location of putative binding sites. *Comput. Syst. Bioinformatics Conf.* **6**: 263–274.

Boobbyer DNA, Goodford PJ, McWhinnie PM, Wade RC. 1989. New hydrogen-bond potentials for use in determining energetically favorable binding sites on molecules of known structure. *J. Med. Chem.* **32**: 1083–1094.

Brady G, Stouten P. 2000. Fast prediction and visualization of protein binding pockets with PASS. *J. Comput. Aided Mol. Des.* **14**: 383–401.

Brenke R, Kozakov D, Chuang G-Y, Beglov D, Hall D, Landon MR, Mattos C, Vajda S. 2009. Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics* **25**: 621–627.

Brooks B, Karplus M. 1983. Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc. Natl Acad. Sci. USA* **80**: 6571–6575.

Brylinski M, Skolnick J. 2008. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc. Natl Acad. Sci. USA* **105**: 129–134.

Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. 2008. Drug target identification using side-effect similarity. *Science* **321**: 263–266.

Carlson HA, Smith RD, Khazanov NA, Kirchhoff PD, Dunbar JB ,Jr, Benson ML. 2008. Differences between high- and low-affinity complexes of enzymes and nonenzymes. *J. Med. Chem.* **51**: 6432–6441.

Cavasotto CN, Kovacs JA, Abagyan RA. 2005. Representing receptor flexibility in ligand docking through relevant normal modes. *J. Am. Chem. Soc.* **127**: 9632–9640.

Chen YZ, Zhi DG. 2001. Ligand–protein inverse docking and its potential use in the computer search of protein targets of a small molecule. *Proteins* **43**: 217–226.

Cheng AC, Coleman RG, Smyth KT, Cao Q, Soulard P, Caffrey DR, Salzberg AC, Huang ES. 2007. Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* **25**: 71–75.

Clark M, Guarnieri F, Shkurko I, Wiseman J. 2006. Grand canonical Monte Carlo simulation of ligand-protein binding. *J. Chem. Inf. Model.* **46**: 231–242.

Coleman RG, Salzberg AC, Cheng AC. 2006. Structure-based identification of small molecule binding sites using a free energy model. *J. Chem. Inf. Model.* **46**: 2631–2637.

Connolly M. 1983a. Analytical molecular surface calculation. *J. Appl. Cryst.* **16**: 548–558.

Connolly ML. 1983b. Solvent-accessible surfaces of proteins and nucleic acids. *Science* **221**: 709–713.

Cruciani G, Goodford PJ. 1994. A search for specificity in DNA-drug interactions. *J. Mol. Graph.* **12**: 116–129.

de Groot BL, van Aalten DM, Scheek RM, Amadei A, Vriend G, Berendsen HJ. 1997. Prediction of protein conformational freedom from distance constraints. *Proteins* **29**: 240–251.

Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J. 2006. CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res.* **34**: W116–W118.

Edelsbrunner H, Facello M, Fu P, Liang J. 1995. Measuring proteins and voids in proteins. *Proc. 28th Ann. Hawaii Intl. Conf. Syst. Sci.* **5**: 256–264.

Egner U, Hillig RC. 2008. A structural biology view of target drugability. *Expert Opin. Drug Discov.* **3**: 391–401.

Eisenberg D, Wesson M, Yamashita M. 1989. Interpretation of protein folding and binding with atomic solvation parameters. *Chem. Scripta* **29A**: 217–221.

Eisenberg D, Weiss RM, Terwilliger TC, Wilcox W. 1982. Hydrophobic moments and protein structure. *Faraday Symp. Chem. Soc.* **17**: 109–120.

Ericksen SS, Cummings DF, Weinstein H, Schetz JA. 2009. Ligand selectivity of D2 dopamine receptors is modulated by changes in local dynamics produced by sodium binding. *J. Pharmacol. Exp. Ther.* **328**: 40–54.

Eyrisch S, Helms V. 2007. Transient pockets on protein surfaces involved in protein-protein interaction. *J. Med. Chem.* **50**: 3457–3464.

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

Feldman-Salit A, Wade RC. 2008. Molecular recognition: computational analysis and modelling. Wiley *Encyclopedia of Chem. Biol.* DOI: 10.1002/9780470048672.webc354

Fischer E. 1894. Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der deutschen chemischen Gesellschaft* **27**: 2985–2993.

Frembgen-Kesner T, Elcock AH. 2006. Computational sampling of a cryptic drug binding site in a protein receptor: explicit solvent molecular dynamics and inhibitor docking to p38 MAP kinase. *J. Mol. Biol.* **359**: 202–214.

Fujita T, Iwasa J, Hansch C. 1964. A new substituent constant, *IT*, derived from partition coefficients. *J. Am. Chem. Soc.* **86**: 5175–5180.

Gabdoulline RR, Wade RC, Walther D. 2003. MolSurfer: a macromolecular interface navigator. *Nucleic Acids Res.* **31**: 3349–3351.

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

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

Gold ND, Jackson RM. 2006. SitesBase: a database for structure-based protein-ligand binding site comparisons. *Nucleic Acids Res.* **34**: D231–D234.

Gonzalez-Ruiz D, Gohlke H. 2006. Targeting protein-protein interactions with small molecules: challenges and perspectives for computational binding epitope detection and ligand finding. *Curr. Med. Chem.* **13**: 2607–2625.

Goodford PJ. 1985. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **28**: 849–857.

Guarnieri F, Mezei M. 1996. Simulated annealing of chemical potential: a general procedure for locating bound waters. application to the study of the differential hydration propensities of the major and minor grooves of DNA. *J. Am. Chem. Soc.* **118**: 8493–8494.

Günther J. 2003. Entwicklung einer Datenbank und wissensbasierter Vorhersagemethoden zur Untersuchung von Wassermolekülen in Proteinstrukturen sowie ihrer Rolle in der Protein-Liganden-Bindung. Marburg: Philipps-Universität Marburg.

Halgren TA. 2009. Identifying and characterizing binding sites and assessing druggability. *J. Chem. Inf. Model.* **49**: 377–389.

Hayward S, de Groot BL. 2008. Normal modes and essential dynamics. *Meth. Mol. Biol.* **443**: 89–106.

Heiden W, Moeckel G, Brickmann J. 1993. A new approach to analysis and display of local lipophilicity/hydrophilicity mapped on molecular surfaces. *J. Comput. Aided Mol. Des.* **7**: 503–514.

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

Henrich S, Richter S, Wade RC. 2008. On the use of PIPSA to guide target-selective drug design. *ChemMedChem* **3**: 413–417.

Hilser VJ, Freire E. 1996. Structure-based calculation of the equilibrium folding pathway of proteins. Correlation with hydrogen exchange protection factors. *J. Mol. Biol.* **262**: 756–772.

Ho BK, Agard DA. 2009. Probing the flexibility of large conformational changes in protein structures through local perturbations. *PLoS Comput. Biol.* **5**: e1000343.

Horn JR, Shoichet BK. 2004. Allosteric inhibition through core disruption. *J. Mol. Biol.* **336**: 1283–1291.

Huang B, Schroeder M. 2006. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol.* **6**: 19.

Huang B. 2009. MetaPocket: a meta approach to improve protein ligand binding sites prediction. *Omics* **13**: 325–330.

Hyde J, Braisted AC, Randal M, Arkin MR. 2003. Discovery and characterization of cooperative ligand binding in the adaptive region of interleukin-2. *Biochemistry* **42**: 6475–6483.

Kastenholz MA, Pastor M, Cruciani G, Haaksma EE, Fox T. 2000. GRID/CPCA: a new computational tool to design selective ligands. *J. Med. Chem.* **43**: 3033–3044.

Kawabata T, Go N. 2007. Detection of pockets on protein surfaces using small and large probe spheres to find putative ligand binding sites. *Proteins* **68**: 516–529.

Kellogg GE, Semus SF, Abraham DJ. 1991. HINT: a new method of empirical hydrophobic field calculation for CoMFA. *J. Comput. Aided Mol. Des.* **5**: 545–552.

Kelly MD, Mancera RL. 2005. A new method for estimating the importance of hydrophobic groups in the binding site of a protein. *J. Med. Chem.* **48**: 1069–1078.

Krebs WG, Alexandrov V, Wilson CA, Echols N, Yu H, Gerstein M. 2002. Normal mode analysis of macromolecular motions in a database framework: developing mode concentration as a useful classifying statistic. *Proteins* **48**: 682–695.

Kuhn D, Weskamp N, Schmitt S, Hullermeier E, Klebe G. 2006. From the similarity analysis of protein cavities to the functional classification of protein families using cavbase. *J. Mol. Biol.* **359**: 1023–1044.

Kuhn D, Weskamp N, Hullermeier E, Klebe G. 2007. Functional classification of protein kinase binding sites using Cavbase. *ChemMedChem* **2**: 1432–1447.

Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. 1982. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **161**: 269–288.

Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben-Tal N. 2005. ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.* **33**: W299–302.

Landon MR, Lancia DR ,Jr, Yu J, Thiel SC, Vajda S. 2007. Identification of hot spots within druggable binding regions by computational solvent mapping of proteins. *J. Med. Chem.* **50**: 1231–1240.

Landon MR, Amaro RE, Baron R, Ngan CH, Ozonoff D, McCammon JA, Vajda S. 2008. Novel druggable hot spots in avian influenza neuraminidase H5N1 revealed by computational solvent mapping of a reduced and representative receptor ensemble. *Chem. Biol. Drug. Des.* **71**: 106–116.

Laskowski R. 1995. SURFNET: a program for visualizing molecular surfaces, cavities and intermolecular interactions. *J. Mol. Graph.* **13**: 323–330.

Laskowski R, Luscombe N, Swindells M, Thornton J. 1996. Protein clefts in molecular recognition and function. *Protein Sci.* **5**: 2438–2452.

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

Le Guilloux V, Schmidtke P, Tuffery P. 2009. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics* **10**: 168.

Lee B, Richards FM. 1971. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**: 379–400.

Lei M, Zavodszky MI, Kuhn LA, Thorpe MF. 2004. Sampling protein conformations and pathways. *J. Comput. Chem.* **25**: 1133–1148.

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

Li H, Gao Z, Kang L, Zhang H, Yang K, Yu K, Luo X, Zhu W, Chen K, Shen J. others., 2006. TarFisDock: a web server for identifying drug targets with docking approach. *Nucleic Acids Res.* **34**: W219–224.

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

Lin JH, Perryman AL, Schames JR, McCammon JA. 2002. Computational drug design accommodating receptor flexibility: the relaxed complex scheme. *J. Am. Chem. Soc.* **124**: 5632–5633.

Lockless SW, Ranganathan R. 1999. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286**: 295–299.

Ma J. 2005. Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure* **13**: 373–380.

Madura JD, Briggs JM, Wade RC, Davis ME, Luty BA, Ilin A, Antosiewicz J, Gilson MK, Bagheri B, Scott LR. others., 1995. Electrostatics and diffusion of molecules in solution - simulations with the university of houston brownian dynamics program. *Comput. Phys. Commun.* **91**: 57–95.

Mattos C, Bellamacina CR, Peisach E, Pereira A, Vitkup D, Petsko GA, Ringe D. 2006. Multiple solvent crystal structures: probing binding sites, plasticity and hydration. *J. Mol. Biol.* **357**: 1471–1482.

Mattos C, Ringe D. 1996. Locating and characterizing binding sites on proteins. *Nat. Biotech.* **14**: 595–599.

Mattos C, Rasmussen B, Ding X, Petsko GA, Ringe D. 1994. Analogous inhibitors of elastase do not always bind analogously. *Nat. Struct. Mol. Biol.* **1**: 55–58.

Miranker A, Karplus M. 1991. Functionality maps of binding sites: a multiple copy simultaneous search method. *Proteins.* **11**: 29–34.

Morita M, Nakamura S, Shimizu K. 2008. Highly accurate method for ligand-binding site prediction in unbound state (apo) protein structures. *Proteins* **73**: 468–479.

Motiejunas D, Wade RC. 2007. Structural, Energetic, and Dynamic Aspects of Ligand–Receptor Interactions. In: Triggle DJ, Taylor JB (editors.) Comprehensive Medicinal Chemistry II. Oxford; Elsevier: 193–213.

Najmanovich R, Kurbatova N, Thornton J. 2008. Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites. *Bioinformatics* **24**: i105–i111.

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

Nicholls A, Sharp KA, Honig B. 1991. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**: 281–296.

Nielsen JE. 2009. Analyzing enzymatic pH activity profiles and protein titration curves using structure-based p*K*a calculations and titration curve fitting. *Meth. Enzymol.* **454**: 233–258.

Pupko T, Re RB, Mayrose I, Glaser F, Ben T. 2002. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* **18**: s71–s77.

Richter S, Wenzel A, Stein M, Gabdoulline RR, Wade RC. 2008. webPIPSA: a web server for the comparison of protein interaction properties. *Nucleic Acids Res.* **36**: W276–W280.

Ringe D. 1995. What makes a binding site a binding site? *Curr. Opin. Struct. Biol.* **5**: 825–829.

Rocchia W, Alexov E, Honig B. 2001. Extending the applicability of the nonlinear Poisson-Boltzmann equation: multiple dielectric constants and multivalent Ions. *J. Phys. Chem. B* **105**: 6754–6754.

Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH. 1985. Hydrophobicity of amino acid residues in globular proteins. *Science* **229**: 834–838.

Ruppert J, Welch W, Jain AN. 1997. Automatic identification and representation of protein binding sites for molecular docking. *Protein Sci.* **6**: 524–533.

Sanschagrin PC, Kuhn LA. 1998. Cluster analysis of consensus water sites in thrombin and trypsin shows conservation between serine proteases and contributions to ligand specificity. *Protein Sci.* **7**: 2054–2064.

Schames JR, Henchman RH, Siegel JS, Sotriffer CA, Ni H, McCammon JA. 2004. Discovery of a novel binding trench in HIV integrase. *J. Med. Chem.* **47**: 1879–1881.

Schymkowitz JW, Rousseau F, Martins IC, Ferkinghoff-Borg J, Stricher F, Serrano L. 2005. Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *Proc. Natl Acad. Sci. USA* **102**: 10147–10152.

Seco J, Luque FJ, Barril X. 2009. Binding site detection and druggability index from first principles. *J. Med. Chem.* **52**: 2363–2371.

Seeliger D, Haas J, de Groot BL. 2007. Geometry-based sampling of conformational transitions in proteins. *Structure* **15**: 1482–1492.

Shuker SB, Hajduk PJ, Meadows RP, Fesik SW. 1996. Discovering high-affinity ligands for proteins: SAR by NMR. *Science* **274**: 1531–1534.

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

Soga S, Shirai H, Kobori M, Hirayama N. 2008. Chemocavity: specific concavity in protein reserved for the binding of biologically functional small molecules. *J. Chem. Inf. Model.* **48**: 1679–1685.

Stultz CM, Karplus M. 1999. MCSS functionality maps for a flexible protein. *Proteins: Structure, Function, and Genetics* **37**: 512–529.

Summa V, Petrocchi A, Bonelli F, Crescenzi B, Donghi M, Ferrara M, Fiore F, Gardelli C, Gonzalez Paz O, Hazuda DJ, others., 2008. Discovery of raltegravir, a potent, selective orally bioavailable HIV-integrase inhibitor for the treatment of HIV-AIDS infection. *J. Med. Chem.* **51**: 5843–5855.

Tama F, Sanejouand YH. 2001. Conformational change of proteins arising from normal mode calculations. *Protein Eng.* **14**: 1–6.

Vajda S, Guarnieri F. 2006. Characterization of protein–ligand interaction sites using experimental and computational methods. *Curr. Opin. Drug Discov. Devel.* **9**: 354–362.

Verdonk ML, Cole JC, Taylor R. 1999. Superstar: a knowledge-based approach for identifying interaction sites in proteins. *J. Mol. Biol.* **289**: 1093–1108.

von Itzstein M, Wu WY, Kok GB, Pegg MS, Dyason JC, Jin B, Van Phan T, Smythe ML, White HF, Oliver SW. *et al.* 1993. Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature* **363**: 418–423.

Vriend G. 1990. What if: a molecular modeling and drug design program. *J. Mol. Graph.* **8**, 52–6, 29.

Wade RC, Goodford PJ. 1993. 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. *J. Med. Chem.* **36**: 148–156.

Wade RC, Gabdoulline RR, De Rienzo F. 2001. Protein interaction property similarity analysis. *Int. J. Quantum Chem.* **83**: 122–127.

Wade RC, Clark KJ, Goodford PJ. 1993. Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 1. Ligand probe groups with the ability to form two hydrogen bonds. *J. Med. Chem.* **36**: 140–147.

Wade RC. 2005. Calculation and Application of Molecular Interaction Fields. In: Cruciani G (editor). Molecular Interaction Fields: Applications in Drug Discovery and ADME Prediction. Weinheim; WILEY-VCH: 27–42.

Wang R, Fang X, Lu Y, Wang S. 2004. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.* **47**: 2977–2980.

Weisel M, Proschak E, Kriegl JM, Schneider G. 2009. Form follows function: shape analysis of protein cavities for receptor-based drug design. *Proteomics* **9**: 451–459.

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

Wells S, Menor S, Hespenheide B, Thorpe MF. 2005. Constrained geometric simulation of diffusive motion in proteins. *Phys. Biol.* **2**: S127–S136.

Wong CF, Kua J, Zhang Y, Straatsma TP, McCammon JA. 2005. Molecular docking of balanol to dynamics snapshots of protein kinase A. *Proteins* **61**: 850–858.

Xie L, Wang J, Bourne PE. 2007. In silico elucidation of the molecular mechanism defining the adverse effect of selective estrogen receptor modulators. *PLoS Comput. Biol.* **3**: e217.

Zavodszky MI, Lei M, Thorpe MF, Day AR, Kuhn LA. 2004. Modeling correlated main-chain motions in proteins for flexible molecular recognition. *Proteins* **57**: 243–261.

Zhang Z, Shi Y, Liu H. 2003. Molecular dynamics simulations of peptides and proteins with amplified collective motions. *Biophys J* **84**: 3583–3593.

Zhong S, MacKerell AD. 2007. Binding response: a descriptor for selecting ligand binding site on protein surfaces. *J. Chem. Inf. Model.* **47**: 2303–2315.

