Published in final edited form as:
J Struct Funct Genomics. 2011 July ; 12(2): 109–117. doi:10.1007/s10969-011-9110-6.

Beyond structural genomics: computational approaches for the identification of ligand binding sites in protein structures

Dario Ghersi and **Roberto Sanchez**
Department of Structural and Chemical Biology, Mount Sinai School of Medicine, 1425 Madison Avenue, New York, NY 10029, USA

Abstract

Structural genomics projects have revealed structures for a large number of proteins of unknown function. Understanding the interactions between these proteins and their ligands would provide an initial step in their functional characterization. Binding site identification methods are a fast and cost-effective way to facilitate the characterization of functionally important protein regions. In this review we describe our recently developed methods for binding site identification in the context of existing methods. The advantage of energy-based approaches is emphasized, since they provide flexibility in the identification and characterization of different types of binding sites.

Keywords

Binding site; Function; Interaction; Ligand; Prediction; Structure

Introduction

Structural genomics projects have revealed structures for a large number of proteins of unknown function. According to the PSI Structural Genomics Knowledgebase as of November 2010 the PSI Structural Genomics Centers had determined the structures of more than 2,800 such proteins [1]. Hence, computational approaches that contribute to elucidating the function of these proteins would add value to structural genomics efforts.

At the heart of protein function lies the most fundamental of all biological mechanisms, namely the interactions between proteins and their ligands. Visualizing the 3D structure of protein–ligand complexes provides a bridge between protein structure and protein function. It facilitates rational experimental validation of the functional contribution of different protein residues, and it enables structure-based drug design. Thus, modeling protein–ligand complexes is one way to contribute to the functional characterization of protein structures of unknown function. The task of identifying ligand-binding sites can be considered as a precondition to achieve the goal of modeling these complexes.

While comparison of global similarities in protein structures and evolutionary relationships between proteins as inferred from global sequence comparisons are clearly very useful and represent one of the main achievements of molecular bioinformatics, it is possible to find numerous examples of proteins that possess distinct evolutionary histories but that carry out similar functions. In many of these instances one can find striking similarities at the level of

roberto@sanchezlab.org; roberto.sanchez@mssm.edu.
Present Address: D. Ghersi, Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08540, USA

the active site (in the case of enzymes) or more generally in the binding site (for recognition domains). One example of similar binding sites found in vastly different structural folds is that of aromatic cages involved in methyl-lysine recognition [2]. Another example is that of phylogenetically unrelated microbial hydrogenases that possess similar features in their active site [3]. These cases might represent instances of convergent evolution, where unrelated protein domains acquired similar recognition motifs that were particularly effective and were therefore retained by selection. An additional layer of complexity is given by the fact that many proteins whose function has already been determined are actually endowed with more than one function. This phenomenon has been called "moonlighting" and is likely to play an important role in central processes like catalysis, transcription, and gene expression [4, 5]. The concept of moonlighting opens up new spaces to the application of algorithms for the prediction of function from structure and justifies the application of tools for the discovery of functional sites even to previously characterized proteins. Computational methods that are specifically tailored to address the problem of binding site identification and characterization are therefore much needed, and have the potential to go beyond the traditional description of global sequence and structural similarities.

We can envision two possible scenarios where indentifying a protein binding site can provide valuable information in the context of functional annotation. If knowledge of the physiological ligand is available, binding site identification can increase the reliability of docking approaches [6, 7] and therefore help determine the binding mode and the most crucial residues for the interaction. In those cases where the physiological ligand is unknown, identifying putative binding sites represents a time and cost effective way to prioritize residues for mutagenesis experiments. Furthermore, it is still possible to resort to docking (with the advantage that derives from focusing the search on predicted sites) and perform virtual screening of physiological compounds or ligand fragments to generate hypotheses about the possible cognate ligand of the protein under investigation. Finally, knowledge of the binding site allows a comparison of the predicted site against known binding sites for functional annotation even in the absence of evolutionary relationships. A recent review of computational approaches to compare binding sites can be found in Perot et al. [8].

Here we review our work on binding site identification methods in the context of other computational methods for the identification of functionally important regions. We highlight the advantages of structure-based methods, and in particular the benefits of an energy-based approach, which enables the recognition of physicochemical properties that distinguish different types of binding sites.

Computational methods for binding site identification

The most widespread computational methods to carry out binding site identification exploit sequence and structural information, used in isolation or in combined form. Existing methods can be roughly divided into sequence-based, template-based, geometric, and energy-based. The last three make use of structural information.

Sequence-based methods

One of the simplest yet effective ideas behind sequence-based approaches to identify functionally important residues is to exploit the evolutionary information contained in Multiple Sequence Alignments (MSAs) of homologous sequences and extract a subset of residues that show a high degree of conservation. The assumption behind this idea is that the evolutionary pressure acting on functionally important residues will reduce their variability in a protein family. Different conservation measures have been employed, with the majority

of them being cast in the information theoretic framework [9]. An alternative approach that takes advantage of phylogenetic analysis is the “evolutionary trace” method [10]. The idea behind the method is to consider the degree of conservation of residue positions in a protein family in phylogenetically distinct groups. The assumption is that functionally important residues may be conserved in a subgroup but can vary across different subgroups, since these subgroups may have evolved to perform slightly different functions. “Rate4Site” [11] is another approach that takes advantage of phylogenetic information. It relies on estimates of site-specific mutation rates by using a Bayesian approach that, by including prior information into the model, is less sensitive to the number of sequences in the alignment than other conservation-based methods. On the other hand, a clear disadvantage of “Rate4Site” compared to simple information theoretic measures of conservation is the speed of execution, which is substantially lower [9].

Despite their usefulness to infer functionally important residues, all the sequence-based methods suffer from the fundamental limitation of not being able to discriminate between residues that are conserved as part of a binding site from residues that are crucial to protein stability, regulation, or folding. In other words, while binding residues are usually conserved across a protein family, conservation alone is not always a sufficiently specific criterion to identify a binding site, since residues can be conserved for reasons other than binding. Sequence-based methods also do not provide geometric and physicochemical information about the binding site such as area, volume, shape, and molecular interaction properties. To overcome these limitations other approaches have been devised that explicitly take structural information into account.

Template-based and structural-similarity based methods

Template-based methods identify binding sites by comparing them with predefined patterns based on known binding sites. A graph theoretic method for identifying 3D patterns of amino acid side chains was applied to the screening of a set of proteins for the Ser-His-Asp catalytic triad [12]. A similar approach was used to build a network of binding site similarities [13]. A different approach to compare specific arrangements of residues is the TESS algorithm [14], that uses a 3D template acquired by mining the primary literature and containing all the atoms that are essential for an enzyme to perform its function; then, given a query structure the algorithm looks for a match between the query and the 3D template using a geometric hashing formalism. Using a 3D template that contained information for the serine protease active site (again with the well known Ser-His-Asp catalytic triad), the TESS algorithm was able to detect the active site of all the serine proteases, acetylcholinesterase and haloalkane dehalogenase [14].

Recently, a template-based approach has been developed to predict binding sites for phosphorylated ligands [15].

The necessity to provide a template with a well-defined structural arrangement of residues limits, in a sense, the applicability of the comparative approaches described above to enzymes or other molecules with a very conserved active site. Proteins whose function is to bind other proteins or ligands (especially in the case of low affinity binding) are less suitable to the generation of a well-defined template, since they will generally lack a highly conserved arrangement of residues in the binding region.

An alternative approach that also takes advantage of the information available in the Protein Data Bank (PDB) [16] is to identify proteins that are structurally related to a query protein and map the known binding sites onto the uncharacterized sequence. One method that exploits this idea is the threading-based approach FINDSITE [17, 18]. 3DLigandSite [19] automatically builds a model for a given sequence and matches the model against the PDB,

looking for structurally similar proteins with a ligand, which is then superimposed onto the model to infer the binding residues.

Geometric methods

One way to move away from templates is to focus on features of the binding site other than the residues, for example shape. Most of the geometric approaches to identify binding sites in protein structures rely on the assumption that a binding site is usually a cleft or a pocket. For example, a study of 67 protein structures determined that the largest cleft corresponded to a binding site in over 83% of the cases [20]. One of the earliest approaches employed by cleft detection algorithms is the "protein-solvent-protein" concept, used in the POCKET [21] and LIGSITE algorithms [22]. The main idea consists of embedding the protein in a 3D lattice and assigning the grid points to either the protein (if within a predefined distance from an atom center) or the solvent. Pockets are defined as the regions in space that contain points assigned to the "solvent" category and that are surrounded by "protein" points. Later versions of LIGSITE replaced the protein-solvent-protein approach with surface-solvent-surface events (LIGSITE^{cs}), and incorporated a conservation measure to re-rank the putative pockets (LIGSITE^{csc}) [23]. Another well established algorithm for pocket detection is implemented in the SURFNET program [24]. The approach places spheres between all pairs of atoms in such a way that no two atoms are contained inside the spheres. The clustered spheres with the largest volume define the putative pocket. Other methods that rely on the concept of alpha-spheres to identify cavities are APROPOS [25], PASS [26], CAST [27], GHECOM [28] (which identifies pockets by looking for regions on the protein VdW surface that can accommodate small spheres but not large ones), and Fpocket [29, 30].

The program CAVER [31] is specifically tailored to identify channels in proteins, defined as void pathways that connect a cavity buried inside a protein with the solvent on the surface. Two grid-based approaches that evaluate the degree of "buriedness" of points to define cavities are PocketDepth [32] and PocketPicker [33].

Another geometric approach (SplitPocket [34, 35]) exploits the fact that a ligand binding to a pocket will reduce its empty space and perturb the continuity of its surface, thereby creating a 'split pocket'.

Huang and Schroeder carried our a systematic comparison of LIGSITE, CAST, POCKET and SURFNET using a dataset of 210 bound proteins plus 48 proteins for which an unbound form was available [23]. The performance of the methods ranged from 80 to 87% for the bound dataset and from 71 to 77% for the unbound cases. Recently, Huang and colleagues combined several geometric approaches with an energy based approach (see next section) into a metaserver named "MetaPocket" [36], yielding an improvement over each of the individual methods used in isolation.

Despite their usefulness for binding site identification, one of the major shortcomings of all the geometric approaches is represented by the fact that not all binding sites are deep pockets (Fig. 1). Additionally, geometric approaches are not able to distinguish different types of sites, such as hydrophobic versus polar, which may provide additional insights into the possible function of a protein.

Energy-based methods

Energy-based approaches to binding site identification work on the assumption that a binding site is characterized by energetic properties, which stand out from the rest of the protein surface and can be reliably identified. One of the earliest attempts to characterize binding sites using energetic rather than geometric properties is the GRID program [37], that computes a semi-empirical interaction energy between the protein and a set of chemical

probes parameterized to mimic atom types and chemical fragments of pharmaceutical and biological interest. The GRID program is not a binding site identification tool per se, but the interaction energy maps (also known as Molecular Interaction Fields) that are produced by the program can be used for that purpose, with appropriate manipulations. As an example, Q-SiteFinder [38] uses the GRID forcefield to compute an interaction energy map between the protein and a methyl ($-CH_3$) probe and carries out cluster analysis to identify the regions that have the highest total interaction energy. These regions usually correspond to binding sites for drug-like molecules. More recently, Morita et al. [39] improved the performance of this approach by using the AMBER force field for the interaction energy calculations and a more sophisticated two-steps algorithm for clustering. Another recently implemented method which uses the AutoDock [40] forcefield is AutoLigand [41]. Similarly, Ghersi and Sanchez improved on the Q-SiteFinder algorithm by using the GROMOS forcefield and different clustering algorithms [42] and, more importantly, extended the approach beyond the use of the methyl probe to improve the detection of binding sites for non-hydrophobic ligands [42-44]. An alternative energy-based approach to carry out binding site identification on protein structures builds on the experimental technique introduced by Mattos and Ringe called Multiple Solvent Crystal Structures (MSCS) [45]. The idea behind MSCS is to repeatedly soak the protein with different organic solvents and identify the regions involved in binding to these solvents by X-ray crystallography. Vajda and Guarnieri have proposed an equivalent of this procedure, where the solvent mapping is carried out computationally and a consensus site, where different solvents bind favorably, is identified as the putative binding site [46].

Energy-based approaches have the ability to identify different types of binding sites if different chemical probes are used to compute interactions. The use of these multiple probes also has the advantage of providing a preliminary characterization of a binding site in which regions with different chemical characteristics within the same site can be identified [44].

Software

Irrespective of the method used for binding site identification and characterization, the calculations require specialized software. While many methods for binding site identification have been published they are not all equally available. Ideally, all methods should be available as web servers for straightforward analysis of individual proteins, and as downloadable software that can be run in an automated fashion to analyze large sets of proteins. Most of the methods to carry out structure-based identification and characterization of protein binding sites are either provided as web servers or require a commercial license (Table 1). More importantly, no currently available tool provides a combined framework in which one can perform binding site identification and characterization using an energy-based approach. This was the main motivation behind the development of our EasyMIFs and SiteHound tools [42], which provide a comprehensive solution to the energy-based binding site identification including standalone and web server versions [43]. Below we describe the most important characteristics of these tools.

EasyMIFs and SiteHound

EasyMIFs and SiteHound are two software tools that in combination enable the identification and characterization of binding sites in protein structures using an energy-based approach.

EasyMIFs is a simple Molecular Interaction Field (MIF) calculator; and SiteHound, a post processing tool for MIFs that identifies interaction energy clusters corresponding to putative binding sites [42]. EasyMIFs can be used to calculate MIFs for binding site characterization, Quantitative Structure–Activity Relationship (QSAR) studies, selectivity analysis of protein

families, pharmacophoric search, and other applications that require MIFs [47]. It aims to provide a simple and rapid way to characterize a protein structure from a chemical standpoint at the global or local level (e.g. around an active site), returning maps that can be loaded in molecular graphics software. The calculations are carried out *in vacuo* using the GROMOS force field and a distance dependent dielectric [42].

The purpose of SiteHound is to manipulate the output of the EasyMIFs program, and other programs such as Autogrid [40] and GRID [37], in order to predict regions on protein structures that are likely to be involved in binding to ligands. The approach is based on the Q-SiteFinder algorithm [38], but uses a different force field and clustering algorithms suited to ligands of different shapes. The most important difference however lies in the use of multiple probes for the detection of different types of binding sites [42, 44]. The program first filters off all the grid points that have energy values above a user-specified threshold (a negative value) and clusters them according to spatial proximity using single or average linkage agglomerative clustering. Subsequently, the Total Interaction Energy (TIE) of each cluster is computed and this value is used to rank the clusters, from the most negative to the least negative. A test on 77 protein–ligand complexes containing drug-like molecule showed that the correct site is identified among the top three SiteHound clusters in 95% of the cases (79% for unbound proteins) when using the ‘methyl’ probe [6]. Similar accuracy was observed in a set of more than 200 proteins that bind to phosphorylated ligands when using a ‘phosphate oxygen’ probe for binding site identification [44]. One of the advantages of using alternative clustering algorithms is that the binding site identification can be tailored to pre-existing knowledge about the ligand. For example, if the ligand is known to be elongated (such as a peptide) the single-linkage clustering algorithm may results in clusters that more closely resemble the binding site. Conversely, binding sites for smaller, more spherical ligand may be better defined with the average-linkage clustering algorithm (Fig. 2).

The usefulness of binding site prediction by SiteHound was illustrated by using it to guide protein–ligand docking [6]. We developed an automated docking protocol that relies on the SiteHound algorithm to predict putative binding sites, and then carries out docking on the predicted sites. The advantages of isolating the predicted sites and docking the ligands one site at a time lie in improved accuracy and faster running times compared to the blind docking approach, making the protocol suitable for reverse virtual screening experiments. The study showed that not only does binding site identification improve the docking results, but the docking results also facilitate the identification of the correct binding site for a give ligand among the top-three ranking clusters [6].

As mentioned above, one of the most relevant aspects of the EasyMIFs/SiteHound toolkit is its ability to use different types of probes for binding site identification [43, 44]. As shown in Fig. 1, not all binding sites are well-defined pockets or clefts, which can be problematic for geometric approaches. Even energy-based approaches that rely mostly on van der Waals contacts (e.g. when using a methyl group as probe) have difficulty identifying shallow binding sites (Fig. 3a). However, some of these binding sites are still identifiable if other characteristics, such as electrostatics are exploited (Fig. 3b). The use of different probes not only improves the identification of binding sites, but also allows distinguishing different types of binding sites, and different regions within one binding site (Fig. 2) [43, 44]. As such, it provides a much finer tool for the identification and characterization of binding sites, which takes the characterization of protein structures closer to a description of its functional implications, at least at the level of the fundamental mechanism of protein-ligand interaction. The versatility of this approach may prove useful in bridging the gap between structural and functional characterization of the many proteins with known structure but unknown function (Fig. 4).

Conclusions

Computational methods for binding site identification can be of great value in the context of Structural Genomics since they provide a fast and cost-effective way of adding value to protein structures. This is particularly true for the many proteins that have had their structure determined, but remain under-characterized at the functional level. Structure-based approaches to binding site identification are the natural choice in this context since they provide greater accuracy than sequence-based methods. Among the structure-based methods, energy-based approaches provide maximal flexibility in term of identifying and characterizing different types of binding sites. The combination of tools described here (EasyMIFs & SiteHound) provides a freely available framework to carry out binding site identification and characterization through an easy to use web-interface or downloadable software available at http://sitehound.sanchezlab.org. Inclusion of these tools in structure characterization and modeling pipelines may provide additional guidance towards the elucidation of the function of proteins.

References

1. Berman HM, Westbrook JD, Gabanyi MJ, Tao W, Shah R, Kouranov A, Schwede T, Arnold K, Kiefer F, Bordoli L, Kopp J, Podvinec M, Adams PD, Carter LG, Minor W, Nair R, La Baer J. The protein structure initiative structural genomics knowledgebase. Nucleic Acids Res. 2009; 37:D365–D368. [PubMed: 19010965]
2. Brent MM, Marmorstein R. Ankyrin for methylated lysines. Nat Struct Mol Biol. 2008; 15:221–222. [PubMed: 18319736]
3. Shima S, Pilak O, Vogt S, Schick M, Stagni MS, Meyer-Klaucke W, Warkentin E, Thauer RK, Ermler U. The crystal structure of [Fe]-hydrogenase reveals the geometry of the active site. Science. 2008; 321:572–575. [PubMed: 18653896]
4. Copley SD. Enzymes with extra talents: moonlighting functions and catalytic promiscuity. Curr Opin Chem Biol. 2003; 7:265–272. [PubMed: 12714060]
5. Jeffery CJ. Moonlighting proteins: old proteins learning new tricks. Trends Genet. 2003; 19:415–417. [PubMed: 12902157]
6. Ghersi D, Sanchez R. Improving accuracy and efficiency of blind protein-ligand docking by focusing on predicted binding sites. Proteins. 2009; 74:417–424. [PubMed: 18636505]
7. Hetenyi C, van der Spoel D. Towards prediction of functional protein pockets using blind docking and pocket search algorithms. Protein Sci. 2011; 20:880–893. [PubMed: 21413095]
8. Perot S, Sperandio O, Miteva MA, Camproux AC, Villoutreix BO. Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. Drug Discov Today. 2010; 15:656–667. [PubMed: 20685398]
9. Capra JA, Singh M. Predicting functionally important residues from sequence conservation. Bioinformatics. 2007; 23:1875–1882. [PubMed: 17519246]
10. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol. 1996; 257:342–358. [PubMed: 8609628]
11. Mayrose I, Graur D, Ben-Tal N, Pupko T. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. Mol Biol Evol. 2004; 21:1781–1791. [PubMed: 15201400]
12. Artymiuk PJ, Poirrette AR, Grindley HM, Rice DW, Willett P. A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. J Mol Biol. 1994; 243:327–344. [PubMed: 7932758]
13. Zhang Z, Grigorov MG. Similarity networks of protein binding sites. Proteins. 2006; 62:470–478. [PubMed: 16299776]
14. Wallace AC, Borkakoti N, Thornton JM. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. Protein Sci. 1997; 6:2308–2323. [PubMed: 9385633]

15. Parca L, Gherardini PF, Helmer-Citterich M, Ausiello G. Phosphate binding sites identification in protein structures. Nucleic Acids Res. 2011; 39:1231–1242. [PubMed: 20974634]
16. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000; 28:235–242. [PubMed: 10592235]
17. Brylinski M, Skolnick J. FINDSITE: a threading-based approach to ligand homology modeling. PLoS Comput Biol. 2009; 5:e1000405. [PubMed: 19503616]
18. Skolnick J, Brylinski M. FINDSITE: a combined evolution/structure-based approach to protein function prediction. Brief Bioinform. 2009; 10:378–391. [PubMed: 19324930]
19. Wass MN, Kelley LA, Sternberg MJ. 3DLigandSite: predicting ligand-binding sites using similar structures. Nucleic Acids Res. 2010; 38:W469–W473. [PubMed: 20513649]
20. Laskowski RA, Luscombe NM, Swindells MB, Thornton JM. Protein clefts in molecular recognition and function. Protein Sci. 1996; 5:2438–2452. [PubMed: 8976552]
21. Levitt DG, Banaszak LJ. POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. J Mol Graph. 1992; 10:229–234. [PubMed: 1476996]
22. Hendlich M, Rippmann F, Barnickel G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph Model. 1997; 15:359–63. 389. [PubMed: 9704298]
23. Huang B, Schroeder M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. BMC Struct Biol. 2006; 6:19. [PubMed: 16995956]
24. Laskowski RA. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph. 1995; 13(323–30):307–308.
25. Peters KP, Fauck J, Frommel 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. [PubMed: 8609611]
26. Brady GP Jr, Stouten PF. Fast prediction and visualization of protein binding pockets with PASS. J Comput Aided Mol Des. 2000; 14:383–401. [PubMed: 10815774]
27. 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. [PubMed: 16844972]
28. Kawabata T. Detection of multiscale pockets on protein surfaces using mathematical morphology. Proteins. 2010; 78:1195–1211. [PubMed: 19938154]
29. Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection. BMC Bioinformatics. 2009; 10:168. [PubMed: 19486540]
30. Schmidtke P, Le Guilloux V, Maupetit J, Tuffery P. fpocket: online tools for protein ensemble pocket detection and tracking. Nucleic Acids Res. 2010; 38:W582–W589. [PubMed: 20478829]
31. Petrek M, Otyepka M, Banas P, Kosinova P, Koca J, Damborsky J. CAVER: a new tool to explore routes from protein clefts, pockets and cavities. BMC Bioinformatics. 2006; 7:316. [PubMed: 16792811]
32. Kalidas Y, Chandra N. PocketDepth: a new depth based algorithm for identification of ligand binding sites in proteins. J Struct Biol. 2008; 161:31–42. [PubMed: 17949996]
33. Weisel M, Proschak E, Schneider G. PocketPicker: analysis of ligand binding-sites with shape descriptors. Chem Cent J. 2007; 1:7. [PubMed: 17880740]
34. Tseng YY, Dupree C, Chen ZJ, Li WH. SplitPocket: identification of protein functional surfaces and characterization of their spatial patterns. Nucleic Acids Res. 2009; 37:W384–W389. [PubMed: 19406922]
35. Tseng YY, Li WH. Identification of protein functional surfaces by the concept of a split pocket. Proteins. 2009; 76:959–976. [PubMed: 19326458]
36. Huang B. MetaPocket: a meta approach to improve protein ligand binding site prediction. OMICS. 2009; 13:325–330. [PubMed: 19645590]
37. Goodford PJ. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J Med Chem. 1985; 28:849–857. [PubMed: 3892003]

38. Laurie AT, Jackson RM. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics. 2005; 21:1908–1916. [PubMed: 15701681]
39. Morita M, Nakamura S, Shimizu K. Highly accurate method for ligand-binding site prediction in unbound state (apo) protein structures. Proteins. 2008; 73:468–479. [PubMed: 18452211]
40. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem. 1998; 19:1639–1662.
41. Harris R, Olson AJ, Goodsell DS. Automated prediction of ligand-binding sites in proteins. Proteins. 2008; 70:1506–1517. [PubMed: 17910060]
42. Ghersi D, Sanchez R. EasyMIFS and SiteHound: a toolkit for the identification of ligand-binding sites in protein structures. Bioinformatics. 2009; 25:3185–3186. [PubMed: 19789268]
43. Hernandez M, Ghersi D, Sanchez R. SITEHOUND-web: a server for ligand binding site identification in protein structures. Nucleic Acids Res. 2009; 37:W413–W416. [PubMed: 19398430]
44. Ghersi D, Sanchez R. Automated identification of binding sites for phosphorylated ligands in protein structures. 2011 submitted.
45. Mattos C, Ringe D. Locating and characterizing binding sites on proteins. Nat Biotechnol. 1996; 14:595–599. [PubMed: 9630949]
46. Vajda S, Guarnieri F. Characterization of protein-ligand interaction sites using experimental and computational methods. Curr Opin Drug Discov Devel. 2006; 9:354–362.
47. Cruciani, G. Molecular interaction fields applications in drug discovery and ADME prediction. Wiley; Weinheim: 2006.
48. Silvaggi NR, Zhang C, Lu Z, Dai J, Dunaway-Mariano D, Allen KN. The X-ray crystal structures of human alpha-phosphomannomutase 1 reveal the structural basis of congenital disorder of glycosylation type 1a. J Biol Chem. 2006; 281:14918–14926. [PubMed: 16540464]
49. Olson LJ, Dahms NM, Kim JJ. The N-terminal carbohydrate recognition site of the cation-independent mannose 6-phosphate receptor. J Biol Chem. 2004; 279:34000–34009. [PubMed: 15169779]
50. Lee KA, Fuda H, Lee YC, Negishi M, Strott CA, Pedersen LC. Crystal structure of human cholesterol sulfotransferase (SULT2B1b) in the presence of pregnenolone and 3′-phosphoadenosine 5′-phosphate. Rationale for specificity differences between prototypical SULT2A1 and the SULT2BG1 isoforms. J Biol Chem. 2003; 278:44593–44599. [PubMed: 12923182]
51. Biswal BK, Au K, Cherney MM, Garen C, James MN. The molecular structure of Rv2074, a probable pyridoxine 5′-phosphate oxidase from Mycobacterium tuberculosis, at 1.6 angstroms resolution. Acta Crystallogr Sect F Struct Biol Cryst Commun. 2006; 62:735–742.
52. Biswal BK, Cherney MM, Wang M, Garen C, James MN. Structures of Mycobacterium tuberculosispyridoxine 5′-phosphate oxidase and its complexes with flavin mononucleotide and pyridoxal 5′-phosphate. Acta Crystallogr D Biol Crystallogr. 2005; 61:1492–1499. [PubMed: 16239726]
53. Ladner JE, Obmolova G, Teplyakov A, Howard AJ, Khil PP, Camerini-Otero RD, Gilliland GL. Crystal structure of Escherichia coli protein ybgI, a toroidal structure with a dinuclear metal site. BMC Struct Biol. 2003; 3:7. [PubMed: 14519207]
54. Zhong S, Mackerell AD Jr. Binding response: a descriptor for selecting ligand binding site on protein surfaces. J Chem Inf Model. 2007; 47:2303–2315. [PubMed: 17900106]
55. Brenke R, Kozakov D, Chuang GY, Beglov D, Hall D, Landon MR, Mattos C, Vajda S. Fragment-based identification of druggable ‘hot spots’ of proteins using Fourier domain correlation techniques. Bioinformatics. 2009; 25:621–627. [PubMed: 19176554]
56. Gelpi JL, Kalko SG, Barril X, Cirera J, De La Cruz X, Luque FJ, Orozco M. Classical molecular interaction potentials: improved setup procedure in molecular dynamics simulations of proteins. Proteins. 2001; 45:428–437. [PubMed: 11746690]
57. Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comput Biol. 2009; 5:e1000585. [PubMed: 19997483]

58. Kawabata T, Go N. Detection of pockets on protein surfaces using small and large probe spheres to find putative ligand binding sites. Proteins. 2007; 68:516–529. [PubMed: 17444522]

59. An J, Totrov M, Abagyan R. Pocketome via comprehensive identification and classification of ligand binding envelopes. Mol Cell Proteomics. 2005; 4:752–761. [PubMed: 15757999]

60. Till MS, Ullmann GM. McVol - a program for calculating protein volumes and identifying cavities by a Monte Carlo algorithm. J Mol Model. 2010; 16:419–429. [PubMed: 19626353]

61. 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. [PubMed: 16477622]

62. Halgren T. New method for fast and accurate binding-site identification and analysis. Chem Biol Drug Des. 2007; 69:146–148. [PubMed: 17381729]

63. Halgren TA. Identifying and characterizing binding sites and assessing druggability. J Chem Inf Model. 2009; 49:377–389. [PubMed: 19434839]

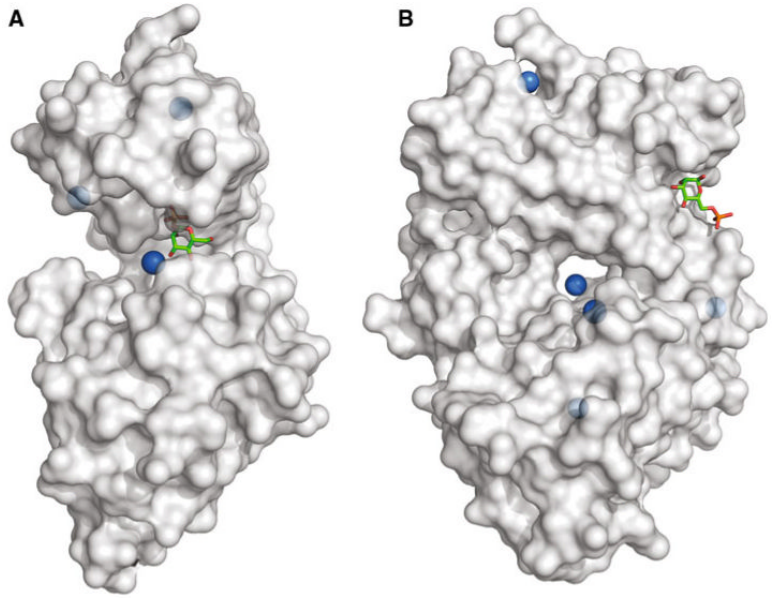

Fig. 1.

Geometric identification of binding sites. **a** Human alpha-Phosphomannomutase in complex with D-mannose 1-phosphate (PDB code: 2fue [48]). The top three binding sites identified by LIGSITEcsc are represented by blue spheres. The ligand binds in a deep crevice that is correctly identified as the largest pocket. **b** Mannose 6-phosphate receptor in complex with mannose 6-phosphate (PDB code: 1sz0 [49]). The binding site is a shallow pocket and in this case is not among the top five sites predicted by LIGSITEcsc. *The blue spheres* show the pockets identified by LIGSITEcsc

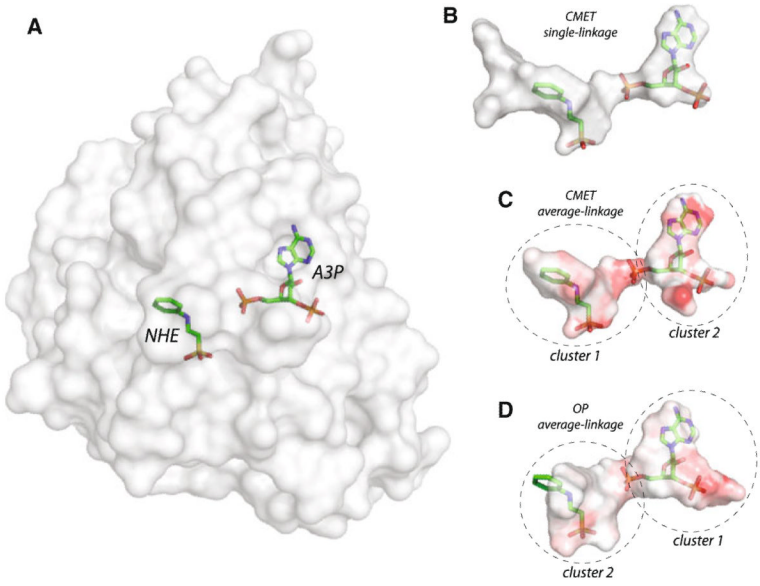

Fig. 2.

Energy-based identification of binding sites using SiteHound. **a** Structure of human pregnenolone sulfotransferase bound to Adenosine-3′-5′-diphosphate (A3P) and 2-[N-cyclohexylamino] ethane sulfonic acid (NHE) (PDB code: 1q1q [50]). **b** Identification of the ligand binding site region using the methyl (CMET) probe and single-linkage clustering. A single cluster covers the two ligands and the entrance to the ligand-binding channel. **c** Identification of the binding sites for NHE (cluster 1) and A3P (cluster 2) using the CMET probe and average linkage clustering. The two binding-sites are identified as separate ligands. The clusters are colored according to the local interaction energy, with red corresponding to stronger interactions. **d** Identification of the binding sites for A3P (cluster 1) and NHE (cluster 2) using the phosphate oxygen (OP) probe. Note the reversal of the cluster ranking, with the cluster for the phosphate-containing ligand (A3P) ranking first, and the most favorable interaction energy spots (*red regions*) being located around the phosphate and sulfonate groups of the ligands

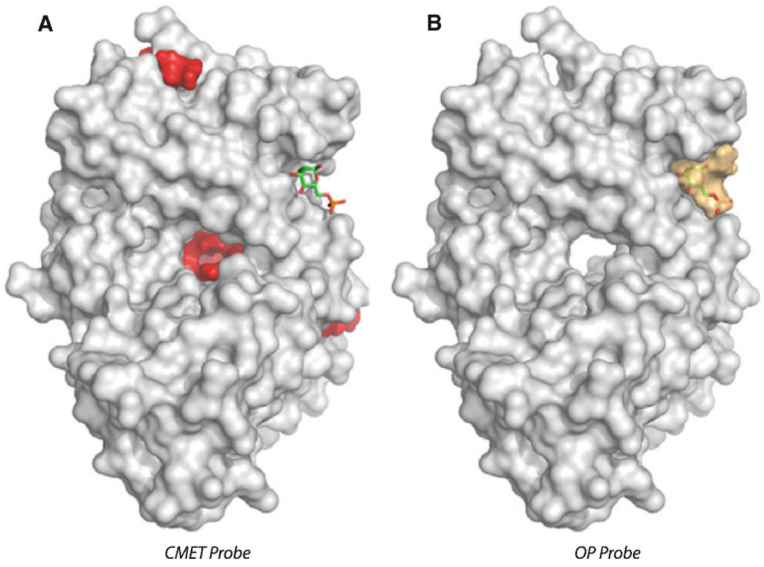

Fig. 3.

Identification of a shallow binding site using the OP probe in SiteHound. Mannose 6-phosphate receptor in complex with mannose 6-phosphate (see Fig. 1). **a** The top 5 clusters identified with the methyl (CMET) probe of SiteHound are shown as red surfaces. Two clusters are located on the opposite side of the structure and not visible. **b** Top 3 clusters identified with the phosphate oxygen (OP) probe of SiteHound are shown as orange surfaces. Two clusters are located on the opposite side of the structure and not visible

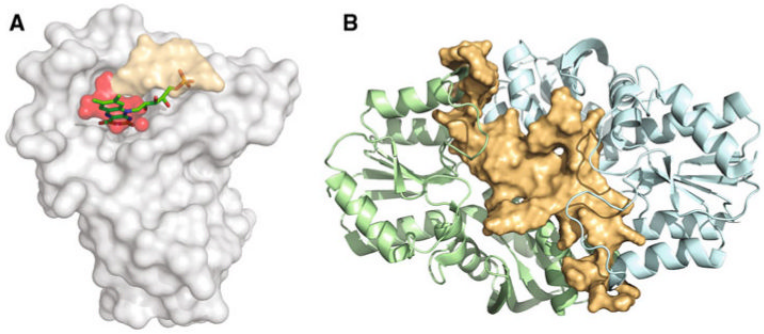

Fig. 4.

Examples of binding site identification in proteins of unknown function. **a** Structure of Rv2074 from *Mycobacterium tuberculosis* (PDB code: 2asf [51]) showing a binding site identified by the CMET probe (*red*) and OP probe (*orange*) of SiteHound. Rv2074 has been suggested as a probable pyridoxine 5′-phosphate oxidase [51]. The clusters delineate a putative binding site that occupies the same position as the Flavin mononucleotide molecule (shown) in the structurally similar Rv1155 pyridoxine 5′-phosphate oxidase [52]. **b** Structure of *Escherichia coli* protein ybgI (PDB code: 1nmo [53]) showing a large OP probe cluster identified using the single-linkage clustering algorithm in SiteHound. This large cluster would seem to be compatible with a suggested role of ybgI in DNA metabolism [53]

Table 1

Software for binding site identification and characterization

Name	Description and reference	Type
APROPOS	Binding site identification (geometrical)—[25]	Currently unavailable
AutoLigand	Binding site identification (energy-based)	Standalone
Binding-response	Binding site identification (geometrical and energy based)—[54]	Standalone
CASTp	Binding site identification (geometrical)—[27]	Web server
CAVER	Binding site identification (geometrical)—[31]	Web server and PyMol plugin
CMIP	Energy-based binding site characterization—[56]	Currently unavailable
ConCavity	Binding site identification (combined)—[57]	Web server and standalone
Evolutionary Trace	Functional residues identification (sequence)—[10]	Web server
FINDSITE	Binding site identification (structural similarity-based)—[17, 18]	Web server and standalone
Fpocket	Binding site identification (geometric)—[29, 30]	Web server and standalone
FTMAP	Fragment-based identification of hot spots—[55]	Web server
GHECOM	Binding site identification (geometrical)—[28, 58]	Web server and standalone
GRID	Energy-based binding site characterization—[37]	Commercial standalone
ICM-PocketFinder	Binding site identification (energy-based)—[59]	Commercial standalone
LIGSITE	Binding site identification (geometrical)—[23]	Web server and standalone
McVol	Binding site identification (geometrical)—[60]	Standalone
PASS	Binding site identification (geometrical)—[26]	Standalone
Pfinder	Phosphate binding site identification (template-based)—[15]	Web server
PocketDepth	Binding site identification (geometrical)—[32]	Web server
PocketFinder	Binding site identification (geometrical)—[22]	Web server
PocketPicker	Binding site identification (geometrical)—[33]	Web server and PyMol plugin
Q-SiteFinder	Binding site identification (energy-based)—[38]	Web server
Screen	Binding site identification (geometrical)—[61]	Web server
SiteHound	Binding site identification (energy-based)—[42]	Web server and standalone
SiteMap	Binding site identification (geometric and energy-based)—[62, 63]	Commercial standalone
SplitPocket	Binding site identification (geometrical)—[34]	Web server
SURFNET	Binding site identification (geometrical)—[24]	Standalone
3DLigandSite	Binding site identification (structural similarity-based)—[19]	Web server