OMICS A Journal of Integrative Biology
Volume 13, Number 4, 2009
© Mary Ann Liebert, Inc.
DOI: 10.1089/omi.2009.0045

MetaPocket: A Meta Approach to Improve Protein Ligand Binding Site Prediction

Bingding Huang

Abstract

The identification of ligand-binding sites is often the starting point for protein function annotation and structure-based drug design. Many computational methods for the prediction of ligand-binding sites have been developed in recent decades. Here we present a consensus method metaPocket, in which the predicted sites from four methods: LIGSITE^{cs}, PASS, Q-SiteFinder, and SURFNET are combined together to improve the prediction success rate. All these methods are evaluated on two datasets of 48 unbound/bound structures and 210 bound structures. The comparison results show that metaPocket improves the success rate from ~70 to 75% at the top 1 prediction. MetaPocket is available at http://metapocket.eml.org.

Introduction

In most cellular processes, proteins interact with other molecules to perform their biological functions. Therefore, knowledge about these interaction sites helps us to understand protein functions. Knowing the location of the functional sites (e.g., substrate or ligand-binding sites of enzymes or receptor proteins) on protein surfaces makes it possible to design inhibitors or antagonists and to introduce targeted mutations aimed at improving the protein function. The protein surface can form pockets that are binding sites of small molecule ligands. Therefore, the identification of pocket sites on the protein surface is often the starting point for protein function annotation and structure-based drug design. Also, proper ligand-binding site detection is a prerequisite for protein–ligand docking. In recent decades, many computational methods have been developed to predict protein–ligand binding sites based on detection of cavities on protein surface. These methods 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), and PocketPicker (Weisel et al., 2007), all of which use pure geometric characteristics and do not require any knowledge of the ligands.

One of the first geometric methods, POCKET (Levitt and Banaszak, 1992), introduced the idea of protein–solvent–protein events as the key concept for the identification (see Fig. 1a). The protein is mapped onto a 3D grid. A grid point is part of the protein if it is within 3 Å of an atom coordinate; otherwise, it is solvent. Next, the x-, y-, and z-axes are scanned for pockets that are characterized as a sequence of grid points, starting and ending with the label "protein" and having a period of solvent grid points in between.

These sequences are called protein–solvent–protein events. Only grid points that exceed a threshold 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. PocketFinder is another implementation of LIGSITE (Laurie and Jackson, 2005). In our previous work (Huang and Schroeder, 2006), we made two extensions to LIGSITE. The first extension is LIGSITE^{cs}, in which we capture more accurate surface–solvent–surface events using the protein's Connolly surface (Connolly, 1983), instead of capturing protein–solvent–protein events. The second extension is LIGSITE^{csc}(LIGSITE^{cs} + Conservation), in which we rerank the pockets identified by the surface–solvent–surface events by the degree of conservation of the involved surface residues. PocketPicker (Weisel et al., 2007) is another extension of LIGSITE using a finer scanning approach to calculate the buriedness-index of grid probes. The buriedness-index is calculated by scanning the protein surroundings along 30 search rays having length of 10 Å and width of 0.9 Å. Then the clustering of grid probes for pocket identification is restricted to those probes with buriedness-indices ranging from 16 to 26. However, the performance of PocketPicker is not much better than that of LIGSITE^{cs}, although it scans more directions (Weisel et al., 2007).

The other geometric approaches for pocket detection are SURFNET, CAST, and PASS. In SURFNET (Laskowski, 1995),

EML Research gGmbH, Schloss-Wolfsbrunnenweg 33, 69118, Heidelberg, Germany.

the key idea is that a sphere that separates two atoms and does not contain any atoms defines a pocket (see Fig. 1b). First, a sphere is placed so that the two given atoms are on opposite sides of the sphere's surface. If the sphere contains any other atoms, it is reduced in size until no more atoms are contained. Only spheres with a radius of 1 to 4 Å are kept. The result of this procedure is a number of separate groups of interpenetrating spheres, called gap regions, both inside the protein and on its surface, which correspond to the protein's cavities and clefts. CAST (Binkowski et al., 2003) computes a triangulation (see Fig. 1c) of the protein's surface atoms using alpha shapes (Edelsbrunner et al., 1995). In the next step, 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. PASS (Brady and Stouten, 2000) uses probe spheres to fill cavities layer by layer (see Fig. 1d). First, an initial coating of the protein with probe spheres is calculated. Each probe has a burial count that 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 each probe is assigned a probe weight, which is proportional to the number of probe spheres in the vicinity and the extent to which they are buried. A small number of active site points (ASP) are then selected by identifying the central probes in regions that contain many spheres with a high burial count. Finally, the retained active site points are ranked by the probe weight.

Besides the purely geometric methods mentioned above, there are other energetic methods. In Q-SiteFinder (Laurie and Jackson, 2005), the protein surface is coated with a layer of methyl (—CH3) 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 energetically most favorable cluster is then ranked first and considered as a potential ligand-binding site. Morita et al. (2008) refined Q-SiteFinder to achieve a higher success rate by using a better probe distribution technique and more suitable force field parameters to calculate interaction energies.

Among these above methods, some are freely available for academic users. The source codes of LIGSITE^{cs}, PocketPicker, and SURFNET are also freely available. For CAST, PocketFinder and Q-SiteFinder, a Web server, is available through which the users can submit a protein structure and visualize the predicted ligand binding sites. PASS provides executable binaries for various operating systems. Therefore, it is of great interest to put all those available methods together to check whether they identify the same pocket sites for the same protein. In this work, we follow the idea of metaPPI, in which five protein–protein binding site predictors were combined together to improve the prediction success rate (Huang and Schroeder, 2008), and propose a meta method called metaPocket that includes four protein–ligand binding site predictors: LIGSITE^{cs}, PASS, Q-SiteFinder, and SURFNET. In all these four methods, the probes around the pocket sites on a protein surface are identified and predicted as potential ligand-binding sites. PocketFinder, PocketPicker, and LIGSITE^{csc} are discarded to avoid biasing because of their similarity to LIGSITE^{cs}. CAST is not taken into account in metaPocket because it identifies the protein atoms forming a pocket rather than the probes around the pocket sites. We will describe the metaPocket approach in detail in the following section.

Materials and Methods

MetaPocket algorithm

For each protein structure, we first use LIGSITE^{cs}, PASS, Q-SiteFinder, and SURFNET to identify pocket sites. For LIGSITE^{cs}, PASS, and SURFNET, we use the executable program to search for the pocket sites on a protein surface. Each identified pocket site is represented as a single probe and has a ranking score. A python script is implemented to submit the protein structures to the Q-SiteFinder server and retrieve the predicted binding sites (probes) automatically. These predicted pocket sites from Q-SiteFinder are represented by probes and are already clustered. For each cluster, the mass center of the probes within it is calculated and is represented as a pocket site ranked by their size. The pocket sites identified by these four methods have different ranking scoring functions. Therefore, it is hard to compare and evaluate the predicted pocket sites directly. To make the ranking scores comparable, a z-score is calculated separately for each site in different methods. Afterward, only the top three pocket sites in each method are taken into further consideration. Therefore, we have a total of 12 pocket sites, which are clustered using a simple hierarchical clustering algorithm, according to their spatial similarity (distance based). Probes within a certain distance threshold (8 Å used here) are grouped together as a cluster. Then each cluster is ranked by a scoring function metaZScore, which is the sum of the z-scores of the pocket sites in a cluster.

Test dataset

In this study, we use the same datasets as those in our previous work. One is a dataset of 48 unbound/bound structures in which both ligand-bound and unbound structures are present. The other one is a nonredundant dataset of 210 ligand-bound only structures, which is derived from the PLD database (Puvanendrampillai and Mitchell, 2003). For a detailed description of these two datasets, see our previous work (Huang and Schroeder, 2006). For the first dataset, the predictions are made for the unbound (apo) structures and checked against the bound structures. In the case of the 210 bound proteins, the ligands are taken away when making predictions and then put back for the evaluation. For a realistic evaluation, we should use the same criteria for all the methods. Each pocket site identified by different methods is represented

Table 1. Success Rate (%) of the Top Three Predictions by Different Methods for 48 Bound/Unbound Structures

	Unbound			*Bound*		
Method	*Top 1*	*Top 2*	*Top 3*	*Top 1*	*Top 2*	*Top 3*
MetaPocket	75	85	90	83	94	96
LIGSITE^{cs}	71	79	85	81	90	92
PASS	58	67	75	58	81	85
Q-SiteFinder	52	60	75	75	83	90
SURFNET	42	58	62	42	56	60

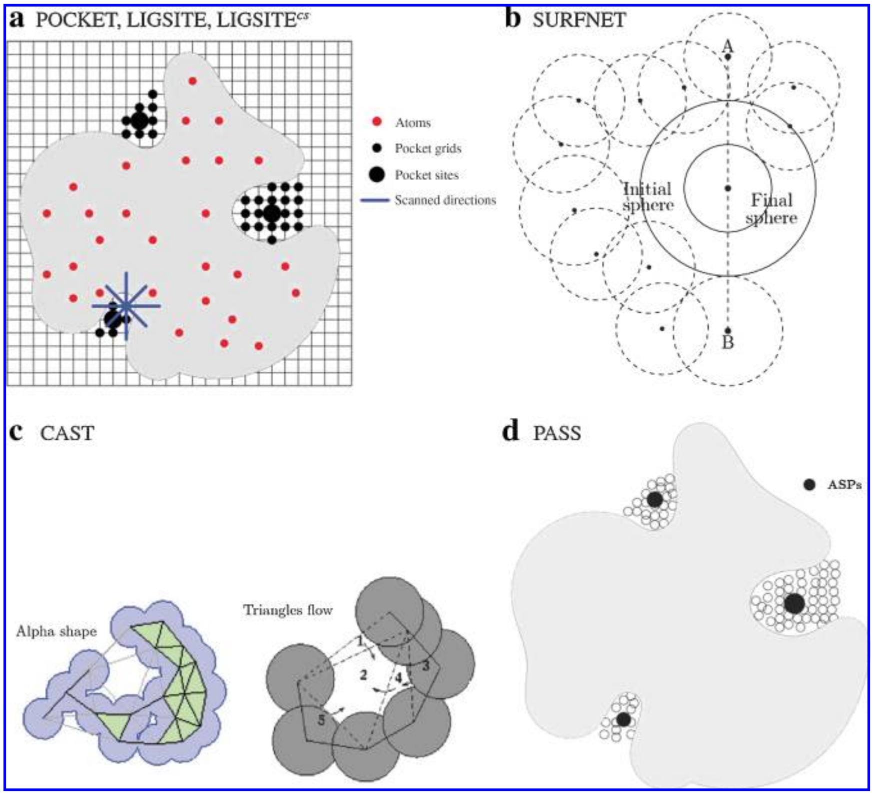

FIG. 1. Illustration of different pocket identification methods, taken from Huang and Schroeder, (2006). (**a**) POCKET, LIGSITE, and LIGSITE^{cs} scan the grid for protein–solvent–protein and surface–solvent–surface events, respectively. POCKET uses three, LIGSITE and LIGSITE^{cs} seven directions. POCKET and LIGSITE use atom coordinates, while LIGSITE^{cs} uses the Connolly surface. (**b**) SURFNET places a sphere, which must not contain any atoms, between two atoms. The 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**) PASS coats the protein with probe spheres, selects the probes with many atom contacts, and then repeats coating until no new probes are kept. The pockets, or active site points, are the probes with the largest number of atom contacts. Q-SiteFinder is similar to LIGSITE, but the ranking of the pocket sites is the sum of the van der Waals interaction energy between the probes and protein atoms.

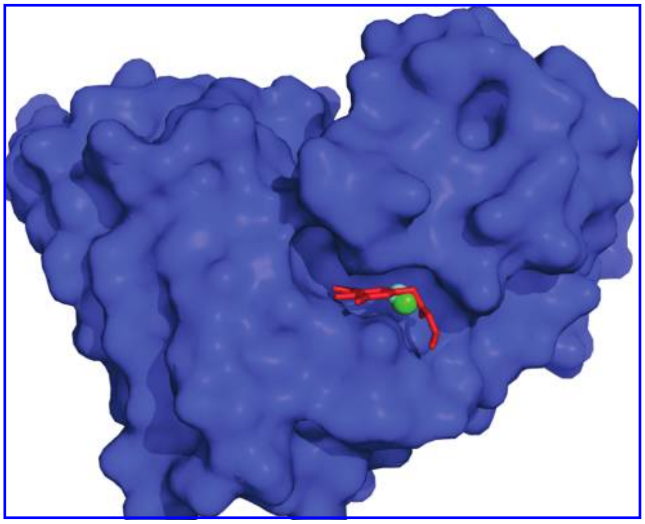

FIG. 2. Ligand binding site on protein 1a6u (unbound)/1a6w (bound). The ligand NIP (red) is bound to a pocket site, which is predicted as the top one by Q-SiteFinder (green sphere) and top five by LIGSITE^{cs} (cyan). LIGSITE, PASS, and SURFNET fail to identify this ligand binding site within the top three predictions.

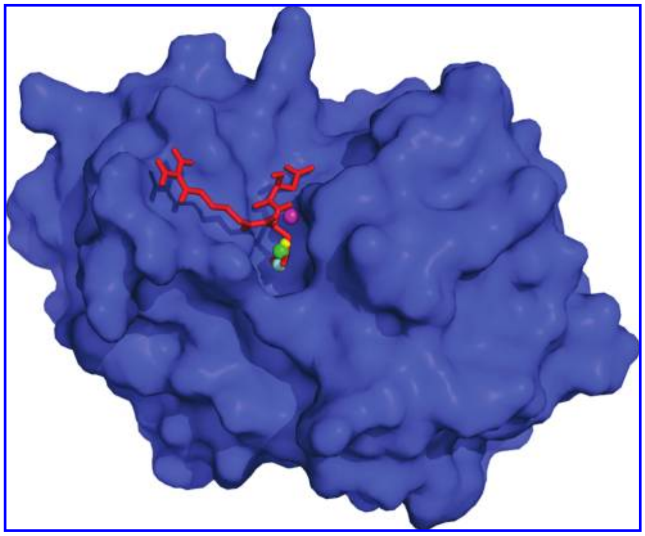

FIG. 3. The ligand (in red) binding site and identified pockets on the surface of the protein structure 1aec. The sites predicted by LIGSITE^{cs} (cyan sphere), PASS (yellow), Q-SiteFinder (green), and SURFNET (magenta) are all in the top one prediction. These four top one sites are spatially similar and identify the same ligand binding site.

Table 2. Success Rate (%) of the Top Three Predictions by Different Methods for 210 Bound Structures

Method	*Top 1*	*Top 2*	*Top 3*
MetaPocket	75	88	93
LIGSITE^{cs}	70	80	86
PASS	51	71	80
Q-SiteFinder	70	85	90
SURFNET	42	52	57

Table 4. Number of Proteins with Different Cluster Sizes [Number of Pocket Site (ps)] in MetaPocket for the Top Three Predictions in the Case of 210 Bound Structures

	First prediction	*Second prediction*	*Third prediction*
4 ps	119	5	1
3 ps	31	11	2
2 ps	6	10	7
1 ps	2	0	1

as a single probe in the center of the pocket. One way to decide whether the identified pocket site is the real ligand-binding site is to check whether it is within 4 Å of any atom of the ligand. If there are multiple ligands bound to the proteins, the best hit is picked up. This is how we evaluated the prediction method in our previous work (Huang and Schroeder, 2006), and here we just simply adapt it for this work.

Results

Table 1 shows the success rates using these five methods on the 48 bound/unbound structures. For unbound structures, metaPocket achieves the best overall success rate for all the top three predictions. Among the four single methods, LIGSITE^{cs} is the best, and can identify ligand-binding sites at 71 and 85% accuracy for the top one and top three pocket sites, respectively. By taking all the top three sites from these single methods into account, metaPocket improves the success rate from 71 to 75%, 85 to 90% (43 cases), for the top one and top three predictions, respectively. Among the five cases where metaPocket fails to succeed, there are three cases (5cpa, 3app, and 6ins), where none of the four single methods can identify the real ligand binding sites correctly within the top three predictions. In the rest two cases (1a6u and 2tga), only Q-SiteFinder predicts correctly the binding site for 1a6u at top 1. LIGSITE^{cs} can also identify the same pocket site but the ranking is top five (see Fig. 2). The reason is that the bound ligand NIP is rather small and does not bind to the largest pocket site. Q-SiteFinder succeeds in this case because it uses the probe energy as ranking schema rather than the size of the pocket. In 2tga, the loops near the binding site stretch significantly to allow ligand binding. LIGSITE^{cs} predicts the site at top three, but the other three methods fail. However, this ligand binding site is the biggest pocket on the bound structure 1mtw.

Table 2 shows the success rates of the five methods for the dataset of 210 bound structures. Overall, metaPocket achieves a slightly better success rate: a three percentage improvement for the top two and three; a five percentage improvement for the top one prediction. In this larger dataset, LIGSITE^{cs} and Q-SiteFinder both get a 70% success rate for the top one prediction. Our method metaPocket improves it to 75%. The success rate is comparable to that of LIGSITE^{csc} method in which the top three predictions were reranked using the degree of conservation score of the residues around the pocket site (Huang and Schroeder, 2006). One can see that the success rates present in this work are slightly different from those in our previous work (Huang and Schroeder, 2006). The reason for this small difference is the different parameters used here. The predicted pocket sites are classified into four classes: the actual ligand binding site, the second and third pocket, or none of these. Table 3 shows the number of proteins in these four classes for all the methods. In the top three pocket sites identified by metaPocket, there are 158 cases (75%) where the ligand binds into the first pocket site, 26 and 11 cases that the ligands choose the second or third pocket site as their binding site, respectively.

Discussion

As described above, the top three pocket sites identified by the four single methods are retained to be further clustered by metaPocket. Thus, we get a total of 12 pocket sites of which each is represented as a single probe. During the clustering procedure, only different pocket sites from different methods can be clustered into the same cluster. Therefore, each cluster contains one to four pocket sites (ps). Table 4 shows the number of proteins with different cluster sizes (1–4 ps) within the top three predictions of the metaPocket approach for 210 bound structures. As shown in the table, among the 158 cases that the ligand binds to the first pocket site, there are 119 cases (75%) where all the four single methods detect this ligand binding site correctly within their top three predictions, and 31 cases (20%) where three single methods predict the site correctly. Figure 3 shows one case (protein structure 1ace) where all the top one pocket sites predicted by these four methods are spatially similar and identify the same ligand binding site. Figure 4 shows the number of proteins with different number of clusters (number of pocket sites in metaPocket) for 210 bound structures after clustering the 12 pocket sites from the four different methods. In most of the cases, the

Table 3. Number of Proteins in Each Pocket Prediction Class For 210 Bound Structures

Class	*metaPocket*	*LIGSITE^{cs}*	*PASS*	*Q-SiteFinder*	*SURFNET*
C1: Binding site (bs) in the first pocket	158	146	108	152	88
C2: Bs in the second pocket	26	22	42	26	21
C3: Bs in the third pocket	11	12	17	10	11
C4: Bs in none of above	15	20	43	22	90

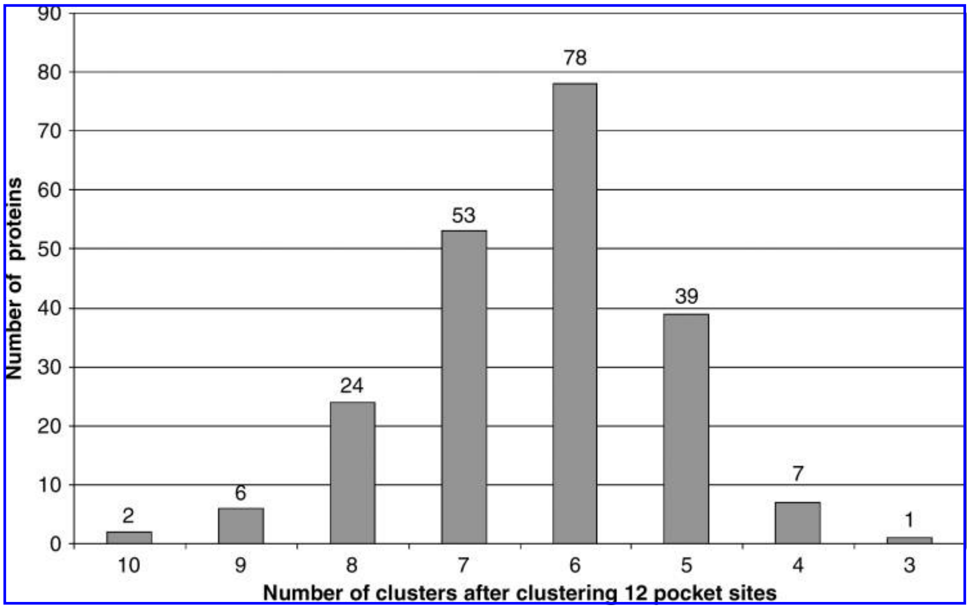

FIG. 4. Distribution of the number of proteins in different number of clusters after clustering the 12 pocket sites in the case of 210 bound structures. There are two cases (1ai5 and 2yhx) where the 12 pocket sites are clustered into 10 clusters, 78 cases for 6 clusters, and only 1 case (1ppi) for 3 clusters, that is, all the four methods identify the same three pocket sites in their top three predictions.

top three pocket sites from the four methods overlap somehow, that is, they form five to eight clusters from the 12 sites (probes). There are two cases where the 12 probes are clustered into 10 clusters. One case is the protein structure 2yhx (Fig. 5), in which LIGSITE^{cs}, PASS, and SURFNET identify the same pocket site in their top two prediction, and thus the three top two probes are clustered into the same cluster. However, the rest of the probes occupy different pocket sites. In this case, LIGSITE^{cs} and SURFNET detect the ligand-binding site correctly at their top one prediction. However, the distance between the top one probe from LIGSITE^{cs} and SURFNET is 8.4 Å and the distance threshold we use for hierarchal clustering is 8 Å.

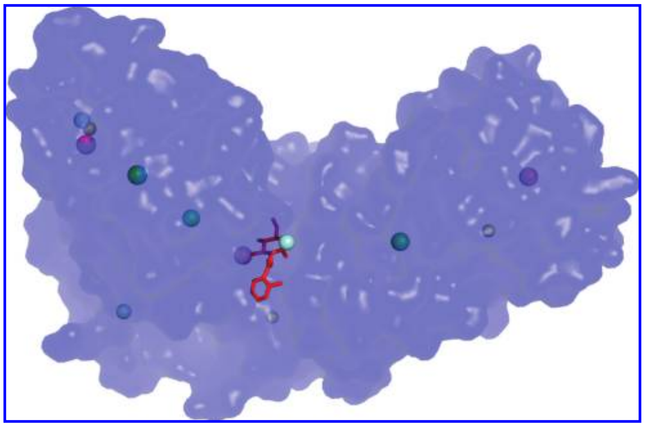

FIG. 5. The identified 12 pocket sites for protein 2yhx. These 12 pocket sites form 10 clusters. The three top two probes from LIGSITE^{cs} (cyan), PASS (yellow), and SURFNET (magenta) are in the same pocket site, as shown in the top left region. The rest of the nine probes occupy different pocket sites. In this case, LIGSITE^{cs} and SURFNET detect the ligand binding site correctly at their top one prediction.

Thus, if we increase the clustering distance threshold, metaPocket will identify this ligand-binding site at its top one prediction. We try different distance thresholds (5 to 10 Å) in the hierarchal clustering and 8 Å returns the best performance for metaPocket (data not shown). Furthermore, there is only one case (1ppi) where metaPocket has only three sites after clustering, that is, all the four methods identify the same three pocket sites in their top three predictions. As shown in Figure 6, all the four methods pick up the ligand-binding site at their top one, and thus, metaPocket at its top one as well. In the dataset of 210 bound proteins, there are 80 cases that have more than one ligand binding sites, for all of which

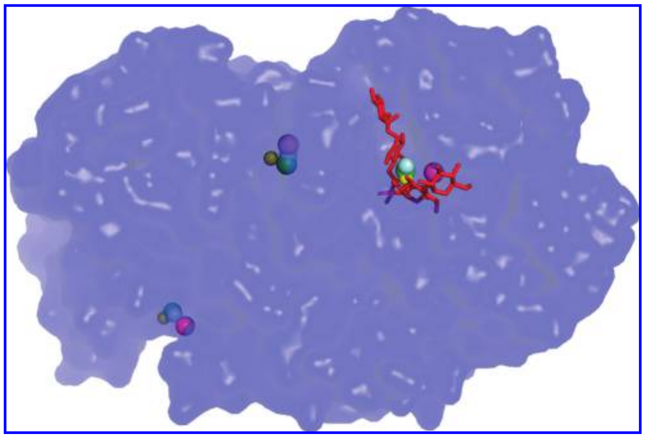

FIG. 6. The 12 identified pocket sites for protein 1ppi. These 12 probes occupy three pocket sites. All the four methods detect the ligand (red) binding site correctly at their top one prediction.

metaPocket can pick up at least one binding site correctly in the top three predictions. And there are 23 cases that metaPocket can identify more than one ligand binding sites in the top three predictions.

Conclusions

In recent decades, many computational efforts have been done to predict protein functional sites based on protein structures. These efforts include methods for prediction of protein–protein interaction sites and protein–ligand binding sites. A number of tools are available to identify pockets on protein surfaces and predict ligand-binding sites from the pockets. In this article, we propose a method called metaPocket, which combines the predictions done by LIGSITE^{cs}, PASS, Q-SiteFinder, and SURFNET. We compare metaPocket to the individual methods on a dataset of 48 unbound/bound and 210 bound-only protein–ligand complexes using the same evaluation criteria. The comparison results show that metaPocket performs slightly better than the other approaches and correctly predicts the ligand-binding site in 75% of the cases at top one, and 93% at top three predictions.

MetaPocket is online at http://metpocket.eml.org. Users can submit PDB files or enter a PDB ID and specify the chain ID. It returns the pocket sites identified by different methods in a standard PDB file format as well as a python script for visualizing the pockets using PyMol (Delano, 2002).

Acknowledgments

We thank the authors of PASS, Q-SiteFinder, and SURFNET for making their tools publicly available. We thank Rebecca Wade for discussions and Outi Salo-Ahen for reading the manuscript.

Author Disclosure Statement

The authors declare that no conflicting financial interests exist.

References

Binkowski, T., Naghibzadeh, S., and Liang, J. (2003). CASTp: computed atlas of surface topography of proteins, Nucleic Acids Res 31, 3352–3355.

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

Connolly, M. (1983). Analytical molecular surface calculation, J Appl Crystallogh, 16, 548–558.

Delano, W. (2002). The PyMOL Molecular Graphics System.

Edelsbrunner, H., Facello, M., Fu, P., and Liang, J. (1995). Measuring proteins and voids in proteins. Proc 28th Annu Hawaii Int Conf Syst Sci 5, 256–264.

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

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

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

Huang, B., and Schroeder, M. (2008). Using protein binding site prediction to improve protein docking, Gene 422, 14–21.

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

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

Levitt, D., and 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.

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

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

Puvanendrampillai, D., and Mitchell, J. (2003). Protein Ligand Database (PLD): additional understanding of the nature and specificity of protein–ligand complexes, Bioinformatics 19, 1856–1857.

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

Address correspondence to:
Dr. Bingding Huang
EML Research gGmbH
Schloss-Wolfsbrunnenweg 33
69118, Heidelberg, Germany

E-mail: bingding.huang@eml-r.villa-bosch.de