BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 15 2011, pages 2083–2088
doi:10.1093/bioinformatics/btr331

Structural bioinformatics

Advance Access publication June 2, 2011

Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction

Zengming Zhang[1], Yu Li[1], Biaoyang Lin[1], Michael Schroeder[2] and Bingding Huang[1,2,*]

[1]Systems Biology Division, Zhejiang-California International NanoSystems Institute, Zhejiang University, 310029 Hangzhou, China and [2]Bioinformatics Group, Biotechnology Center, Technical University of Dresden, 01307, Dresden, Germany

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Protein–ligand binding sites are the active sites on protein surface that perform protein functions. Thus, the identification of those binding sites is often the first step to study protein functions and structure-based drug design. There are many computational algorithms and tools developed in recent decades, such as LIGSITE$^{cs/c}$, PASS, Q-SiteFinder, SURFNET, and so on. In our previous work, MetaPocket, we have proved that it is possible to combine the results of many methods together to improve the prediction result.

Results: Here, we continue our previous work by adding four more methods Fpocket, GHECOM, ConCavity and POCASA to further improve the prediction success rate. The new method MetaPocket 2.0 and the individual approaches are all tested on two datasets of 48 unbound/bound and 210 bound structures as used before. The results show that the average success rate has been raised 5% at the top 1 prediction compared with previous work. Moreover, we construct a non-redundant dataset of drug–target complexes with known structure from DrugBank, DrugPort and PDB database and apply MetaPocket 2.0 to this dataset to predict drug binding sites. As a result, >74% drug binding sites on protein target are correctly identified at the top 3 prediction, and it is 12% better than the best individual approach.

Availability: The web service of MetaPocket 2.0 and all the test datasets are freely available at http://projects.biotec.tu-dresden.de/metapocket/ and http://sysbio.zju.edu.cn/metapocket.

Contact: bhuang@biotec.tu-dresden.de

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on March 14, 2011; revised on May 10, 2011; accepted on May 27, 2011

1 INTRODUCTION

Proteins perform their biological functions in biological processes mainly by interacting with other molecules such as other proteins, small molecules, DNAs and RNAs. Usually not all the residues on a protein surface participate in these interactions. Thus, identification of these functional sites is of great importance to understanding the function of a protein and the mechanism of the interactions. In addition, knowledge of these functional sites can be used to guide the mutagenesis experiments. There exist a number of cavities or pockets on protein surface where small molecules bind. Therefore, identification of such cavities is often the starting point in protein–ligand binding site prediction for protein function annotation and structure-based drug design. Proper ligand binding site detection is a prerequisite for protein–ligand docking and high-throughput virtual screening to identify drug candidates in drug discovery processes.

Many computational algorithms and tools have been developed in last two decades to identify pocket for protein–ligand binding site prediction. Most of the existing methods can be classified into two types: geometry based and energy based. The geometry-based methods can be further classified into grid based, sphere based and α-shape based (Kawabata, 2010; Yu *et al.*, 2010). In the grid-based methods, the protein structure is projected into a 3D grid and the grid points are categorized into different types according to their positions related to the protein. Then the solvent grid points are clustered using some geometry attributes and those grid points near the pocket sites can be recognized. LIGSITE (Hendlich *et al.*, 1997), LIGSITE^{CS} (Huang and Schroeder, 2006), PocketPicker (Weisel *et al.*, 2007), GHECOM (Kawabata, 2010) and ConCavity (Capra *et al.*, 2009) are the representatives of this type of method. In the sphere-based approaches, the common strategy is to fulfill protein surface with spheres of different radius layer by layer and a cutting method is applied during the fulfilling process. The final pocket sites are those regions that are rich with fulfilled spheres. This kind of methods include SURFNET (Laskowski, 1995), PASS (Brady and Stouten, 2000), PHECOM (Kawabata and Go, 2007) and POCASA (Yu *et al.*, 2010). Approaches based on α-shape theory (Edelsbrunner and Mucke, 1994) include CAST (Binkowski *et al.*, 2003; Dundas *et al.*, 2006) and Fpocket (Le Guilloux *et al.*, 2009). CAST computes the triangulations of the protein's surface atoms and these triangulations are grouped by letting small-sized ones flow toward the neighboring larger one. The pocket sites are the collection of empty triangles. Different from CAST, Fpocket uses the idea of α-sphere which is a sphere contacting four atoms on its boundary and containing no inside atom. The next step is to identify clusters of spheres close together and those clusters are potential pocket sites. In comparison to geometry-based method, Q-SiteFinder (Laurie and Jackson, 2005) aims to find pocket sites by computing the interaction energy between protein atoms and a small molecule probe. In Q-SiteFinder, layers of methyl (–CH3) probes are initialized on protein surface to calculate the van der Waals interaction energy between the protein atoms and the probes. Then the probes are clustered into many groups and are ranked by the total energy of probes. Those clusters with high energy will be the potential ligand binding sites. SiteHound (Ghersi and Sanchez, 2009; Hernandez

*To whom correspondence should be addressed.

et al., 2009) is similar to Q-SiteFinder but it includes Lennard-Jones and electrostatics energy terms and uses different types of probes to calculate interaction energy. However, it is difficult to compare their performance systematically because of different evaluation criteria and dataset being used. In our previous work (Huang and Schroeder, 2006), we compared LIGSITE^{cs}, SUFNET, PASS and Q-SiteFinder using the same dataset and criteria. Later on, we combined these four methods and introduced a new consensus tool called MetaPocket to improve the prediction success rate (Huang, 2009). Because there are many new tools developed recently, we continued our work on MetaPocket by including four more free available tools: Fpocket, GHECOM, ConCavity and POCASA. These tools were chosen because they are freely available either with source code or executable binary. In this work, we improve the workflow and the way of mapping ligand-binding residues and propose a new dataset for drug–target complexes. The web server design architecture is also improved as we developed a new on-line visualization system. We named the new version MetaPocket 2.0 (MPK2), in contrast to the old version of MetaPocket 1.0 (MPK1).

We demonstrated that MPK2 performed better than MPK1 and each of the individual methods by extensive validation and comparison. First, we applied MPK2 to the original three datasets of 48 bound/unbound and 210 bound complexes as we used before in our previous work (Huang, 2009; Huang and Schroeder, 2006). We proved that MPK2 improved the success rate up to 6% than MPK1. Second, we built a novel dataset of drug–target complexes and applied both MPK1 and MPK2 to this new dataset. MPK2 also showed better performance than its previous version with an improvement of up to 6% for the success prediction rate. Furthermore, we compared MPK2 to each single method and showed that MPK2 achieved >12% success rate over the best single method.

2 METHODS

2.1 MetaPocket algorithm

This section describes the algorithm and workflow of MPK2 for predicting ligand binding sites and mapping binding residues from protein 3D structures, as well as the design and architecture of the web server of MPK2. As mentioned above, MPK2 is a consensus method in which the predicted pocket sites from eight methods, LIGSITE^{CS}, PASS, Q-SiteFinder, SURFNET, Fpocket, GHECOM, ConCavity and POCASA, are combined together to improve the prediction success rate. There are three steps in MetaPocket 2.0 procedure: calling-based methods, generating meta-pocket sites and mapping ligand-binding residues. The whole working procedure of MPK2 is illustrated in Figure 1 and is described in details below.

Calling-based methods: in this step, the given protein structure is sent to all the based methods parallel and separately. For LIGSITE^{CS}, PASS, SURFNET, GHECOM, Fpocket and ConCavity, their executable binary programs are run locally to do the prediction. For Q-SiteFinder and POCASA, python scripts are implemented to submit the protein structure to their web servers and the results are retrieved from the remote servers automatically. As results, LIGSITE^{CS}, PASS and SURFNET output different clusters of grid points and the mass center of these clusters is used to represent the pocket site. For the other five methods, pocket sites are indicated by clustered probes. Thus, the mass center of each cluster is calculated and then is used as the representative point of the identified pocket sites. As we note that, each identified pocket site from every method is ranked by different scoring functions. To make them comparable, the z-score is calculated separately for each site in different methods, as used in our previous work (Huang, 2009).

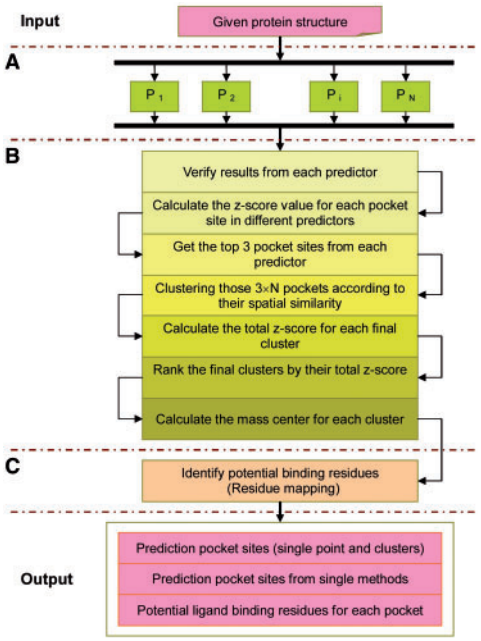

Fig. 1. The illustration of the MetaPocket 2.0 procedure. MPK2 takes the standard PDB file as input and output the prediction meta-pocket sites and also the prediction pocket sites from all the successfully running based single methods. The ligand binding residues for each meta-pocket are also listed. (Step A) Based methods execution. The given protein structure will be sent to all the based methods to do prediction. P1, P2, Pi and PN indicate the based methods (predictors). All the predictors are called in parallel to save running time. (Step B) Meta-pockets generation. This step includes z-score calculating, clustering pocket sites and final clusters ranking. (Step C) Residue mapping: identification of the potential ligand binding residues for each meta-pocket.

Generating meta-pocket sites: after calling each method, MPK2 only take the first three pocket sites from each method into account. Thus, totally there are 24 pocket sites and these pocket sites are somehow overlapped spatially. To identify those overlapped pocket sites, we use hierarchical clustering approach to cluster these 24 sites according to their spatial similarity. The distance cut-off threshold is set to 8 Å here. Then the total z-score for each cluster is calculated and serve as the final scoring function to re-rank the final meta-pocket sites. In the end, the mass center for each final cluster is calculated and is represented as the final meta-pocket site in MPK2.

Mapping ligand-binding residues around the meta-pocket site: the purpose of this step is to identify the functional residues around the identified meta-pocket site which could be the potential ligand binding sites on protein surface. As illustrated in Figure 2, MPK2 uses a synthetical way to identify those residues which might contribute to protein–ligand interaction. As we mentioned above, each method outputs a cluster of probe points for each pocket site. In this step, MPK2 merges the probe points from each single method in the same meta-pocket site. Then a big cluster of probe points is obtained for each meta-pocket site. Those surface residues, which are within a certain distance (5 Å used here) to the probe points in the cluster, are the potential ligand-binding residue. The surface residues are defined using the NACCESS program whose relative solvent accessible surface area is >20%.

2.2 Test datasets

Four different datasets are used in this work. The first three datasets are 48 bound/unbound and 210 bound datasets, which were first introduced in our previous work (Huang and Schroeder, 2006). To compare MPK2 to the other

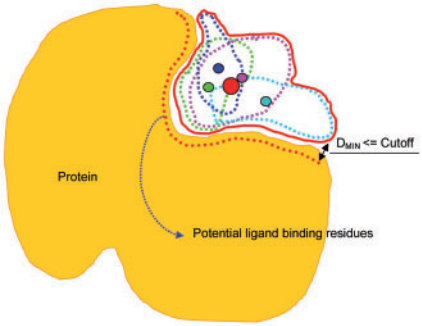

Fig. 2. The ligand-binding residues mapping procedure in MPK2. The smaller spheres are the pocket sites generated by different single methods. The bigger sphere is the meta-pocket site generated by MPK2. The regions surrounded by thin dotted lines out of protein are the original clusters of corresponding pocket sites generated by the corresponding single methods. The region surrounded by the thicker solid line is the cluster for the meta-pocket generated by MPK2 after merging all the clusters of single methods. The dotted line in the protein indicates the potential ligand-binding residues around the meta-pocket site, calculated by a distance threshold D_{MIN}.

methods and previous version of MetaPocket (MPK1), we still use these three datasets. In order to identify drug binding sites, we built a novel dataset of drug–target complex structures available in PDB. To our knowledge, the DrugPort database (http://www.ebi.ac.uk/thornton-srv/databases/drugport/) contains the information of protein–ligand complexes where the bound ligands are approved drugs reported in DrugBank (Wishart *et al.*, 2006, 2008). In the first step, we derive all drug–target pairs from DrugPort web site. For each pair, we retrieve the UniProt ID for the target and link it to PDB and get the PDB file to check whether it contains both protein target and drug ligand. Only one complex structure is selected for each drug–target pair and we only keep the single chain where ligands bind. At the end of this step, we obtained 217 pairs and 96 types of drugs. In the next step, we used CD-HIT (Huang *et al.*, 2010; Li and Godzik, 2006) program to remove the redundancy of protein targets using 40% similarity threshold. Finally 198 drug–target complexes are obtained. This dataset is freely available from the web site of MPK2.

2.3 Evaluation criteria

To evaluate and compare MPK2 with MPK1 and other individual-based methods fairly, the same performance measurement should be used. It is noted that for some proteins in the datasets, more than one ligand is bound. These ligands might be separated in different pocket sites but sometimes occupy the same region on protein surface, for example, those co-factors and substrates. First, we define the real ligand binding sites (RBSs), which are those regions on protein surface where one or more ligands are bound. If two ligands are closed to each other (distance threshold 5 Å), they are defined to share the same RBS. Here, we define that one RBS is predicted correctly if it is located at the identified pocket sites, i.e. any atom of the ligand is within 4 Å to the mass center of this pocket, as we used in our previous work (Huang and Schroeder, 2006). We also define that a prediction is a hit if at least one RBS in the given protein is detected correctly in a certain number of top predictions. The top 1 to top 3 identified pocket sites from MPK2 and other methods are evaluated separately in this work. Thus, to compare the performance of different approaches quantitatively, the success rate (SR) is calculated according to the following formulas:

$$\text{Success\_Rate} = \frac{N_{\text{HIT}}}{N_{\text{P}}}$$

Where N_{P} is the total number of proteins in the dataset; N_{HIT} is the total number of hit prediction. The success rate is calculated for all the methods for the top 1, top 2 and top 3 predictions, respectively.

Table 1. The comparison of MPK2 to MPK1 on success rate (%) for different datasets

Dataset	Version	Top 1	Top 2	Top 3
48 (bound)	MPK2	85	92	96
	MPK1	83	94	96
48 (unbound)	MPK2	80	90	94
	MPK1	75	85	90
210 (bound)	MPK2	81	91	95
	MPK1	75	89	94
198 drug–target	MPK2	61	70	74
	MPK1	55	65	68

3 RESULTS

3.1 MPK2 improves the prediction success rate by combining eight individual prediction methods

In our previous work, only four methods are included in MPK1: LIGSITE^{cs}, SUFNET, PASS and Q-SiteFinder (Huang, 2009). Recently, there are four more free available tools: Fpocket, GHECOM, ConCavity and POCASA, as described above. We therefore developed a MetaPocket 2.0 (MPK2) to combine these eight methods of detection. We evaluated MPK2 and MPK1 on the three old datasets used before (Huang, 2009) and the dataset of 198 drug–target complexes which we developed in this work, and compared the success rates of MPK2 and MPK1. Table 1 shows the detailed comparison results. In the first three old datasets, MPK2 improved the success by up to 6% at the top 1 prediction in 210 bound and in 48 unbound dataset. For the novel dataset of 198 drug–target complexes, the improvement of MPK2 over MPK1 is significant, ranking from 4% to 6% for all the top 3 predictions. Overall, after including four new methods, MPK2 improves the whole performance of prediction.

3.2 MPK2 outperforms all the single methods

Table 2 shows the success rates for MPK2 and the eight single methods for the drug–target dataset. Overall, MPK2 archived better result than each of the eight single methods. In the top 1 and top 2 prediction, LIGSITE^{CS} performed best among the eight single methods and MPK2 increased the success rate by 13%. In the top 3 predictions, Q-SiteFinder is the best method and MPK2 also receives 12% improvements. The reason why MPK2 improves the success rate is that it takes the overlapping prediction results from different approaches. One pocket site has higher probability to be a RBS if it was picked out by multiple methods as top predictions. This is not surprising as different pocket detection methods use different scoring functions to rank these cavities and MPK2 clusters all the identified pocket sites according to their spatial distance and re-ranks them by summing up the z-scores of different methods.

3.3 How many cavities occur on protein surface?

In the combining procedure of MetaPocket 2.0, only the top 3 pocket sites from each of 8 single methods are taken into account, and these 24 pocket sites are clustered into different clusters (so called meta-pocket site) according to their spatial similarity. In the evaluation of MPK2 on the drug–target dataset, the number of final clusters

Table 2. The success rates (%) of the top 3 predictions by MPK2 and eight different methods on the drug–target dataset

Method	Top 1	Top 2	Top 3
MPK2	**61**	**70**	**74**
LIGSITE^{CS}	***48***	***57***	61
PASS	35	50	56
Q-SiteFinder	40	54	***62***
SURFNET	24	30	34
GHECOM	39	51	56
ConCavity	47	53	56
Fpocket	31	48	57
POCASA	43	54	56

The values in bold and italic indicate they are the best values.

Table 3. Number of hit proteins in each pocket prediction class on the drug–target dataset

Method	First pocket	Second pocket	Third pocket	None
MPK2	121	17	9	51
LIGSITE^{CS}	95	18	7	78
PASS	69	30	11	88
Q-SiteFinder	79	28	16	75
SURFNET	46	11	8	133
GHECOM	78	22	10	88
ConCavity	93	12	6	87
Fpocket	61	34	17	86
POCASA	83	23	4	88

Fig. 3. The MetaPocket 2.0 prediction success rates at the top 3 versus the number of clusters (meta-pocket sites). The number of proteins is also indicated.

for each protein and the prediction success rates of MPK2 on those proteins are quite diverse. Figure 3 shows the distribution of the number of proteins with different number of clusters on the drug–target dataset, and the success rates for those proteins having the same number of clusters. Overall, the number of clusters ranges from 4 to 14, which means there are 4 to 14 cavities (meta-pocket sites) on protein surfaces generally. There are 5 cases in which those 24 pockets are clustered into 4 clusters, meaning that those 5 proteins only have 4 big cavities on their surfaces and all the 8 methods correctly picked them up at their top 3 predictions. In these five cases, MPK2 all predicted the ligand binding sites correctly. There is only one case that the number of final clusters is 14, which indicates that this protein has 14 cavities on its surface and each of 8 methods picked up different pockets at their top 3 predictions. The real ligand binds to one of those 14 cavities and MPK2 failed to recognize it correctly at the top 3 predictions. As shown in Figure 3, most of the proteins have 7 (43 cases) or 8 (56 cases) cavities on surface generally and there is no correlation between the number of cavities and the prediction success rate of MPK2.

3.4 Most of ligands bind to large pockets

In order to check whether ligands bind to large pockets on protein surface, we conducted a statistical analysis to assess the possibility that a RBS locates at the top 3 prediction pockets. The identified pocket sites are classified into four different classes: the actual ligand binding site locates at the first, the second, the third pocket or at none of these top 3 pockets (Table 3). In the top 3 predictions of MPK2,

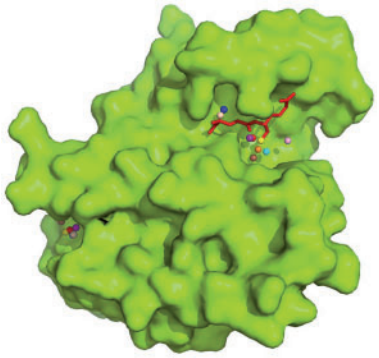

Fig. 4. The real ligand (red) binding site and the identified pockets on glutathione *S*-transferase (PDB code: 1PX7). The pocket sites of LIGSITE^{CS} (purple), PASS (cyan), SURFNET (brown), Q-SiteFinder (blue), Fpocket (pink), ConCavity (orange), GHECOM (yellow) and POCASA (wheat) are all from their top 1 predictions and are located in the same cavity where ligand binds. The meta-Pocket site from MPK2 is shown in red sphere.

there were 121 (61%) cases that the top 1 predicted pocket is the RBS. There were 17 and 9 cases that the RBS was located at the top 2 and top 3 predicted pocket, respectively. However, there were 51 cases for which the MPK2 failed to detect the RBS among the top 3 predictions. Among the 121 cases that ligands were predicted to bind to the first pocket site in MPK2, in 94 (78%) cases, the predictions overlap with one of the top 3 identified pockets identified by all the 8 single methods and in 17 (14%) cases the predictions overlap with one of the top 3 identified pockets identified by 7 out of the 8 single methods. Only in 12 of the 121 cases, the real-ligand binding sites were predicted by all 8 single methods as the top 1 prediction. Figure 4 shows a representative case for such situation for Glutathione S-transferase (PDB code: 1PX7).

3.5 Dealing with difficult cases for which ligand binding does not occur in the large cavities

Although MPK2 significantly outperforms its previous version and each of the individual methods, it could not correctly detect those binding sites where the ligands do not occur in the large cavities on protein surface. We investigate all the 51 cases for which MPK2 fails to detect the RBSs within its top 3 predictions and categorized them into four classes according to the following reason: flat RBS; RBS too small to be detected; RBS at the interface of two domains;

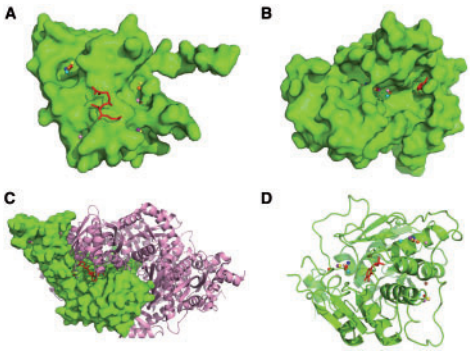

Fig. 5. Examples of difficult structures in drug–target dataset. For each structure, protein is illustrated in green surface or cartoon; ligands are illustrated in red stick; identified pocket sites are illustrated in small spheres. (**A**) The flat binding site (triggering receptor expressed on myeloid cells, PDB code: 1q8m_A). The ligand binds to the flat region on protein surface, not the expected pocket shape region. (**B**) The RBS is too small to be detected in the first three predictions. (Oxidoreductase, PDB code: 1yxm_B). (**C**) Ligand binds at the interface of two chains or domains (HMG-CoA Reductase, PDB code: 1hwk_A, the other chain is also shown in magenta). (**D**) The RBS is inside the protein and thus cannot be detected (cystathionine beta-synthase, PDB code: 1m54_A).

and RBS inside the protein. We show a representative case for each class in Figure 5. In the first class, ligands bind to a flat region on protein surface. Therefore, geometry approaches that identify pockets cannot detect such binding site correctly. Of total, 26 out of 51 proteins belong to this class. In the second class, many cavities on protein surface are all likely to be ligand binding sites but the X-ray structures show that the ligands bind to small pockets rather than to big pockets. Thus, the RBSs were not predicted among the top 3 identified pockets (10 cases). For the third class, two proteins (chains or domains) form a complex and the binding pockets are located at the interface between them. But these pockets do not exist when the two proteins are separated from each other. Because we used the single protein for prediction, MPK2 could not detect such pockets. There are nine such cases in the drug–target dataset. However, when the whole complex structures for such cases were used in MPK2 prediction, the RBSs were correctly recognized for 8 out of 9 cases except PDB code: 1F3A. In the complex structure of 1F3A, there is a big pocket-shape region in the interface between two proteins and MPK2 successfully detected this pocket. The ligand was predicted to bind at the edge of the pocket but not inside the pocket, as shown by X-ray structure. Therefore, MPK2 failed to recognize the RBSs correctly in this case. In the fourth class, the RBSs are inside proteins as shown in the X-ray structure and MPK2 cannot handle this case since it only pick up the pockets on protein surface (6 cases).

4 DISCUSSION

Although many computational approaches have been developed to identify pocket for ligand binding sites prediction, there are a few methods that predict protein druggability (Cheng *et al.*, 2007; Hajduk *et al.*, 2005a; Schmidtke and Barril, 2010; Sugaya and Ikeda, 2009). How to discriminate druggable cavities from non-druggable ones is still a challenge problem (Hajduk *et al.*, 2005b). Nayal and Honig used the program SCREEN (Nayal and Honig, 2006) to locate and analyze the surface cavities of a non-redundant set of 99 proteins co-crystallized with drugs and they found that using cavity size alone as a criterion predicted drug binding sites with 72% coverage. With aid of Random Forests and 408 physicochemical, structural and geometric features, the prediction coverage was improved to 89% (Nayal and Honig, 2006). In another recent work, different pocket descriptors including pocket volume/size, solvent accessible surface area, hydrophobicity score, etc., have been integrated as a drug score in the Fpocket program package to score the druggability of cavities (Schmidtke and Barril, 2010). As shown in Table 2, MetaPocket 2.0 can detect about 74% of the drug binding sites at the top 3 predictions using a simple scoring function (*Z*-Score). In order to gain better druggability prediction accuracy, we are planning to develop new druggability prediction method which will consider many physical–chemical and structural/sequence features. This is beyond the scope of this work and hence is not described here. Nevertheless, we proposed a dataset of drug–target complexes with available structures in this work, which can be further used to evaluate new structure-based drugability prediction methods.

To make our tool available to the community, we developed a new web server for MPK2 with better design and software architecture. In the new web server, eight single methods are called in parallel to reduce computational time. Each of eight single methods is treated as a plug-in in MPK2 and thus it is easy to add other new predictors when available. With this design pattern, the new web server is much more extensible than its previous version. It is important to mention that some of the eight methods might fail to return any prediction results for some reasons. This plug-in pattern makes our server automatically detect the failed methods and the algorithm is only applied to those results from successful methods. This feature makes MPK2 server more robust than MPK1. The users can provide a PDB ID and a chain ID or upload their own structures. The server will output the prediction results from eight single methods and the meta-pocket sites of MPK2 based on those results. The predicted pocket sites and those surrounding residues can be downloaded as standard PDB files or directly be visualized in the server based on JMOL (http://www.jmol.org) plug-in. It only takes about 10 s to 0.5 min to finish pocket identification depending on the size of protein. We envisage that our web server will become an all-in-one tool for protein–ligand binding site prediction to the community and provide useful guide to structure-based functional annotation, site-directed mutagenesis experiments, protein–ligand docking and large-scale virtual screening.

ACKNOWLEDGEMENTS

We thank all the authors who developed the eight single methods and made their tools available. We thank JingNa Si and Wenhan Wang for discussion and useful suggestion.

Funding: Ministry of Science and Technology (MOST) China international cooperation projects (grant no: 2008DFA11320); EU 7th Framework Marie Curie Actions of International Research Staff Exchange Scheme (IRSES) project (grant no: 247097).

Conflict of Interest: none declared.

REFERENCES

Binkowski,T.A. *et al.* (2003) CASTp: Computed Atlas of Surface Topography of proteins. *Nucleic Acids Res.,* **31**, 3352–3355.

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

Capra,J.A. *et al.* (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput. Biol.,* **5**, e1000585.

Cheng,A.C. *et al.* (2007) Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.,* **25**, 71–75.

Dundas,J. *et al.* (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. and Mucke,E. (1994) Three-dimensional alpha shapes. *ACM Trans. Graph.,* **13**, 43–72.

Ghersi,D. and Sanchez,R. (2009) EasyMIFS and SiteHound: a toolkit for the identification of ligand-binding sites in protein structures. *Bioinformatics*, **25**, 3185–3186.

Hajduk,P.J. *et al.* (2005a) Druggability indices for protein targets derived from NMR-based screening data. *J. Med. Chem.,* **48**, 2518–2525.

Hajduk,P.J. *et al.* (2005b) Predicting protein druggability. *Drug Discov. Today*, **10**, 1675–1682.

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

Hernandez,M. *et al.* (2009) SITEHOUND-web: a server for ligand binding site identification in protein structures. *Nucleic Acids Res.,* **37**, W413–W416.

Huang,B. (2009) MetaPocket: a meta approach to improve protein ligand binding site prediction. *OMICS,* **13**, 325–330.

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,Y. *et al.* (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics,* **26**, 680–682.

Kawabata,T. (2010) Detection of multiscale pockets on protein surfaces using mathematical morphology. *Proteins,* **78**, 1195–1211.

Kawabata,T. and 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.

Laskowski,R.A. (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.,* **13**, 323–330, 307–308.

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

Le Guilloux,V. *et al.* (2009) Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics,* **10**, 168.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics,* **22**, 1658–1659.

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

Schmidtke,P. and Barril,X. (2010) Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *J. Med. Chem.,* **53**, 5858–5867.

Sugaya,N. and Ikeda,K. (2009) Assessing the druggability of protein-protein interactions by a supervised machine-learning method. *BMC Bioinformatics,* **10**, 263.

Weisel,M. *et al.* (2007) PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chem. Cent. J.,* **1**, 7.

Wishart,D.S. *et al.* (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.,* **34**, D668–D672.

Wishart,D.S. *et al.* (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.,* **36**, D901–D906.

Yu,J. *et al.* (2010) Roll: a new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. *Bioinformatics,* **26**, 46–52.