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, 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 their positions related to the protein. Then the solvent grid points are clustered into many groups and are ranked by the total energy 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 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 a-shape based (Kawahata, 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 region can be recognized. LIGSITE (Hendlich et al., 1997), LIGSITE (Huang and Schroeder, 2006), PocketPicker (Weisel et al., 2007), GHECOM (Kawahata, 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 the protein surface with spheres of different radius layer by layer and then the solvent grid points are clustered into many groups and are ranked by the total energy 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 such cavities is often the starting point in protein–ligand binding site prediction for protein function annotation and structure-based drug design.

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 the probes. Those clusters with high energy will be the potential ligand binding sites. SiteHound (Ghersi and Sanchez, 2009; Hernandez...
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, LIGSITEcs, 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 LIGSITEcs, 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, LIGSITEcs, 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 points. Thus, the mass center of each cluster is calculated and then 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).

Generating meta-pocket sites: after calling each method, MPK2 only takes 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 syntheatical 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...
Where \( N \) These ligands might be separated in different pocket sites but sometimes
the dotted line in the protein indicates the potential ligand-binding residues
The region surrounded by the thicker solid line is the cluster for the meta-
the bigger sphere is the meta-pocket site generated by MPK2. The regions
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
targets. 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
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
summed up the z-scores of different methods.

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Version</th>
<th>Top 1</th>
<th>Top 2</th>
<th>Top 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>48 (bound)</td>
<td>MPK2</td>
<td>85</td>
<td>92</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>MPK1</td>
<td>83</td>
<td>94</td>
<td>96</td>
</tr>
<tr>
<td>48 (unbound)</td>
<td>MPK2</td>
<td>80</td>
<td>90</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>MPK1</td>
<td>75</td>
<td>85</td>
<td>90</td>
</tr>
<tr>
<td>210 (bound)</td>
<td>MPK2</td>
<td>81</td>
<td>91</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>MPK1</td>
<td>75</td>
<td>89</td>
<td>94</td>
</tr>
<tr>
<td>198 drug-target</td>
<td>MPK2</td>
<td>61</td>
<td>70</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>MPK1</td>
<td>55</td>
<td>65</td>
<td>68</td>
</tr>
</tbody>
</table>

### 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 POCSA, 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, ranging 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

<table>
<thead>
<tr>
<th>Method</th>
<th>Top 1</th>
<th>Top 2</th>
<th>Top 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPK2</td>
<td>61</td>
<td>70</td>
<td>74</td>
</tr>
<tr>
<td>LIGSITE&lt;sub&gt;CS&lt;/sub&gt;</td>
<td>48</td>
<td>57</td>
<td>61</td>
</tr>
<tr>
<td>PASS</td>
<td>35</td>
<td>50</td>
<td>56</td>
</tr>
<tr>
<td>Q-SiteFinder</td>
<td>40</td>
<td>54</td>
<td>62</td>
</tr>
<tr>
<td>SURFNET</td>
<td>24</td>
<td>30</td>
<td>34</td>
</tr>
<tr>
<td>GHECOM</td>
<td>39</td>
<td>51</td>
<td>56</td>
</tr>
<tr>
<td>ConCavity</td>
<td>47</td>
<td>53</td>
<td>56</td>
</tr>
<tr>
<td>Fpocket</td>
<td>31</td>
<td>48</td>
<td>57</td>
</tr>
<tr>
<td>POCASA</td>
<td>43</td>
<td>54</td>
<td>56</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>First pocket</th>
<th>Second pocket</th>
<th>Third pocket</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPK2</td>
<td>121</td>
<td>17</td>
<td>9</td>
<td>51</td>
</tr>
<tr>
<td>LIGSITE&lt;sub&gt;CS&lt;/sub&gt;</td>
<td>95</td>
<td>18</td>
<td>7</td>
<td>78</td>
</tr>
<tr>
<td>PASS</td>
<td>69</td>
<td>30</td>
<td>11</td>
<td>88</td>
</tr>
<tr>
<td>Q-SiteFinder</td>
<td>79</td>
<td>28</td>
<td>16</td>
<td>75</td>
</tr>
<tr>
<td>SURFNET</td>
<td>46</td>
<td>11</td>
<td>8</td>
<td>133</td>
</tr>
<tr>
<td>GHECOM</td>
<td>78</td>
<td>22</td>
<td>10</td>
<td>88</td>
</tr>
<tr>
<td>ConCavity</td>
<td>93</td>
<td>12</td>
<td>6</td>
<td>87</td>
</tr>
<tr>
<td>Fpocket</td>
<td>61</td>
<td>34</td>
<td>17</td>
<td>86</td>
</tr>
<tr>
<td>POCASA</td>
<td>83</td>
<td>23</td>
<td>4</td>
<td>88</td>
</tr>
</tbody>
</table>

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.

Fig. 4. The real ligand (red) binding site and the identified pockets on glutathione S-transferase (PDB code: 1PX7). The pocket sites of LIGSITE<sub>CS</sub> (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;
Although many computational approaches have been developed to predict protein binding sites, there are a few methods that predict protein druggability (Cheng et al., 2007; Hajduk et al., 2005a; Schmidtle 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). Naylor and Honig used the program SCREEN (Naylor 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% (Naylor 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 (Schmidtle 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 druggability 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 these 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 the 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**

Z. Zhang et al.


