

Comprehensive Identification of “Druggable” Protein Ligand Binding Sites

Jianghong An¹ Maxim Totrov² Ruben Abagyan¹
jianghon@scripps.edu max@molsoft.com abagyan@scripps.edu

¹ Department of Molecular Biology, The Scripps Research Institute, La Jolla, CA 92037, USA
² Molsoft, LLC, La Jolla, CA 92037, USA

Abstract

We have developed a new computational algorithm for de novo identification of protein-ligand binding pockets and performed a large-scale validation of the algorithm on two systematically collected datasets from all crystallographic structures in the Protein Data Bank (PDB). This algorithm, called DrugSite, takes a three-dimensional protein structure as input and returns the location, volume and shape of the putative small molecule binding sites by using a physical potential and without any knowledge about a potential ligand molecule. We validated this method using 17,126 binding sites from complexes and apo-structures from the PDB. Out of 5,616 binding sites from protein-ligand complexes, 98.8% were identified by predicted pockets. In proteins having known binding sites, 80.9% were predicted by the largest predicted pocket and 92.7% by the first two. The average ratio of predicted contact area to the total surface area of the protein was 4.7% for the predicted pockets. In only 1.2% of the cases, no “pocket density” was found at the ligand location. Further, 98.6% of 11,510 binding sites collected from apo-structures were predicted. The algorithm is accurate and fast enough to predict protein-ligand binding sites of uncharacterized protein structures, suggest new allosteric druggable pockets, evaluate druggability of protein-protein interfaces and prioritize molecular targets by druggability. Furthermore, the known and the predicted binding pockets for the proteome of a particular organism can be clustered into a “pocketome”, that can be used for rapid evaluation of possible binding partners of a given chemical compound.

Keywords: protein-ligand binding sites, active sites, binding pockets, PDB

1 Introduction

Prediction of ligand-binding sites is a fundamental step in order to investigate the molecular recognition mechanism and function of a protein. Because an increasing number of protein structures are becoming available from high-throughput structural genomics projects prior to biological and functional characterization, computational methods to predict ligand-binding sites are becoming increasingly important. There are three independent sources of information that can be used to infer the location of possible ligand binding sites on the surface of a protein: (i) *protein structure*, (ii) *evolutionary information (sequence alignments)*, and (iii) *ligand/substrate information*. A number of sophisticated algorithms using evolutionary information or algorithms predicting locations of binding sites for specific substrates have been published [9, 20, 21]. Here we attempt to develop a highly accurate algorithm that is based solely on the protein structure and without any prior knowledge about the nature of the substrate. We postulate that the structure itself is informative enough and that the sequence and the ligand signal can be easily added to the “structural signal” when it is needed and available.

Proteins are involved in all kinds of molecular interactions: with other proteins, DNA, RNA, peptides, and small molecules. In this paper we present an algorithm not to identify every kind of inter-molecular interface, but primarily those that can be targeted with small “drug-like” compounds.

In other words, we wanted to elucidate where proteins can be modulated by chemical compounds that resemble small orally-available therapeutics. Once druggability is established, high throughput ligand docking or structure-based drug design [1, 3, 11, 14, 26, 31] can generate a list of initial drug candidates.

The “druggability” requirement excludes from consideration very small ligands, such as metals and small solvent molecules, as well as very large or long substrates. The properties of drug-like molecules are well studied [22, 29] and cover a wide range of sizes and physicochemical properties. Similarly, potential ligand binding pockets may cover an even larger range. For example, a neutral drug-like molecule can successfully compete with a charged substrate. Can the same pocket prediction algorithm account for these diverse pocket types?

A number of structure based pocket prediction algorithms have been published over the last ten years. They can be divided into two general classes: (i) geometric algorithms, and, (ii) probe mapping/docking algorithms. Geometric approaches analyze protein surfaces in order to find clefts. SURFNET [16] detects the gap regions in proteins by fitting spheres into the spaces between protein atoms, this results in a number of separate groups of interpenetrating spheres, which correspond to the protein’s cavities and clefts. LIGSITE [13], an improved version of POCKET [18], identifies clefts by putting the protein in a regular Cartesian grid and scanning along the x, y, and z axes and the cubic diagonals for areas that are enclosed on both sides by protein. APROPOS [24] and CAST [19] are based on the alpha-shape algorithm, they identify pockets by comparing surfaces of the protein generated with different levels of detail. PASS [8] identifies the “active site points” by coating the protein surface with a layer of spherical probes first and then filtering out those that clash with the protein or not sufficiently buried. The active site points are identified from the final probes. Besides those pure geometrical methods, more physically justified mapping procedures based on mapping/docking and scoring of molecular fragments have been published [10, 15, 25, 30]. There are also several docking based methods that used ligands to probe the proteins for binding sites [7, 12]. Two recent reviews of computational tools for identification of small-molecule binding sites in proteins have been published [9, 27].

Pure geometric methods are relatively straightforward but there is no direct physical meaning behind them. On the other hand, methods using molecular fragments mapping and ligand docking are more physically justified but generally too computationally expensive. It is also difficult to obtain an ideal discrimination score. A pitfall of the previously cited methods is that only small datasets were used for a test benchmark. While APROPOS used a relatively large test set of about 300 structures, others only used 10 to 50 selected test cases. This is despite the fact that thousands of X-ray structures have been deposited in the PDB [5, 6]. Another problem is that the goal of binding site prediction methods is to find active sites on uncharacterized structures, whereas most of the methods only tested complexes instead of performing a realistic test using uncomplexed (apo) structures. (PASS was tested on a small dataset of 21 apo-structures.) A benchmark test based on a large, systematic dataset of apo-structures is necessary for evaluating protein-ligand binding site identification methods.

In this paper, we present and validate a fast and accurate algorithm of ligand binding site prediction. The algorithm called DrugSite, is based on a transformation of the van der Waals energy potential. Like pure geometric approaches, DrugSite is very fast and capable of identifying clefts and cavities regardless of the nature of the substrate, while being more sensitive and specific. The method was tested on a systematically collected dataset which is two orders of magnitude larger than previous benchmarks: 5,616 binding sites collected from ligand-protein complexes, and 11,510 apo-binding sites inferred from the complexes by homology.

2 Method and Results

Protein-ligand binding sites are identified based on the grid potential map of van der Waals interaction of the receptor. Prior to computation, we remove all ligands and water molecules from the receptor.

Step1: Create the grid potential map of the van der Waals force field using a carbon probe of 1.7 Å radius in orthogonal parallelepiped surrounding receptor atoms [28]. A margin of 1Å was used.

$$P_i^0 = \sum_{j=1}^N \left(\frac{A_{jC}}{r_{ij}^{12}} - \frac{B_{jC}}{r_{ij}^6} \right),$$

where r_{ij} is the distance between the probe and the atoms. A_{jC} and B_{jC} are calculated according to the ECEPP/3 molecular mechanics force field. P_i^0 values were further truncated into the $[\text{minimum}(P_i^0), -0.8]$ range.

Step2: Smooth (space-average) the potential map 10 times to emphasize the regions with larger cumulative values and to avoid excessive density fragmentation.

Step3: Create potential ligand envelopes by contouring the resulting map at a level calculated as follows:

$$\text{Contouring Level} = \text{Mean}(\text{map}) - \text{Threshold} * \text{Rmsd}(\text{map}),$$

where $\text{Threshold} = 4.6$ was established on the training data set and RMSD is a root-mean-square difference of all map values.

Step4: Sort the created envelopes by their volumes and filter out those smaller than 100 Å³. All parameters for the map transformations have been optimized using a large, diverse set of binding sites. The algorithm was coded with the ICM scripting language [2, 23].

Since our main objective is to develop and validate an accurate algorithm to predict “druggable” pockets, namely protein ligand binding sites to which a typical drug-like small molecule may bind, we first studied the size distribution of known drugs. We then compiled a comprehensive database of appropriate protein ligand binding sites and finally tested the performance of the pocket prediction algorithm.

Compiling an exhaustive benchmark of ligand binding sites. We considered all 17,730 crystallographic protein structures from the October 3, 2003 PDB release compile a data set of observed binding sites from protein-ligand complexes (Liganded-Pocket Set, or LP-Set (Liganded Pocket Set collected from complexes)). We chose structures with better than 2.5Å resolution and with non-peptide ligands larger than 7 heavy atoms. The size limit excludes metals and popular crystallization buffer components. We also considered the crystallographic symmetry. Ligands interacting with the symmetric neighbors were removed because their binding sites are formed between the asymmetric units and building a correct model requires biological information. A detailed description of the procedure is given in the supporting information. The final set consisted of 5,616 protein-ligand binding sites representing 4,711 PDB entries with 2,175 unique ligands.

Collecting inferred protein-ligand binding sites from the uncomplexed structures. Since the main goal of our algorithm is to predict a potential binding envelope from an uncomplexed structure, it was critical to compile a benchmark of Unliganded Pocket sites (UP-Set (Unliganded Pocket Set collected from apo-structures)). This additional data set helped us to validate the pocket prediction algorithm in a more realistic situation. Unliganded pockets may not be as obvious as the liganded pockets due to the ligand induced conformational changes. Side-chains often obstruct a part of the pocket in the absence of the ligand. In order to collect the unliganded pockets, we inferred potential ligand binding sites from a close sequence similarity to a protein complex in the LP-set. We chose a safe limit of 95% sequence identity of the whole receptor chain and 100% identity in the 8Å vicinity of a potential binding site. The number of inferred binding sites of the same liganded pocket is determined by the number of close homologues of the liganded receptor. We preserved all alternative binding sites in the UP-set to avoid an arbitrary choice of one representative unliganded pocket. A total of 11,510 unliganded site projections were collected from the eligible PDB entries by superimposing the receptors. 1,445 binding sites from LP-set corresponded to these 11,510 mapped binding sites. As a result, this set represents on average about 7.9 different mappings to the same binding site of LP-set. A detailed description of the procedure is given in the supporting information.

Representation of the predicted ligand-binding pockets. We calculated transformed version of the three-dimensional van der Waals potential on a 1Å grid surrounding the entire protein surface (see Methods), then contoured at an optimized level to create a pocket envelope. The pocket envelope was represented by a triangulated surface (Fig. 1). We chose this potential because in contrast to purely geometrical methods it has a clear physical meaning, and, at the same time it does not require any knowledge of the chemical nature of a ligand. Additionally, the van der Waals component of the binding energy is present in complexes of various physical natures, including hydrophobic, charged, polar, or mixed complexes.

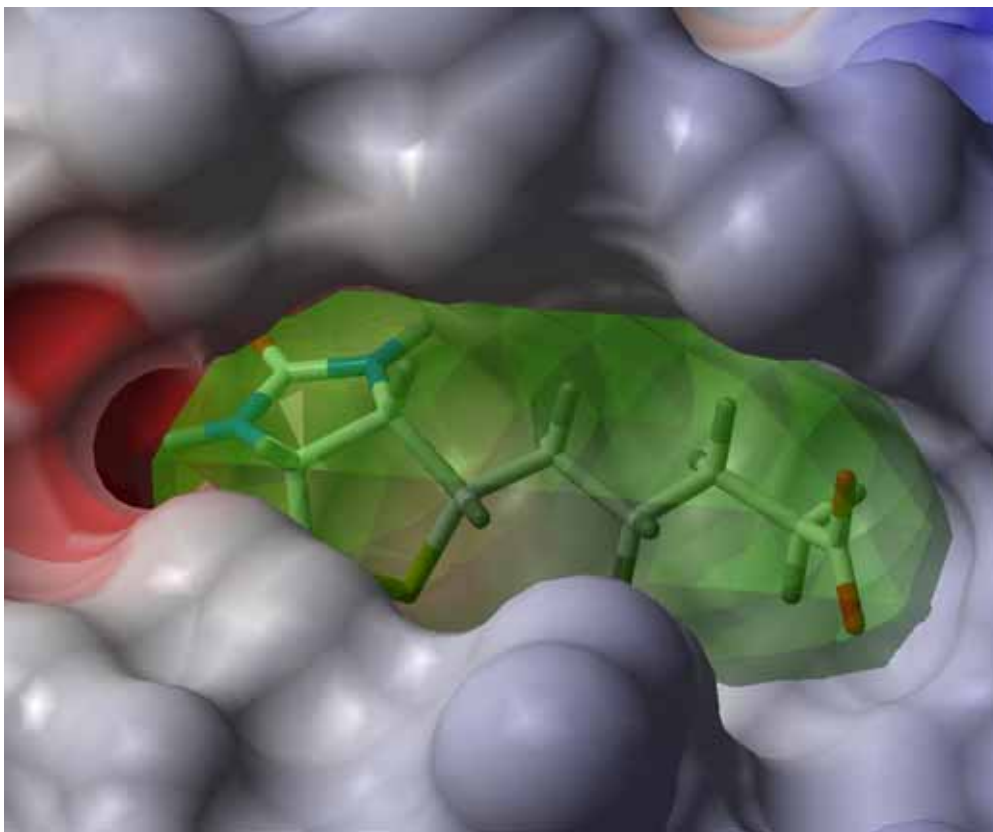


Figure 1: An example of identified biotin-streptavidin binding site (PDB 2izi). The predicted envelope is displayed transparently and colored in green. The bound ligand (biotin) is displayed in stick model.

Evaluating pocket predictions. After predicting the location of potential binding envelopes, we evaluated the quality of those predictions. Ideally we would like to capture the binding area as closely as possible. However, this requirement is not strict because, obviously, the same pocket may bind different ligands which may share a core site but extend in different directions. The accuracy of each prediction was measured by the overlap of protein atoms in contact with the ligand and protein atoms in contact with the predicted envelope. This relative overlap (RO (Relative Overlap of predicted patch to the real binding patch)) parameter was calculated as follows:

$$RO = (A_L \cap A_E) / A_L$$

where, A_L is the solvent accessible area of the receptor atoms within 3.5 Å from a bound ligand, and A_E is the solvent accessible area of the receptor atoms within 3.5 Å from the predicted envelope. A perfect prediction would have RO close to 1.0, and a failed prediction would have RO equal to zero.

Validating the site prediction algorithm. We first applied DrugSite to the LP-set in order to evaluate its performance on complexed structures. The method successfully identified approximately

98.8% of the 5616 ligand binding sites as having a non-zero relative overlap ($RO > 0$) between the observed and predicted sites (see the definition of the RO measure above). Furthermore, 55.2% of binding sites were perfectly identified ($RO = 1.0$) and 85.7% of binding sites were predicted with RO higher than 0.8, i.e. greater than 50.0% coverage (Fig. 2).

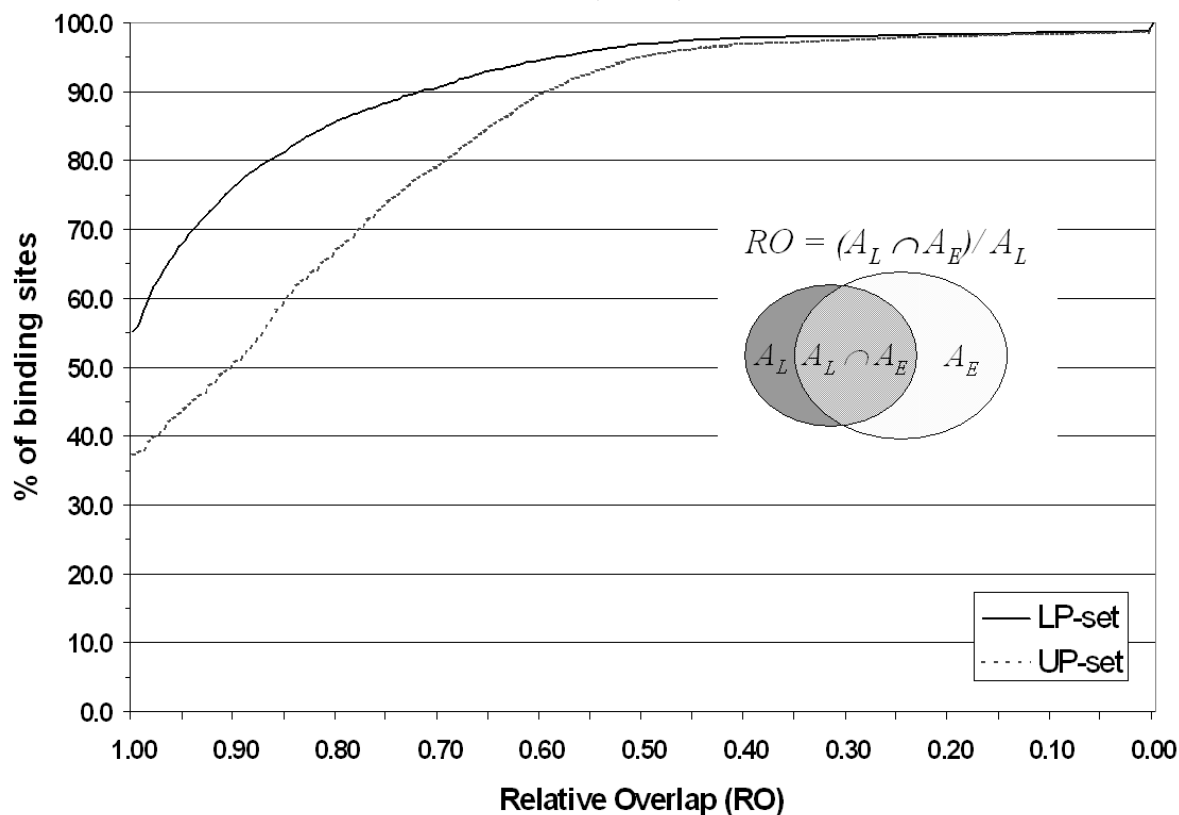


Figure 2: Accuracy of the prediction measured by the relative overlap (RO) between the predicted binding patch A_E (solvent accessible area of the receptor atoms within 3.5 Å from the predicted envelope) and the observed binding patch A_L (solvent accessible area of the receptor atoms within 3.5 Å from a bound ligand). Result of 5,616 binding sites from protein-ligand complexes (LP-Set) and 11,510 binding sites from uncomplexed structures (UP-Set) were sorted separately by RO .

Predicting binding sites from uncomplexed structures. A more realistic test was performed using a dataset consisting of 11,510 binding sites collected from uncomplexed structures (the UP-Set). We expected that due to side-chain movements the predictability of the binding pockets would be somewhat reduced. Surprisingly, 98.6% of binding sites had some overlaps ($RO > 0$), being almost the same as LP-Set. Even for RO higher than 0.5, the difference was very small (95.0% versus 96.7%). These results suggest that the method is not sensitive to the conformational variability of binding sites in uncomplexed structures. However, in the high RO region, only 67.0% of the binding sites showed RO higher than 0.8, while LP-set showed 85.7% (Fig. 2). For the percentage of perfectly predicted binding sites ($RO = 1.0$), it was 37.5% compared to 55.2% for the LP-Set.

We studied the effect of different conformational rearrangements observed in the unliganded-binding sites of the UP-Set upon the prediction success. Because the 11,510 binding sites of the UP-Set were inferred based on 1445 binding sites in the LP-Set, we grouped them into 1445 clusters and sorted the clusters by the RO of LP-Set. The distribution of the RO values of UP-Set for 1445 clusters is shown in Fig.3. From this distribution we observed: (i) Only a few (1.4% of UP-Set) cases with $RO=0$, the overwhelming majority had $RO > 0.6$; (ii) For the clusters with high RO of LP-Set ($RO > 0.9$), the prediction of UP-Set decreased the RO compared with LP-Set, but the majority was still above 0.8. However some cases had much larger deterioration of the ligand coverage; (iii) For the clusters with lower RO of LP-Set ($RO < 0.9$), the picture was mixed. We observed better overlap for

the UP-Set conformations than the LP-set, implying some “opening” of unliganded pockets. However, the majority of all 1445 clusters showed the unliganded pockets appear more closed. Overall, we may conclude that even though the unliganded conformations of receptors were not good for binding site detection, there was only a marginal deterioration of recognition performance.

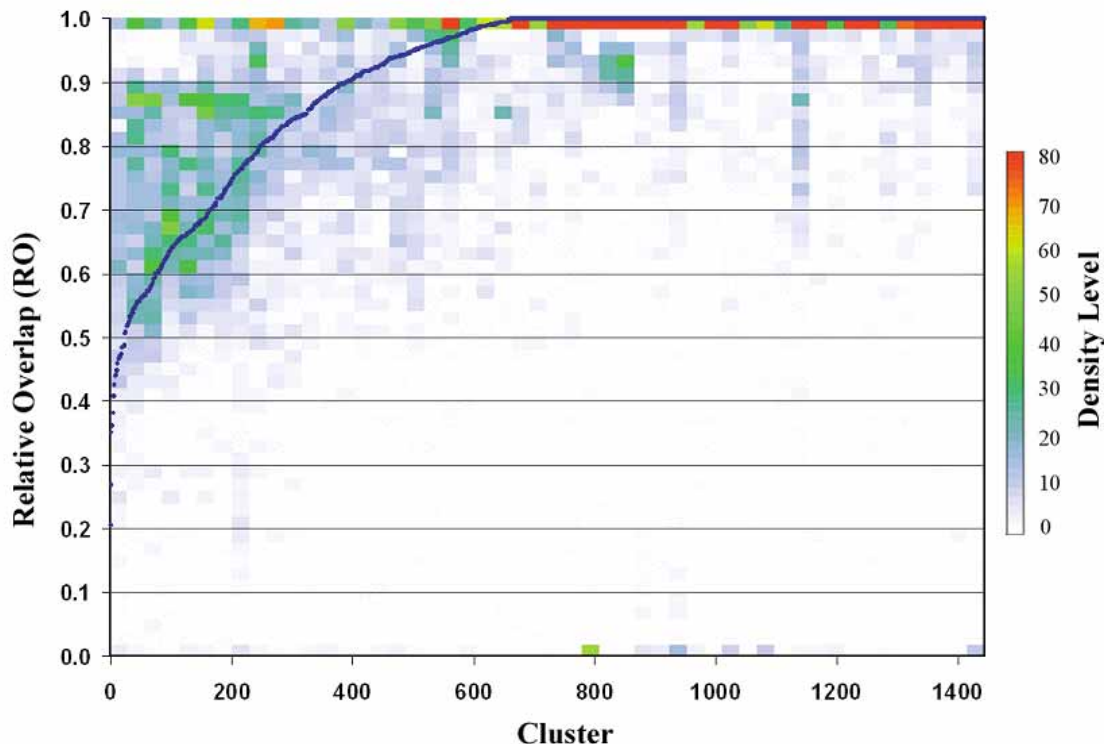


Figure 3: Dependence of the pocket predictions on conformational variations of apo-structures. UP-Set is grouped into 1445 clusters based on the corresponding binding sites in the LP-Set. The results are sorted by the RO values of the LP-set (blue dots). The density of the points of the UP-Set in each mesh (mesh size: 30 clusters by 0.02 RO range) is represented by color. The density higher than 80 was truncated to 80.

Rank of the real binding sites. The algorithm may produce any number of putative binding envelopes depending on the nature of the protein surface and the cutoff volume of a predicted envelope. If several envelopes are generated, it is important to know which envelope corresponds to the real binding site in the receptor. In the majority of cases, we found that the envelopes overlapping with the ligand were the largest envelopes. More specifically, 80.9% of the predicted envelopes overlapping with the ligand were the largest one and 11.8% the second largest. We sorted the envelopes by volume and Fig. 4 shows the rank of real pocket. This implies that although the algorithm may return several putative binding sites, the top two cover as high as 92.7% of the real binding sites. Laskowski and his colleagues reported a similar result based on a dataset of 67 single-chain enzymes [17].

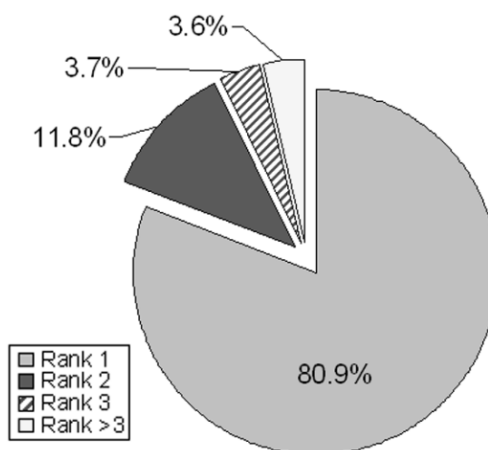


Figure 4: Rank of the real binding sites in the predicted putative binding site lists. The predicted binding sites were sorted by the volumes of the envelopes. The top two largest predicted sites covered 92.7% of the real binding sites.

Size of the predicted binding sites. The size of the predicted envelope is another important criterion of the prediction, because over-sized envelopes contribute to binding site coverage but lose precision by increasing false positives. The DrugSite was optimized to generate envelopes that just fill the binding sites. In order to investigate the false positive of the prediction, we calculated (i) the volumes of the envelopes with respect to the volumes of the bound ligands, and, (ii) the ratio of the predicted binding patch to the whole surface area of the receptor for each binding site in the LP-Set. For the volume, the average value of the ligands was 439.6 \AA^3 while the average of the envelopes was 610.8 \AA^3 . On average, the envelope was approximately 1.4 times larger in volume than the ligand. Considering that most ligands should fit inside their pockets, and the ligands bound to those receptors were just some represents of all possible ligands, this value is very reasonable. For the predicted binding patch, the average ratio to the whole surface of the receptor was 4.7%. Fig. 5 shows the distribution of the ratio of contact area to the whole surface area of the protein for bounded ligands and predicted envelopes. This distribution showed that the size of predicted binding patch was very close to the real binding area. Consider together with the high relative overlap (RO) values of the prediction, we can conclude that the prediction was very accuracy.

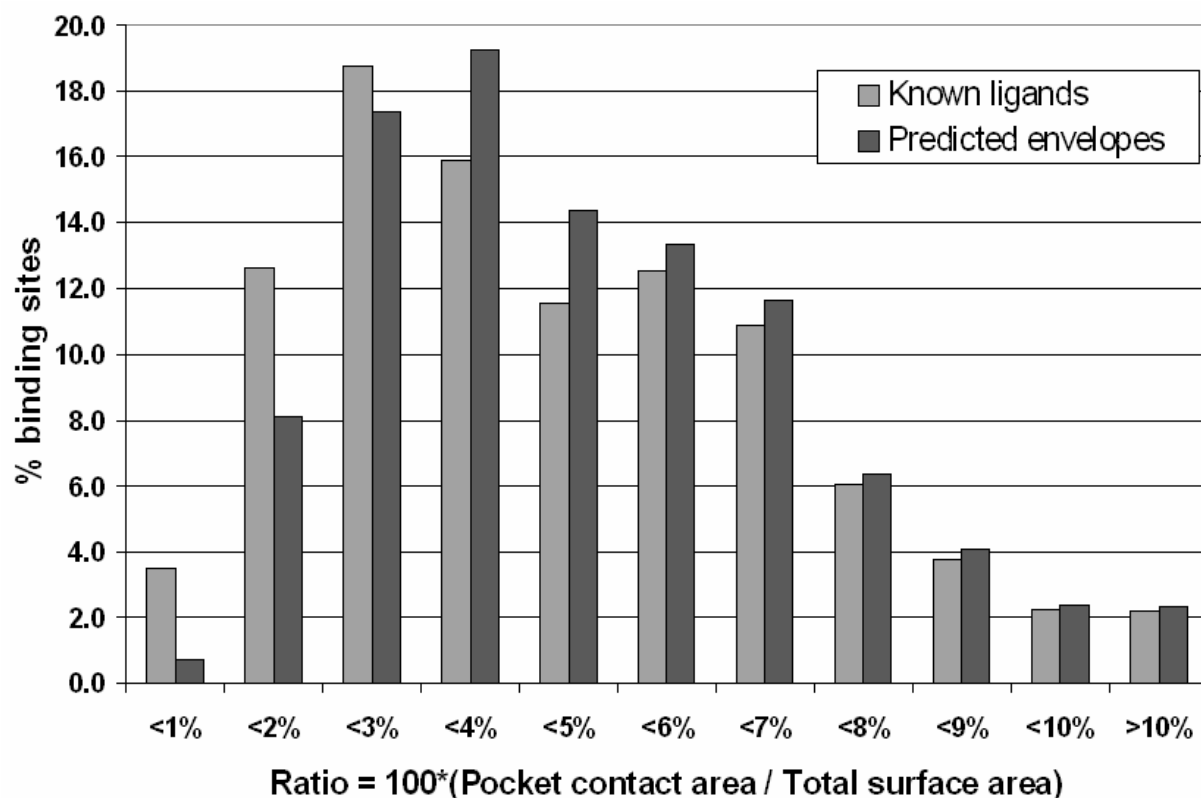


Figure 5: Distribution of the ratio of predicted contact area to total surface area of the receptor (LP-Set).

Pocket collection. We annotated the proteins in the LP-Set based on the SCOP (Structural Classification of Proteins) database [4]. They occupied 10 out of 11 classes of SCOP, showing high diversity of our dataset. They distributed in 261 folds, 347 superfamilies and 589 families. Fig. 6 shows a small selection of the predicted envelopes with the proteins. All the proteins are from different fold and only the first two largest envelopes were displayed for each protein. These 20 representatives were selected based on the following criteria: (i) the envelope size was between 300 \AA^3 and 700 \AA^3 , the most populated envelopes; (ii) it was the only member of the fold, this avoided the artificial picking of representative proteins from folds. We then sorted these folds by their SCOP fold ID and chose the first 20 folds. The complete datasets (LP-Set and UP-Set) and prediction results can be found at <http://abagyan.scripps.edu/pockets/>.

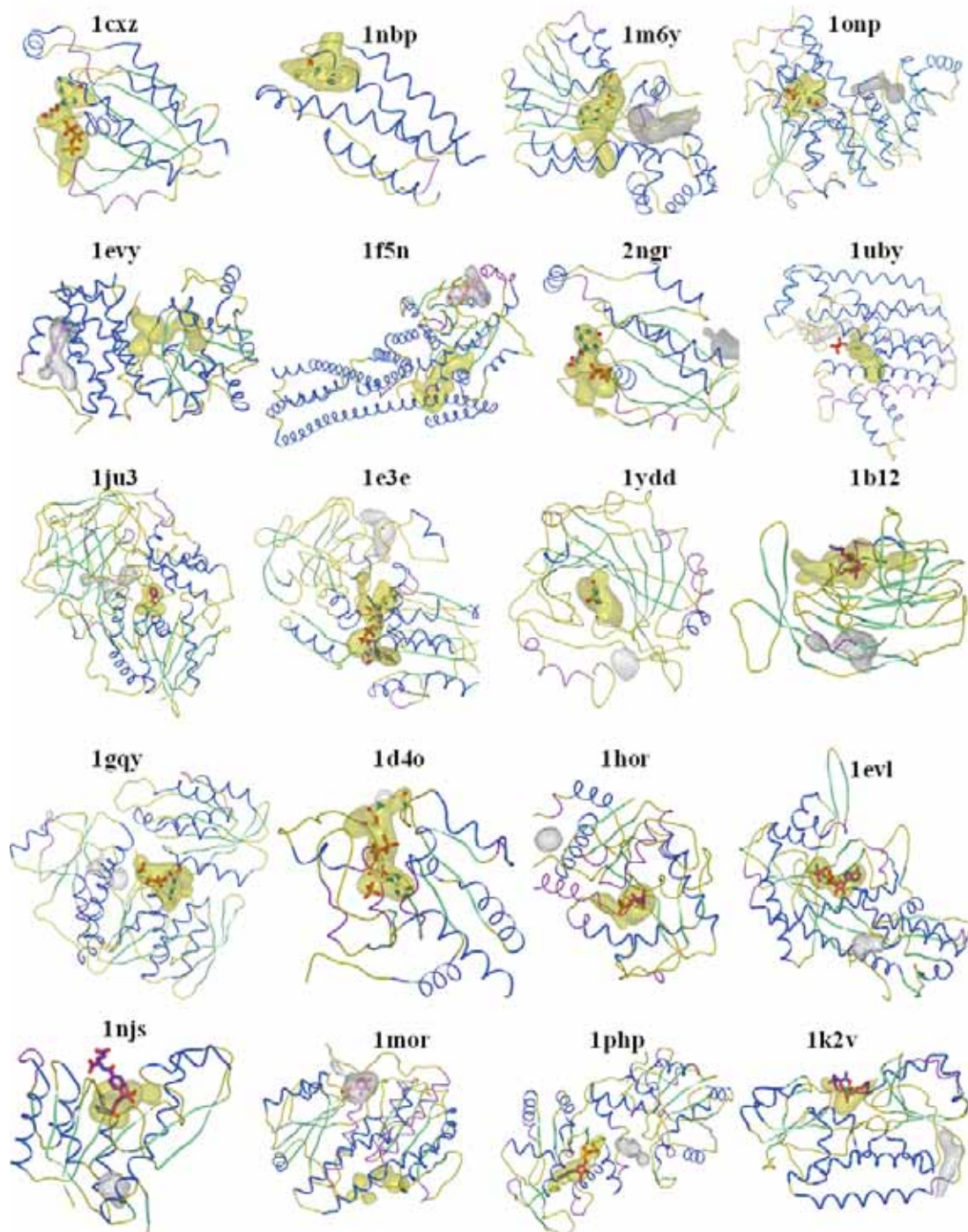


Figure 6: A collection of predicted envelopes. Only two largest envelopes were displayed for clearness. The largest envelopes were colored in yellow and the second largest in gray. The bound ligands were displayed in stick model.

3 Discussion and Conclusion

We have demonstrated that DrugSite can successfully identify and predict protein-ligand binding sites collected systematically from both complexes and apo-structures. There were a small number of binding sites that cannot be identified by the algorithm. For 5,616 binding sites from protein-ligand complexes, only 66 (1.2%) remained unidentified. We visually examined all these cases and found that these binding sites were either very small or very shallow. We successfully identified all of them by adjusting the parameters (contouring level and volume cutoff of the envelope) of the program. However, adjusting the parameters to detect small or shallow binding sites results in higher false positives by producing more putative sites or larger envelopes. Practically, if the program always returns a long list of putative binding sites, the prediction is of little value. For large-scale identification, we use parameters from this study because they were optimized based on a large, diverse training set and this comprehensive identification of the binding sites show they are able to generate good results. However, when applying the method to a specific protein, those “average” parameters can be adjusted based on the intent and the feed back of the result. If no pocket can be founded by using the default “average” parameters or you are going to find some “pocket potential” for peptide binding site (they are small and shallow), you can try to adjust the parameters to find smaller or shallower pockets.

In conclusion, of 5,616 protein-ligand binding sites of complexes we tested, DrugSite correctly identifies 98.8% of the known binding sites. 85.7% of the binding sites showed coverage of the known contact area higher than 80.0%. For correctly identified binding sites, 80.9% were ranked first and 11.8% second. The average ratio of the predicted binding patch to the total surface area of the protein was 4.7%, implying that the prediction was accuracy and with low false positive rate. The prediction rate for 11,510 binding sites from apo-structures (UP-Set) was very close to that of the LP-Set. The relative overlap (RO) was lower than that of the LP-Set but still having 67.0% showed higher than 80.0% coverage of the real binding patch ($RO > 0.8$).

This fast and accurate pocket prediction algorithm DrugSite can be used to identify possible binding site locations for orphan receptors, or for uncharacterized secondary binding sites of known receptors. It can also be used to prioritize novel targets by the “druggability” of identified pockets. In addition, applying the algorithm to separated protein subunits and evaluating the strength of “pocket potential” can help evaluate the feasibility of protein-protein interaction inhibition. Furthermore, directing ligand design can be carried out along the predicted envelope.

Using the algorithm, we can identify and collect all potential small molecule binding pockets in a genome, cluster them into classes and categories according to their size, shape and physicochemical properties. The study of this “pocketome” can help us characterize the genome from a new and critical aspect, because the binding sites are the most important elements contributing to the functions of the structures. Investigating the relationship between small molecules and the pockets, i.e., matching a drug candidate to a pocketome can potentially be used to evaluate the side effects of the drug.

Acknowledgments

We thank Brian Marsden, Sanjay Saldanha and Colin Smith for the discussion of the manuscript. We also thank MolSoft LLC for making the ICM program available for this research. This work was supported by the Department of Energy (ER63042).

References

- [1] Abagyan, R. and Totrov, M., High-throughput docking for lead generation, *Curr. Opin. Chem. Biol.*, 5:375–382, 2001.

- [2] Abagyan, R., Totrov, M., and Kuznetsov, D., ICM: A new method for structure modeling and design: Applications to docking and structure prediction from the distorted native conformation, *J. Comput. Chem.*, 15:488–506, 488–506, 1994.
- [3] Anderson, S. and Chiplin, J., Structural genomics: Shaping the future of drug design?, *Drug Discov. Today*, 7:105–107, 2002.
- [4] Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J., Chothia, C., and Murzin, A.G., SCOP database in 2004: Refinements integrate structure and sequence family data, *Nucleic Acids Res.*, 32 Database issue:D226–229, 2004.
- [5] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E., The protein data bank, *Nucleic Acids Res.*, 28:235–242, 2000.
- [6] Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, Jr., E.E., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M., The protein data bank: A computer-based archival file for macromolecular structures, *J. Mol. Biol.*, 112:535–542, 1977.
- [7] Bliznyuk, A. and Gready, J., Simple method for locating possible ligand binding sites on protein surfaces, *J. Comput. Chem.*, 9:983–988, 1999.
- [8] Brady, Jr., G.P., and Stouten, P.F., Fast prediction and visualization of protein binding pockets with PASS, *J. Comput. Aided Mol. Des.*, 14:383–401, 2000.
- [9] Campbell, S.J., Gold, N.D., Jackson, R.M., and D.R., Westhead Ligand binding: Functional site location, similarity and docking, *Curr. Opin. Struct. Biol.*, 13:389–395, 2003.
- [10] Dennis, S., Kortvelyesi, T., and Vajda, S., Computational mapping identifies the binding sites of organic solvents on proteins, *Proc. Natl. Acad. Sci. USA*, 99:4290–4295, 2002 .
- [11] Gane, P.J. and Dean, P.M., Recent advances in structure-based rational drug design, *Curr. Opin. Struct. Biol.*, 10:401–404, 2000.
- [12] Glick, M., Robinson, D.D., Grant, G.H., and Richards, W.G., Identification of ligand binding sites on proteins using a multi-scale approach, *J. Am. Chem. Soc.*, 124:2337–2344, 2002.
- [13] Hendlich, M., Rippmann, F., and Barnickel, G., LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins, *J. Mol. Graph. Model*, 15:359–363, 389, 1997.
- [14] Klebe, G., Recent developments in structure-based drug design, *J. Mol. Med.*, 78:269–281, 2000.
- [15] Kortvelyesi, T., Silberstein, M., Dennis, S., and Vajda, S., Improved mapping of protein binding sites, *J. Comput. Aided Mol. Des.*, 17:173–186, 2003.
- [16] Laskowski, R.A., SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions, *J. Mol. Graph.*, 13:323–330, 307–308, 1995.
- [17] Laskowski, R.A., Luscombe, N.M., Swindells, M.B., and Thornton, J.M., Protein clefts in molecular recognition and function, *Protein Sci.*, 5:2438–2452, 1996.
- [18] Levitt, D. and Banaszak, L., POCKET: A computer graphics method for identifying and displaying protein cavities and their surrounding amino acids., *J. Mol. Graphics*, 10:229–234, 1992.
- [19] Liang, J., Edelsbrunner, H., and Woodward, C., Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design, *Protein Science*, 7:1884–1897, 1998.

- [20] Lichtarge, O. and Sowa, M.E., Evolutionary predictions of binding surfaces and interactions, *Curr. Opin. Struct. Biol.*, 12:21–27, 2002.
- [21] Lichtarge, O., Yao, H., Kristensen, D.M., Madabushi, S., and Mihalek, I., Accurate and scalable identification of functional sites by evolutionary tracing, *J. Struct. Funct. Genomics*, 4:159–166, 2003.
- [22] Lipinski, C.A., Drug-like properties and the causes of poor solubility and poor permeability, *J. Pharmacol. Toxicol. Methods*, 44:235–249, 2000.
- [23] *MolSoft ICM 2.8 Program Manual*, MolSoft LLC, San Diego, 2000.
- [24] Peters, K.P., Fauck, J., and Frommel, C., The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria, *J. Mol. Biol.*, 256:201–213, 1996.
- [25] Ruppert, J., Welch, W., and Jain, A., Automatic identification and representation of protein binding sites for molecular docking, *Protein Science*, 6:524–533, 1997.
- [26] Shoichet, B.K., McGovern, S.L., Wei, B., and Irwin J.J., Lead discovery using molecular docking, *Curr. Opin. Chem. Biol.*, 6:439–446, 2002.
- [27] Sottriffer, C. and Klebe, G., Identification and mapping of small-molecule binding sites in proteins: Computational tools for structure-based drug design, *Farmaco.*, 57:243–251, 2002.
- [28] Totrov, M. and Abagyan, R., Flexible protein-ligand docking by global energy optimization in internal coordinates, *Proteins*, Suppl 1:215–220, 1997.
- [29] Veber, D.F., Johnson, S.R., Cheng, H.Y., Smith, B.R., Ward, K.W., and Kopple, K.D., Molecular properties that influence the oral bioavailability of drug candidates, *J. Med. Chem.*, 45:2615–2623, 2002.
- [30] Verdonk, M.L., Cole, J.C., Watson, P., Gillet, V., and Willett, P., SuperStar: Improved knowledge-based interaction fields for protein binding sites, *J. Mol. Biol.*, 307:841–859, 2001.
- [31] Walters, W.P., Stahl, M.T., and Murcko, M.A., Virtual screening - an overview, *Drug Discov. Today*, 3:160–178, 1998.