Pocketome via Comprehensive Identification and Classification of Ligand Binding Envelopes*

Jianghong An‡, Maxim Totrov§, and Ruben Abagyan‡¶

We developed a new computational algorithm for the accurate identification of ligand binding envelopes rather than surface binding sites. We performed a large scale classification of the identified envelopes according to their shape and physicochemical properties. The predicting algorithm, called PocketFinder, uses a transformation of the Lennard-Jones potential calculated from a three-dimensional protein structure and does not require any knowledge about a potential ligand molecule. We validated this algorithm using two systematically collected data sets of ligand binding pockets from complexed (bound) and uncomplexed (apo) structures from the Protein Data Bank, 5616 and 11,510, respectively. As many as 96.8% of experimental binding sites were predicted at better than 50% overlap level. Furthermore 95.0% of the asserted sites from the apo receptors were predicted at the same level. We demonstrate that conformational differences between the apo and bound pockets do not dramatically affect the prediction results. The algorithm can be used to predict ligand binding pockets of uncharacterized protein structures, suggest new allosteric pockets, evaluate feasibility of protein-protein interaction inhibition, and prioritize molecular targets. Finally the data base of the known and predicted binding pockets for the human proteome structures, the human pocketome, was collected and classified. The pocketome can be used for rapid evaluation of possible binding partners of a given chemical compound. *Molecular & Cellular Proteomics* 4:752–761, 2005.

Prediction of ligand binding sites is a fundamental step in the investigation of the molecular recognition mechanism and function of a protein. An increasing number of protein structures are becoming available from high throughput structural genomic projects prior to biological and functional characterization. Therefore, 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 (1–3). Here we attempted to develop an algorithm that is based solely on the protein structure and without any prior knowledge about the nature of the substrate. We hypothesized that the structure itself is sufficiently informative, whereas the evolutionary conservation and the nature of the ligand can only be used as optional contributions.

Proteins are involved in several kinds of molecular interactions: with other proteins, DNA, RNA, peptides, and small molecules. In this study we present an algorithm to predict the binding envelopes near potential small ligand binding sites or areas that could be targeted with small “druglike” compounds. Once the ligand binding pocket is predicted, a high throughput ligand docking procedure or structure-based drug design (4–9) can be used to generate a list of the lead molecules. The properties of druglike molecules are well studied (10, 11) and cover a certain range of sizes, typically with molecular mass between 300 and 700 daltons. Therefore, we excluded from consideration very small ligands, such as metals and small solvent molecules, as well as very large substrates. However we wanted to develop an algorithm that within this size range does not depend on the nature of the ligand.

A number of structure-based pocket prediction algorithms have been published over the last 10 years. They can be divided into two general classes: (i) geometric algorithms and (ii) probe mapping/docking algorithms. Geometric approaches analyze protein surfaces to find clefts. SURFNET (12) detects the gap regions in proteins by fitting spheres into the spaces between protein atoms. The sphere fitting process results in a number of separate groups of interpenetrating spheres, which correspond to the cavities and clefts of the protein. LIGSITE (13), an improved version of POCKET (14), 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 (15) and CAST (16) are based on the a-shape algorithm, which identifies pockets by comparing surfaces of the protein generated with different levels of detail. PASS (17) identifies the “active site points” by coating the protein surface with a layer of spherical probes and then filtering out...
those that clash with the protein or are not sufficiently buried. In addition to those pure geometrical methods, some approaches based on mapping/docking and scoring of molecular fragments have been proposed (18–23). Two excellent recent reviews of computational tools for identification of small molecule binding sites in proteins give a good overview of the field (1, 24).

Pure geometric methods are relatively straightforward, but there is no direct physical meaning behind them. On the other hand, methods using molecular fragment mapping and ligand docking are better physically justified but computationally expensive and cannot always provide a good discrimination between correct and incorrect sites. Also the previously published methods were typically tested on relatively small data sets. Although APROPS used a relatively large test set of about 300 structures, others only used 10–50 selected test cases. This is despite the fact that almost 30,000 x-ray structures have been deposited in the Protein Data Bank (25, 26). Finally, because the ultimate goal of the binding site prediction methods is to find active sites on uncharacterized structures, it is important to test and validate the algorithms on large sets of the "unbound" or apo structures. Only the PASS algorithm was tested on a data set of 21 apo structures; the other publications did not test the effect of induced conformational changes on the prediction accuracy. A benchmark test based on a large, systematic data set of apo structures is necessary for evaluating protein-ligand binding site identification methods.

In this study, we present and validate a novel algorithm for prediction of ligand binding pockets. The algorithm called PocketFinder is based on a transformation of the Lennard-Jones potential. Like pure geometric approaches, PocketFinder is fast and capable of identifying clefts and cavities regardless of the nature of the substrate while being more sensitive and specific. Furthermore, in contrast to other methods, the PocketFinder algorithm not only detects the location of the binding pocket but also predicts envelopes representing the shape and size of putative ligand binding volume. The method was tested on a systematically collected data set 2 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. All small molecule binding envelopes from the human structural proteome were collected and clustered into a pocketome. The predicted binding envelopes were hierarchically clustered. The complete pocketome may be useful for understanding a complex network of interactions between small molecules and the cell proteins.

MATERIALS AND METHODS

All three-dimensional protein structures were taken from the October 3, 2003 Protein Data Bank release. Prior to computation, we removed all ligands and water molecules from the structure. Protein-ligand binding pockets are predicted based on the grid potential map of van der Waals interaction of the receptor. To determine the regions of consistently high van der Waals attraction, the following four-step procedure was applied.

Step 1—The first step was to create the grid potential map of the van der Waals force field using a probe atom (parameters for an aliphatic carbon were used) in orthogonal parallelepipeds surrounding receptor atoms (27). The grid had 1.0-Å spacing, and a margin of 1.0 Å beyond the dimensions of the protein was added. The potential was calculated according to the Lennard-Jones formula,

\[ P_{ij}^r = \sum_{i=1}^{N} \left( \frac{A_{xc}}{r_{ij}^{12}} - \frac{B_{xc}}{r_{ij}^{6}} \right) \]

where \( r_{ij} \) is the distance between the probe placed at a grid node and the protein atom \( X_i \). Parameters \( A_{xc} \) and \( B_{xc} \) are taken from the Empirical Conformational Energy Program for Peptides (ECEPP/3 molecular mechanics force field). \( P_{ij}^r \) values were further truncated into the \( \min(P_{ij}^r), -0.8 \) range to retain only the attractive regions.

Step 2—The second step was to smooth (space average) the potential map to emphasize the regions with the van der Waals potential consistently high across significant space span and to avoid excessive density fragmentation. Smoothing was performed by an iterative averaging of the potential \( P_{ijk}^n \) on each grid node \((i,j,k)\) with the adjacent nodes.

\[ P_{ijk}^{n+1} = \left( P_{ijk}^n + (P_{ijk}^n)_{i-1,j,k} + P_{ijk}^n_{i+1,j,k} + P_{ijk}^n_{i,j-1,k} + P_{ijk}^n_{i,j+1,k} + P_{ijk}^n_{i,j,k-1} + P_{ijk}^n_{i,j,k+1} \right) / 6 \]

Iterative application of this transformation closely approximates a convolution with an averaging function.

\[ P(\bar{r}) = \int f(\bar{p}, \bar{r}) P_{ijk}^n d\bar{p}, \quad f(\bar{p}, \bar{r}) = e^{-\left( \frac{\bar{r}^2}{2} \right)}, \quad \lambda = \sqrt{\frac{2}{3} N_{\text{iter}}} \]

Ten iterations were performed, corresponding to the smoothing length parameter (radius) \( \lambda \) of \(-2.6 \) Å.

Step 3—The third step was to create putative ligand envelopes by contouring the resulting map at a level \( P_{\text{cont}} \) calculated as follows,

\[ P_{\text{cont}} = \text{mean}(P_{ijk}) - \sigma \times \text{r.m.s.d.}(P_{ijk}) \]

where the threshold parameter \( \sigma = 4.6 \) was established on the training data set, \( \text{mean}(P_{ijk}) \) and \( \text{r.m.s.d.}(P_{ijk}) \) are an average and a root mean square difference of all map potential values \( P_{ijk} \).

Step 4—The fourth step was to sort the created envelopes by their volumes and filter out those smaller than 100 Å³. All parameters for the map transformations were optimized using a large, diverse set of binding sites. The algorithm was coded with the ICM1 scripting language (28, 29).

RESULTS

Because our main objective was to develop and validate an accurate algorithm for predicting protein-ligand binding sites

1 The abbreviations used are: ICM, internal coordinate mechanics; LP-Set, liganded pocket set collected from complexes; UP-Set, unliganded pocket set collected from apo structures; RO, relative overlap of predicted patch to the real binding patch; SCOP, Structural Classification of Proteins; MDM2, ubiquitin-protein ligase e3 mdm2; LFA-1, lymphocyte function associated antigen 1.
Identification and Classification of Ligand Envelopes

to which a typical druglike 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 for benchmarking the performance of the pocket prediction algorithm.

Compiling an Exhaustive Benchmark of Ligand Binding Sites—We started from the October 30, 2003 release of the Protein Data Bank that contained 17,730 crystallographic structures of proteins. Then we collected all structures that are complexed with heteromolecules (tagged by HET keywords in the Protein Data Bank) to compile a data set of observed binding sites from protein-ligand complexes (liganded-pocket set (LP-Set)). This resulted in 13,431 Protein Data Bank entries and 3,561 unique heteromolecules. Several filters were used to produce the liganded protein-ligand binding site data set. First, we introduced a ligand size and frequency filter. Heteromolecules containing fewer than seven heavy atoms were excluded in this research. This excluded metals and popular crystallization buffer components. To avoid a benchmark bias toward the most frequent ligands, we excluded high frequency cofactors or substrates, such as hemes, N-acetyl-D-glucosamine, α-D-mannose, certain sugars, etc. This filter reduced the number of entries to 7,275 (about 54% of all complexes). Second, a receptor quality filter was used. Only entries with resolution better than 2.5 Å were retained. Proteins that consist of fewer than 50 or more than 2,000 residues were also excluded. This filter reduced the number of entries to 5,736. We applied other filters that did not reduce the size of the data set significantly but cleaned up the data as follows: (i) heteromolecules that are far away from the receptors (no atoms of the heteromolecules were within 3.5 Å from the receptor) were removed; (ii) heteromolecules that contact the symmetric parts of the receptor (within 3.5 Å in distance) were removed because their binding sites are formed between the asymmetric units, and building a correct model requires biological information; (iii) ion clusters were removed; and (iv) duplicate combinations of Protein Data Bank entries and ligands were removed. The final data set consisted of 5,616 protein-ligand binding sites (LP-Set). That is the combination of 4,711 Protein Data Bank entries with 2,175 unique ligands.

Collecting Protein-Ligand Binding Sites from the Uncomplexed Structures—Because our main goal was to predict a potential binding envelope from an uncomplexed structure, it was critical to compile a benchmark of unliganded pocket sites (UP-Set). 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. The UP-Set was collected by superimposing the protein structure in the LP-Set onto its uncomplexed homologues and mapping the ligand to them. The proteins (apo structures) in the UP-Set had to meet the following criteria. (i)

The sequence identity between apo structure and complexed protein is over 95%. (ii) The resolution of the structure is better than 2.5 Å. (iii) There are no mutations among the surface residues within 8 Å around the mapped ligand. (iv) There are no other ligands within 8 Å around the mapped ligand in the apo structure.

Only the single chain receptors in the LP-Set were used to find the apo structures. The only reason for using the single chain receptors was to simplify the process of identifying homologues and superimposing the structures. We preserved all alternative binding pockets in the UP-Set to avoid an arbitrary choice of one representative unliganded pocket. A total of 11,510 unliganded pockets were collected from the eligible Protein Data Bank entries. The pocket projection from LP-Set was used to find the unliganded pockets in the UP-Set. As a result, the UP-Set contains on average about eight different mappings of the same binding site in LP-Set.

Representation of the Predicted Ligand Binding Pockets—We calculated transformed version of the three-dimensional Lennard-Jones potential on a 1-Å grid surrounding the entire protein surface (see “Materials and Methods”) and then contoured at an optimized level to create pocket envelopes. 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.

Evaluating Pocket Predictions—After predicting the location of potential binding envelopes, we evaluated the quality of those predictions. 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 pocket.
envelope. This relative overlap (RO) parameter was calculated as follows,

\[
RO = \frac{(A_L \cap A_E)}{A_L} \tag{Eq. 5}
\]

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 completely failed prediction would have RO equal to zero, whereas a perfect prediction would have RO close to 1.0. However, this requirement would be too strict because obviously the same pocket may bind different ligands, and all these ligands may share a core site but extend in different directions. We define the threshold of successful prediction as RO \(\geq\) 0.5, i.e. at least 50% of the accessible area \(A_L\) has to be overlapped by the predicted area \(A_E\).

Validating the Pocket Prediction Algorithm—We first applied PocketFinder to the LP-Set to evaluate its performance on holo structures. The method successfully identified 96.8% of the 5,616 ligand binding sites as having a relative overlap RO \(\geq\) 0.5 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), i.e. at least 100% of the accessible area \(A_L\) has to be overlapped by the predicted area \(A_E\). The results 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.

The prediction accuracy measured by the RO between the predicted binding patch \(A_E\), defined as the solvent-accessible surface of the receptor atoms within 3.5 Å from the predicted envelope, and the observed binding patch \(A_L\), defined as the solvent-accessible surface of the receptor atoms within 3.5 Å from the bound ligand. The results 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 data set 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 95.0% of the binding sites were predicted with RO \(>\) 0.5. These results suggest that the method is tolerant to the conformational variability of binding sites in uncomplexed structures. However, in the high RO region, only 67.0% of the apo binding sites showed RO higher than 0.8 as compared with 85.7% in the LP-Set (Fig. 2). For the percentage of completely overlapped binding sites (RO = 1.0), it was 37.5% compared with 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 1,445 binding sites in the LP-Set, we grouped them into 1,445 clusters and sorted the clusters by the RO of LP-Set. The distribution of the RO values of UP-Set for 1,445 clusters is shown in Fig. 3. From this distribution we observed the following. (i) The overwhelming majority of UP-Set had RO \(>\) 0.5. (ii) For the clusters with high RO of LP-Set (RO \(>\) 0.9), the relative overlap of the UP-Set decreased compared with the LP-Set, but the majority were still above 0.8. However, some cases had much larger deterioration of the relative overlap. (iii) For the clusters with lower RO of LP-Set (RO \(<\) 0.8), the picture was mixed. We observed better overlap for the UP-Set conformations than the LP-Set, implying some “opening” of unliganded pockets. However, for the majority of 1,445 clusters the unliganded pockets appear more difficult to predict than the liganded pockets. Overall we may conclude that although the unliganded conformations of receptors were not as good as the liganded conformations for the binding site identification, there was only a marginal deterioration of recognition performance.

To investigate the difference of the predicted envelopes between holo and apo structures, we applied our algorithm to a set of five enzymes with induced domain movements as described by Hayward (30). For each pair, we superimposed the core domains of the proteins. For enzymes with relatively small domain movements (Fig. 4, A and B), the envelopes predicted from both holo and apo conformations are very similar and perfectly captured the bound ligands. For the enzymes with large domain movements (Fig. 4, C, D, and E),
The envelopes show different sizes and shapes between the holo and apo structures. However, the ligand in case C was still successfully captured, and the ligands in cases D and E were partially captured.

Number of Predicted Envelopes and 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. Fig. 5A depicts the relationship between the protein size and the number of predicted pockets. The number of envelopes predicted per protein is roughly proportional to the volume of the protein (one pocket per 10,000 Å³); for the majority of the proteins (volume is below or close to 50,000 Å³) in the data set, less than four envelopes were predicted. If several envelopes are generated, it is important to know which envelope corresponds to the real binding site(s) 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 ones, and 11.8% were the second largest. We sorted the envelopes by volume, and Fig. 5B shows the rank of real pocket. This implies that although the algorithm may return several additional putative binding sites, the top two largest predicted binding sites cover as many as 92.7% of the real binding sites. Laskowski and colleagues (31) reported a similar result based on a data set of 67 single chain enzymes.

Size of the Predicted Envelopes—The size of the predicted envelope is another important criterion of the prediction because oversized envelopes contribute to binding site coverage but are less precise at pointing to its exact location. The PocketFinder was optimized to generate envelopes that just
fill the binding sites. To investigate the extent of overprediction, 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. The average value of the ligands was 439.6 Å³, whereas the average volume of the envelopes was 610.8 Å³. On average, the predicted envelope was ~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 representative examples of all possible ligands, this value seems reasonable. For the predicted binding patch, the average ratio to the whole surface of the receptor was 4.7%.

Fig. 5C shows the distribution of the ratio of contact area to the whole surface area of the protein for bound ligands and predicted envelopes. This distribution shows that the size of the predicted binding patch was close to the real binding area. Considered together with the high RO values of the prediction, we can conclude that the prediction was accurate.

Sample of Predicted Envelopes—Unfortunately there are too many pockets in the benchmark to show each prediction graphically. However, to leave a visual impression of the results, we created an unbiased subset of the benchmark pockets using the following selection procedure. We first annotated the proteins in the LP-Set based on the Structural Classification of Proteins (SCOP) data base (32). They belonged to 10 of 11 classes in SCOP, showing high diversity of our data set. Structures from 261 folds, 347 superfamilies, and 589 families were represented. These 20 representatives in Fig. 6 were selected from different folds based on the following criteria: (i) the envelope size between 300 and 700 Å³, the range of sizes that was most populated, and (ii) the folds that had only one representative in LP-Set were chosen to avoid any subjective selection of representative proteins from folds. We then sorted these folds by their SCOP fold ID and chose the first 20 folds. For each protein, only the largest and the second largest (if it existed) envelopes were displayed. Fig. 6 demonstrates the accuracy and reliability of the envelope prediction. It is important to note that in most cases the predicted envelopes captured the ligands correctly with size and shape being close to the bound ligands. The complete data sets (LP-Set and UP-Set) and prediction results are available upon request.

Pocketome: Collection and Classification of Binding Envelopes from a Specific Organism—As the size of structural proteome grows, we can start compiling a pocketome, defined as the collection of all possible small molecule envelopes present in a cell. Here we collected, compared, and classified all envelopes from human protein-ligand complexes with known three-dimensional structure. There are 943 binding pockets from human proteins in the LP-Set. For the envelope clustering, five shape descriptors (volume, surface area, and the three principal axis lengths) and two physicochemical descriptors (Kyte and Doolittle hydrophobicity pa-
rameter (33) and electrostatic charge) were used to characterize the envelope. After representing each envelope by a seven-element vector, a Euclidean distance matrix was calculated. A hierarchical clustering algorithm was used to build a tree from the distance matrix. We compared the cluster tree of the human ligand envelopes with the tree based on the

![A sample of predicted envelopes.](image)

Fig. 6. A sample of predicted envelopes. Only the two largest envelopes are displayed for clearness. The largest envelopes are colored in yellow, and the second largest are in gray. The bound ligands are displayed as stick models. IL-2, interleukin 2; DXR, 1-deoxy-D-xylulose 5-phosphate reductoisomerase; GPD, glycerol-3-phosphate dehydrogenase; GBP1, interferon-induced guanylate-binding protein 1; FPPS, farnesyl diphosphate synthase; COCE, cocaine esterase; ADH, alcohol dehydrogenase; HCAi, carbonic anhydrase II; PGK, 3-phosphoglycerate kinase; RHOA, transforming protein RhoA; MRAW, α-adenosyl methyltransferase; G2SP, gtp binding protein; LEP, signal peptidase I; MURC, udp-n-acetylmuramate-l-alanine ligase; NNTM, nadp(h) transhydrogenase; NAGB, glucosamine 6-phosphate deaminase; SYT, threonyl-tRNA synthetase; PUR2, phosphoribosylpyrimidinamidoformyltransferase; GLMS, glucosamine 6-phosphate synthase; FHUD, ferrichrome-binding periplasmic protein.
A snapshot of the human pocketome. 160 human protein structures from 943 human complexes in the LP-Set were used. The sequence similarity, the chemical similarity, and the envelope similarity were used to cluster the protein-ligand binding sites. The binding sites were labeled by the Protein Data Bank code and ligand ID. The resulting “envelope” clustering tree was compared with the chemical and sequence trees. Five groups of ligands in the ligand tree were marked in different colors first and then tracked to the other two trees to reveal a complex many-to-many relationship among sequence, ligand, and envelope. Some representative envelopes for each group with corresponding bound ligands were illustrated on the right side of the figure. A poster of the pocketome derived from 943 human complexes is available upon request.
Identification and Classification of Ligand Envelopes

![Diagram](image1)

**DISCUSSION AND CONCLUSIONS**

We demonstrated that the PocketFinder algorithm can successfully identify or predict protein-ligand binding envelopes from complexed and apo structures. We also compiled the first draft of a human "pocketome" consisting of 943 envelopes. Albeit incomplete, this pocketome can grow along with the structural proteome.

A small number of binding sites could not be identified by the algorithm. For 5,616 binding sites from the LP-Set, 66 cases were completely missed (no relative overlaps, RO = 0) by the identification. We visually examined all these cases and found that these binding sites were either very small or very shallow. All of them could still be identified 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 may result in higher number of false positives. Practically if the program always returns a long list of putative binding sites, the prediction is of little value. However, when applying the method to a specific protein, the "average" optimal parameters can be adjusted based on the additional biological and structural data such as the flexibility of B-factors and the feedback from the inspection of the prediction result. If no pocket can be found by using the default average parameters or if the pockets are likely to be small/shallow, for example in the case of peptide binding sites, one can try to adjust the parameters to find smaller or shallower pockets if there are compelling reasons to expect their existence.

In conclusion, of the 5,616 protein-ligand binding sites of complexes we tested, the PocketFinder method correctly identified 96.8% 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 the largest envelopes, and 11.8% were the second largest. The average ratio of the predicted binding patch to the total surface area of the protein was 4.7%, implying that the prediction was quite specific and had a low false positive rate. The prediction rate for 11,510 binding sites from complexed (LP-Set) was 95.0%, close to that of the LP-Set. The RO in general was somewhat lower than that of the LP-Set, but 67.0% of the binding sites still had 80% or better overlap.

The proposed pocket prediction algorithm can be used to identify possible binding site locations for orphan receptors or for uncharacterized secondary binding sites of known receptors. Furthermore ligand design can be directed by the predicted envelope. 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" at the interface patches may...
help evaluate the feasibility of protein-protein interaction inhibition. Fig. 8 shows some examples of successful protein-protein interaction inhibition by small molecules targeting the transcriptional regulator MDM2 and LFA-1. We found pockets in the protein-protein interfaces from the apo structures that overlapped the bound ligands.

Using our algorithm, we can identify and collect potential small molecule binding envelopes in a structural proteome; cluster them into classes and categories according to their size, shape, and physicochemical properties; and compare this classification with the classification of chemical substrates or ligands. We expect the pocketome to grow along with the structural proteome and the improvements of the envelope prediction and classification methods.

Acknowledgments—We thank Brian Marsden, Colin Smith, Andy Orry, and Sanjay Saldanha for the discussion of the manuscript and Eugene Raush and Levon Budagyan for help in clustering and preparation of the illustrations. J. H. A. and R. A. also thank MolSoft for making the ICM program available for this research.

This work was supported in part by Department of Energy Grant ER63042. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

REFERENCES

1. Campbell, S. J., Gold, N. D., Jackson, R. M., and Westhead, D. R. (2003) Ligand binding: functional site location, similarity and docking. Curr. Opin. Struct. Biol. 13, 389–395
2. Lichtarge, O., Yao, H., Kristensen, D. M., Madabushi, S., and Mihalek, I. (2000) Accurate and scalable identification of functional sites by evolutionary tracing. J. Struct. Funct. Genomics 4, 159–166
3. Lichtarge, O., and Sowa, M. E. (2002) Evolutionary predictions of binding surfaces and interactions. Curr. Opin. Struct. Biol. 12, 21–27
4. Walters, W. P., Stahl, M. T., and Murcko, M. A. (1998) Virtual screening—an overview. Drug Discov. Today 3, 160–178
5. Klebe, G. (2000) Recent developments in structure-based drug design. J. Mol. Med. 78, 269–281
6. Gane, P. J., and Dean, P. M. (2000) Recent advances in structure-based rational drug design. Curr. Opin. Struct. Biol. 10, 401–404
7. Abagyan, R., and Totrov, M. (2001) High-throughput docking for lead generation. Curr. Opin. Chem. Biol. 5, 375–382
8. Shoichet, B. K., McGovern, S. L., Wei, B., and Irwin, J. J. (2002) Lead discovery using molecular docking. Curr. Opin. Chem. Biol. 6, 439–446
9. Anderson, S., and Chiplin, J. (2002) Structural genomics: shaping the future of drug design? Drug Discov. Today 7, 105–107
10. Veber, D. F., Johnson, S. R., Cheng, H. Y., Smith, B. R., Ward, K. W., and Kopple, K. D. (2002) Molecular properties that influence the oral bioavailability of drug candidates. J. Med. Chem. 45, 2615–2623
11. Lipinski, C. A. (2000) Drug-like properties and the causes of poor solubility and poor permeability. J. Pharmacol. Toxicol. Methods 44, 235–249
12. Laskowski, R. A. (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J. Mol. Graph. 13, 323–330
13. 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
14. Levitt, D., and Banaszak, L. (1992) POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. J. Mol. Biol. 10, 229–234
15. Peters, K. P., Fauck, J., and Frommell, C. (1996) The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. J. Mol. Biol. 256, 201–213
16. 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
17. Brady, G. P., Jr., and Stouters, P. F. (2000) Fast prediction and visualization of protein binding pockets with PASS. J. Comput.-Aided Mol. Des. 14, 383–401
18. Dennis, S., Kortvelyesi, T., and Vajda, S. (2002) Computational mapping identifies the binding sites of organic solvents on proteins. Proc. Natl. Acad. Sci. U. S. A. 99, 4290–4295
19. Kortvelyesi, T., Silberstein, M., Dennis, S., and Vajda, S. (2003) Improved mapping of protein binding sites. J. Comput.-Aided Mol. Des. 17, 173–186
20. Ruppert, J., Welch, W., and Jain, A. (1997) Automatic identification and representation of protein binding sites for molecular docking. Protein Sci. 6, 524–533
21. Verdonk, M. L., Cole, J. C., Watson, P., Gillet, V., and Willett, P. (2001) SuperStar: improved knowledge-based interaction fields for protein binding sites. J. Mol. Biol. 307, 841–859
22. Bliznyuk, A., and Greedy, J. (1999) Simple method for locating possible ligand binding sites on protein surfaces. J. Comput. Chem. 9, 983–988
23. Glick, M., Robinson, D. D., Grant, G. H., and Richards, W. G. (2002) Identification of ligand binding sites on proteins using a multi-scale approach. J. Am. Chem. Soc. 124, 2337–2344
24. Sotriffer, C., and Klebe, G. (2002) Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design. Farmaco 57, 243–251
25. Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. E., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. J. Mol. Biol. 112, 535–542
26. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weisig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank. Nucleic Acids Res. 28, 235–242
27. Totrov, M., and Abagyan, R. (1997) Flexible protein-ligand docking by global energy optimization in internal coordinates. Proteins Suppl. 1, 20, 215–220
28. Abagyan, R., Totrov, M., and Kuznetsov, D. (1994) 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
29. MolSoft (2000) ICM 2.8 Program Manual, MolSoft, LLC, San Diego, CA
30. Hayward, S. (2004) Identification of specific interactions that drive ligand-induced closure in five enzymes with classic domain movements. J. Mol. Biol. 339, 1001–1021
31. Laskowski, R. A., Luscombe, N. M., Swindells, M. B., and Thornton, J. M. (1996) Protein clefts in molecular recognition and function. Protein Sci. 5, 2438–2452
32. Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., Chothia, C., and Murzin, A. G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res. 32, D226–D229
33. Kyte, J., and Doolittle, R. F. (1982) A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 106–132

Identification and Classification of Ligand Envelopes