Characterization of Local Geometry of Protein Surfaces with the Visibility Criterion

Bin Li[1], Srinivasan Turuvekere[2], Manish Agrawal[2], David La[3], Karthik Ramani[2]

& Daisuke Kihara[3,1,4,5]*

[1] Department of Computer Science, College of Science
[3] Department of Biological Sciences, College of Science
[2] Department of Mechanical Engineering, College of Engineering
[4] Markey Center for Structural Biology
[5] The Bindley Bioscience Center

Purdue University, West Lafayette, IN, 47907, USA

* Corresponding Author
E-mail: dkihara@purdue.edu
Tel: 1-(765)496-2284
Fax: 1(765)496-1189

Running Title:
Local geometry of protein surfaces

KEYWORDS
protein surface shape, binding site, active site, protein surface signature, protein-ligand docking, voxelization, visibility criterion, protrusion, local geometry

Accepted for Proteins, Structure, Function and Bioinformatics (July 2007)

Abstract

Experimentally determined protein tertiary structures are rapidly accumulating in a database, partly due to the structural genomics projects. Included are proteins of unknown function, whose function has not been investigated by experiments and was not able to be predicted by conventional sequence-based search. Those uncharacterized protein structures highlight the urgent need of computational methods for annotating proteins from tertiary structures, which include function annotation methods through characterizing protein local surfaces. Toward structure-based protein annotation, we have developed VisGrid algorithm which uses the visibility criterion to characterize local geometric features of protein surfaces. Unlike existing methods which only concerns identifying pockets which could be potential ligand binding sites in proteins, VisGrid is aimed to identify also large protrusions, hollows, and flat regions which can characterize geometric features of a protein structure. The visibility used in VisGrid is defined as the fraction of visible directions from a target position on a protein surface. A pocket or a hollow is recognized as a cluster of positions with a small visibility. A large protrusion in a protein structure is recognized as a pocket in the negative image of the structure. VisGrid correctly identified 95.0% of ligand binding sites as one of the three largest pockets in 5616 benchmark proteins. To examine how natural flexibility of proteins affects pocket identification, VisGrid was tested on distorted structures by molecular dynamics simulation. Sensitivity decreased approximately 20% for structures of a root mean square deviation of 2.0 Å to the original crystal structure, but specificity was not much affected. Due to its intuitiveness and simplicity, the visibility criterion will lay the foundation of characterization and function annotation of local shape of proteins.

Introduction

Urgent tasks in bioinformatics are the interpretation of massive genomics and the other omics data[1,2], such as protein-protein interaction [3,4,5,6] and gene expression data [7,8]. Prediction of gene function is of a particular importance because it is an indispensable step of genome sequence annotation[9,10] and thus the basis of a wide variety of biology research [11,12]. Function prediction to genes in a genome is mainly done by employing sequence comparison-based methods, which ranges sequence database search methods [13,14,15], functional domain assignments [16,17,18], and sequence motif searches [19,20]. Recent developments are aimed to have a larger coverage in a genome scale function annotation by employing comparative genomics approaches [21,22,23] and extensive mining of more information from PSI-BLAST search results [24,25,26,27].

Protein tertiary structure can be an additional useful source of information for function prediction when available [28,29]. Structure-based approaches are becoming more and more important as structural genomics projects [30,31,32] have been accumulating an increasing number of tertiary structures of proteins of unknown function. In fact more than 1200 protein structures of unknown function are solved and deposited to Protein Data Bank (PDB) [33] by structural genomics projects, which await functional elucidation[34].

Because the global structure of proteins are better conserved than sequence during evolution in general[35,36], weak homology between proteins could be better detected by comparing protein folds [37,38]. Special attention should be paid, however, to "superfolds", which is adopted by various proteins of different functions [39,40]. More directly, function of a protein can often be predicted by identifying its characteristic local tertiary structure responsible for the function. This applies especially well to enzymes, because they have distinct conserved catalytic residues[41]. Thus a straightforward way to predict an enzymatic function of a protein is to identify catalytic residues by applying known templates of active sites[42,43], by finding characteristic local environment of functional residues[44,45], or by identifying the surface shape of active sites[46,47,48,49]. For example, eF-Site[50] compares surface geometries described as graphs, where nodes of a graph being vertices of triangles of the Connolly surface representation[46]. Sequence conservation is also valuable information to identify functionally important region in a protein structure[51,52,53].

From a structural point of view, functional sites, especially active sites of enzymes can be often identified by detecting a pocket region, because in many cases a catalytic site locates at a large pocket in a protein structure[54,55]. Several methods have been developed for identifying pockets as potential active sites of proteins: CAST uses Delaunay tessellation of protein residues; here, empty triangles construct the cavity regions of the protein[54]. LIGSITE places a protein on a grid and identifies a binding pocket as solvent-accessible grids which are enclosed on both sides by the protein[56]. POCKET detects cavities by placing a trial sphere of a given radius on a 3D grid, and grids with a sphere which does not make a contact in either x, y, or z directions are considered to be cavities[57]. SURFNET identifies cavities by fitting spheres between atoms[58]. Identifying pockets in a protein structure also has its application in ligand-protein docking prediction, as an initial step to narrow the search space for docking [59,60,61,62,63].

The existing algorithms mentioned above focus their application on identification of pocket regions for structure-based ligand design and characterization/classification of active sites of enzymes. However, in the structural genomics era when structures of totally unknown function await investigation, we need more versatile methods which can “annotate” characteristic surface shapes in a given protein structure and link them to functional information. Ultimately, we would like to develop a database and matching algorithms for three-dimensional (3D) local surface motifs which are the signature of each class of protein families. Those 3D surface motifs would be represented as a combination of geometric feature, physicochemical properties, and/or evolutionary information (e.g. sequence conservation). As a step toward this end, here we have developed a novel algorithm, VisGrid, which can identify protrusions, flat regions, hollows, together with pocket regions in a protein surface. Flat regions are important because it is known that many of the known protein-protein interaction interfaces are actually flat[64]. In this manuscript we first introduce the visibility which is the key idea of the VisGrid algorithm. Next, we show examples of characteristic local geometry of proteins identified by VisGrid. Then, we have tested the performance of VisGrid in identifying active sites of enzymes in large benchmark databases of ligand bound and unbound structures. Considering the the possibility of applying VisGrid to predicted structures, we have carefully checked the robustness of the algorithm by carrying out the benchmark on distorted structures from the crystal structure by a molecular dynamics (MD) simulation program[65].

Materials and Methods

The Visibility Criterion

The visibility is defined as the fraction of visible directions, *i.e.* directions which are not blocked by protein atoms, from a point in a 3D grid where a target protein structure is projected (Fig. 1). The first step is voxelization of the target protein structure. A protein is projected onto a 3D grid which is large enough to contain all the atoms of the protein. The unit grid size used is 0.9 Å. Then any voxel (cell of the grid) within a sphere centering a protein atom with the radius of the van der Waals radius of the atom plus the radius of a water molecule (1.4 Å) is marked as filled by the protein, otherwise marked as empty. The size of a water molecule is added to take the accessible surface area of a protein[66] into account. Subsequently, surface voxels are defined among the filled voxels as ones which are adjacent to at least one empty voxel. Thus each voxel in the system is marked as either surface, filled (but not surface), or empty.

Now the visibility of each surface voxel is computed. The total number of possible directions of a voxel is 26 when one surrounding layer of the voxel is considered and 98 when the second layer is counted. A direction is considered to be visible when a ray shot from the target voxel toward that direction does not hit any filled or surface voxel up to 20 steps.

Pockets, hollows, protrusions and flat regions

After the visibility for each voxel is computed, all the voxels whose visibility falls into a predefined range are collected. Then, among the collected voxels, those which are within 2.0 Å to each other are grouped. A voxel is merged into a group if the voxel is closer than 2.0 Å to any one of voxels in the group. Two groups are merged into one group if any pair of voxels from the two groups is closer than 2.0 Å to each other.

A pocket is identified as a set of grouped voxels with a certain visibility below a threshold, and a hollow is an empty space surrounded by filled voxels with a low visibility below a threshold. The procedure of identifying pockets and hollows is the same. As for identifying a protrusion region, we used the negative image of a protein volume, *i.e.* a protrusion in an original protein is identified as a pocket region of the negative image of the protein. This is because identifying pockets in the negative image turned out to be less sensitive to small local bumps of a protein surface, rather than directly grouping voxels with a high visibility in the original protein surface.

Identification of pockets and protrusions are completed using voxels. Then, to be able to name atoms and residues which form identified pockets or protrusion regions, atoms which have a pocket/protrusion voxel within the radius of the van der Waals radius of the atom plus the radius of a water molecule are determined to be included in the pocket/protrusion region.

Finally, from the rest of the regions which are not identified neither of pockets, hollows, nor protrusions, the most flat circular region of the radius of 10 Å centering a surface atom is identified. For a given group of *N* surface atoms (x_i, y_i, z_i) (i = 1 .. N), which are within a circle of 10 Å, a plane, *ax* + *by* + *cz* + *d* = *0*, is fitted in the following way:

The sum of the square of the distance from each atom to the plane, E^2, is

$$E^2 = \sum_{i=1}^{N} \left(\frac{|ax_i + by_i + cz_i + d|}{\sqrt{a^2 + b^2 + c^2}} \right)^2 \qquad (1)`$$

$$E^2 = \sum_{i=1}^{N} (Ax_i + By_i + z_i + D)^2 \qquad (2)$$

when $\sqrt{a^2 + b^2 + c^2} = 1$, and where $A \equiv -\frac{a}{c}$, $B \equiv -\frac{b}{c}$, $D \equiv -\frac{d}{c}$.

A, B, and D are determined such that E^2 is minimized.

Assuming $\frac{\partial E^2}{\partial A} = \frac{\partial E^2}{\partial B} = \frac{\partial E^2}{\partial D} = 0$ (3)

yields

$$\begin{pmatrix} \sum_{i=1}^{N} x_i^{\;2} & \sum_{i=1}^{N} x_i y_i & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i y_i & \sum_{i=1}^{N} y_i^{\;2} & \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} y_i & n \end{pmatrix} \begin{pmatrix} A \\ B \\ D \end{pmatrix} = \begin{pmatrix} -\sum_{i=1}^{N} x_i z_i \\ -\sum_{i=1}^{N} y_i z_i \\ -\sum_{i=1}^{N} z_i \end{pmatrix}, \tag{4}$$

from which A, B, D, hence a, b, c, d which define the plane can be computed.

The grid size of 0.9 Å and the grouping parameter of 2.0 Å were determined in our previous work which used the same tertiary grid to represent a protein structure in voxels for protein docking prediction[67,68].

Benchmark data sets for pocket identification

Because a ligand binding site of an enzyme usually locates at a pocket or a hollow of the enzyme, finding a geometrical pocket and hollow itself can be used as a prediction method of binding sites of enzymes. Recent studies show that combining with the other information, such as sequence conservation, hydrophobicity or charge distribution can further improve the accuracy of the prediction [69,70]. Although the purpose of VisGrid is not only identifying pockets and hollows but also protrusions and flat regions, as a way of demonstrating usefulness of VisGrid, here we applied it for prediction of ligand binding sites. Note that benchmarking performance of protrusion or flat region identification is not possible because there aren't a golden standard dataset of protrusion and flat regions of proteins. Primarily we used Liganded-Pocket (L-P) Set[71] as the benchmark database, which consists of 5,616 protein structures taken from Protein Data Bank (PDB)[33]. L-P set includes naturally occurring hetero molecules of a reasonable size. Redundant pairs of a protein and a ligand were removed, but the original L-P set does include identical proteins with different ligands, because pocket shapes are different when different ligands are bound. To eliminate potential bias of homologous proteins in the benchmark results, we have further tested VisGrid on two additional datasets created from the L-P set, where redundant entries in terms of sequence and structure are removed. In the first set, by referring a representative protein set created by PDB-RERPDB server[72] Rel.#2005_05_29 (essentially proteins with more than 30% sequence identity and a root mean square deviation (RMSD) of less than 10Å are grouped in a family), only one representative protein in each protein family are chosen from the L-P set. This set is named L-P-reprdb set, which contains 1139 proteins. In the second set, one protein from each topology of proteins defined by CATH database[73] version 2.6.0 is chosen from the L-P set. This set is named L-P-cath set, which contains 273 proteins. In addition to these benchmark tests on the the L-P set, we have tested VisGrid on ligand unbound structures to examine the performance in a more realistic situation of predicting ligand binding site in proteins. Performance of VisGrid on datasets of unbound proteins are compared with four other existing programs. See the section entitled "Comparison with the other existing methods" below.

Moreover we have further tested VisGrid on distorted structures of 164 proteins in the L-P set. The purpose of this experiment is to mimic the situation of predicting ligand binding sites in computationally predicted protein structures. Considering the recent situation that protein structure prediction methods, especially template-based structure prediction methods[74,75,76], have improved to make reasonable models in many cases, it is of great interest to investigate how robustly VisGrid performs on predicted structures whose RMSD are in a realistic range of "successful" predictions. To generate hypothetical predicted protein structures which are distorted from its native structure but still maintain "protein-like" stereochemical characteristics, we used a MD simulation package, NAMD[65]. For each of the 164 proteins, a series of structures are generated with an RMSD to the original crystal structure of 0.5 Å (148), 1.0 Å (159), 1.5 Å (163), 2.0 Å (156), 2.5 Å (113), and 3.0 Å (45). These distorted structures and files of input parameters used in this calculation are found at http://dragon.bio.purdue.edu/visgrid_suppl/. In the parentheses, the number of available proteins among the 164 proteins in the RMSD range is shown. Some of the proteins were not available for a certain RMSD range because there wasn't a structure within that range in the simulation trajectories generated. Especially, only 45 proteins were available for the range of 3.0 Å because about three quarters of the simulated proteins did not distort that far in the simulations of about 24 hours at the temperature of 310K with implicit water. Ligand molecules are removed from protein structures when molecular dynamics simulation is performed. Note here that this benchmark on the dataset of distorted structures is much more challenging than benchmark of unbound proteins, because in most of the cases the RMSD between an unbound structure to its bound form is around 1 Å or less. In the two datasets of pairs of bound and unbound proteins, the average RMSD of pairs are 0.51 and 0.52 Å (see below).

Definition of accurate prediction of binding sites

In the benchmark data sets above, the actual binding site residues are defined as all the residues which are closer than 4.5 Å to the ligand. The distance between a residue to a ligand is defined as the minimum distance between all the possible pairs of heavy atoms of the residue and the ligand. The accuracy of predictions is computed in both residue-based and the binding site-based. The sensitivity and the specificity of the residue-based accuracy are computed as follows:

$$Sensitivity = {}^{TP}\!\big/_{TP+FN} \tag{1}$$

$$Specificity = {}^{TP}\!\big/_{TP+FP} \tag{2}$$

where TP is the number of actual binding site residues which are correctly predicted to be binding site residues; FN is the number of actual binding site residues which are missed in a binding site residue prediction; FP is the number of residues not included in ligand binding sites of a protein but wrongly predicted to be binding site residues.

A predicted binding site is considered to be correct if a certain fraction of the residues involved in the actual binding site are covered in the prediction (Fig. 5).

Comparison with the other existing methods

The performance of binding site prediction by VisGrid is compared with four other programs, LIGSITE[56], SURFNET[58], CAST[54], and PASS[77]. PASS rolls a probe sphere on a protein surface and identifies pockets as regions where probes have higher number of contacts with atoms. The algorithms of LIGSITE, SURFNET, and CAST are briefly mentioned above in Introduction. The LIGSITE source code was downloaded from http://scoppi.biotec.tu-dresden.de/pocket/download.html. The SURFNET source code was obtained from http://www.biochem.ucl.ac.uk/~roman/surfnet/surfnet.html. We used the CAST web server at http://sts.bioengr.uic.edu/castp/calculation.php.

We have chosen the four methods for comparison because all of them use solely geometry of protein surface as input information. However, a fair comparison is still not trivial because these methods use different algorithms and were evaluated (thus probably tuned) in different criteria in their original papers. Considering this difficulty, the comparison is performed in the following two stages. First, because CAST also predicts pockets defined by binding residues in the pockets, primarily VisGrid is compared with CAST (Table 2, Fig. 9). A predicted binding site by CAST and VisGrid is considered to be correct if 50% or more of the residues in the predicted binding site overlap with the residues of the actual binding site. Second, to compare all the four methods with VisGrid, we follow the approach taken by Huang & Schroeder[78]. Essentially the performance of an algorithm on a test protein is evaluated by the distance between the center of mass of the binding ligand and a point specified by each method is computed (Table 3, Fig. 10). For LIGSITE, the point is the geometric center of the pocket site's grid points. PASS predicts an active site point. SURFNET outputs each detected gap region represented by an ATOM record, which we use as the point. As for VisGrid and CAST, we computed the center of mass of atoms of a predicted binding site as the point to be used for evaluation. Note that this is not a fair comparison especially for CAST and VisGrid, but we took this approach because LIGSITE, SURFNET, and PASS don't predict binding residues, but a point in an empty space of a pocket. If the distance between the center of the binding ligand and the point representing a predicted binding site is equal to or less than 4.0 Å, that prediction is considered to be correct. In contrast to the residue-based and the binding site-based accuracy introduced above, we call this accuracy metric the ligand center-based accuracy.

We used three datasets for this comparison. The first dataset is 48 pairs of ligand bound and unbound proteins taken from Table 4 in the paper by Huang & Schroeder[78]. The average RMSD between a pair of bound and unbound proteins is 0.52 Å computed by the CE program[79]. The second data set used is 86 pairs of bound and unbound proteins. These pairs are taken from the L-P set mentioned above and its associated dataset called the U-P set which is the counterparts of unbound proteins to the bound proteins in the L-P set[71]. The average RMSD between a bound and unbound protein pair is 0.51 Å. The list of proteins in these two datasets can be found at our website, http://dragon.bio.purdue.edu/visgrid_suppl/. In addition to the two sets of bound and unbound proteins, we used the data set of the distorted structures described above as well. For an unbound protein and a distorted protein structure, the (hypothetical) ligand center to be predicted is determined by superimposing its counterpart bound protein, transferring the ligand center of the bound protein to the unbound or to the distorted structure. Ligand binding residues of an unbound protein or a distorted protein is assumed to be the same as its counterpart protein with the bound ligand.

Results

Identified pockets, protrusions and flat regions

Figure 2 shows examples of identified pocket regions, protrusions and flat regions by VisGrid. Fig. 2A is HSP90, which has an adenosine triphosphate (ATP) binding pocket. All the ligand binding residues are correctly identified. A large protrusion on the left of Fig. 2A1 is the C-terminal tail which stretches outwards. Another protrusion locating on the right side of the figure is a loop region between two helices. The other small protrusions in the protein are side-chains which point outwards. The identified flat region is formed by a side of a kinked small helix. Fig. 2B is argininosuccinate synthase, another enzyme with an ATP binding site. The large protrusion is the C-terminal long helix which flips out from the mass. The flat region is formed by a kinked helix of ten residue long. 18 ATP binding residues among 19 total are correctly identified. Fig. 2C is 3-α-hydroxysteroid dehydrogenase, which has a large pocket for nicotinamide adenine dinucleotide (NAD). 24 among 27 ligand binding residues are correctly identified in this structure. Protrusions identified are parts of helices which lie at the entrance of the NAD binding pocket. The flat region on the left side in the figure consists of two parallel helices. Fig 2D is a dehydrogenase, which binds NAD. The identified large protrusion is the C-terminal tail of the protein. 31 among 34 NAD binding residues are correctly identified. The flat region, which is on the opposite side of the pocket, consists of ten residues at the end of two helices and a strand. Fig. 2F is a dehydrogenase, which has a large NAD binding site. 24 out of 25 NAD binding residues are correctly identified in the binding pocket. The identified large protrusion next to the NAD binding pocket is a long loop region which hangs over the pocket. The flat region is formed by a helix-turn-helix of eleven residues. Fig 2F is ferredoxin NADP+ reductase, which binds flavin adenine dinucleotide (FAD). The two large protrusions are loops which stick out and hold FAD. The most flat region of this protein surface locates at the back side of the FAD binding pocket, consisting of one side of a helix, a part of a strand, and a loop region.

To summarize this section, firstly, ligand binding residues are very well predicted as ones in pockets by VisGrid in these examples. Moreover, a large characteristic protrusion in each protein structure was identified, successfully ignoring many small bumps on the surface. The flat regions of the size of radius of 10 Å identified here consist of one side of helices and a part of a loop or strand. Although computing an "accuracy" of identified protrusions and flat regions is not possible in the same way as we do for predicted ligand binding sites, the protrusions and flat regions shown in Figure 2 would be intuitive and reasonable by visual inspection. Combined with other properties such as sequence conservation, charges, and hydrophobicity, local geometric features VisGrid can identify will be useful 3D local surface motifs of protein families.

Identification of ligand binding sites: residue-based accuracy

We have tested if VisGrid can detect ligand binding sites of proteins as pockets or hollows, because ligand binding pockets are where geometric aspect of protein surface shape and protein function is directly connected. A practical convenience of this exercise is that the accuracy can be well defined as the fraction of correctly identified ligand binding residues and the fraction of the correctly predicted binding sites in a dataset. First, we examined how the sensitivity and the specificity of predicting ligand binding residues changes as the threshold value of visibility changes. The sensitivity and the specificity are defined in the equations (1) and (2). 5,616 proteins in the L-P set are used. Increasing the visibility threshold makes the sensitivity higher (Fig. 3A) and the specificity lower (Fig. 3B), because in most of the cases a ligand binds to a pocket or a hollow, which have smaller visibility. Almost all the binding residues are captured when the visibility threshold of 0.35 or lower is used. While the sensitivity monotonically increases as the visibility threshold is raised, a peak is observed for the specificity at around the visibility of 0.08. This is partly because using a too small visibility threshold will detect small local deep depressions which are not included in a large global pocket. It is shown that using 98 directions (filled symbols) improves the specificity over results using 26 directions (empty symbols). At the visibility threshold of 0.08, considering the top three largest pockets (and choose the one with the highest sensitivity; triangles) rather than just the largest pocket (circles) improves the specificity approximately by 0.08 and sensitivity by 0.07. The best specificity, 0.52, is achieved at the visibility criterion of 0.08 using 98 visible directions, when the top three largest pockets are taken into account. The sensitivity achieved with the same visibility threshold is 0.75. Figure 4 shows the ROC curve of predicting ligand binding residues. L-P-reprdb set and L-P-cath set gave almost the same results compared with the results on the original full L-P set.

Binding site-based accuracy

Next, the accuracy is computed based on the number of correctly identified binding sites in the benchmark set. In Table 1, a detected pocket as a predicted binding site is counted as correct if it overlaps more than 50% of the actual ligand binding residues. But because the accuracy depends on the criterion of the overlap, we show the accuracy computed with different overlap thresholds (Fig. 5). Here one of the three detected largest pockets which gives the largest overlap to the actual binding site is used to compute the accuracy. Consistent with the results of residue-based accuracy (Fig. 4), considering top 0.8% voxels with the smallest visibility gives the best accuracy (sensitivity), then using top 150 smallest visible voxels follows. With the 50% overlap threshold, the accuracy is 0.95 (by considering 0.8% top smallest visible voxels). The accuracy further increases almost to 1.0 if the overlap threshold is relaxed to 0.3 or less. On the other hand, the accuracy drops rapidly when the threshold is raised to 0.8 or higher. This plot gives important implication in prediction of ligand binding sites: First, almost all (97.2%) of the actual binding site is overlapped by more than 30% with one of the three largest pockets identified by VisGrid. In other words, geometrical characteristic alone (pockets and hollows) can significantly restrict search space without taking a risk of missing the correct binding site. On the other hand, approximately only in half of the cases a pocket out of three largest identified has more than 90% overlap to the actual site. Therefore some additional information, such as residue conservation, charge distribution etc. should be combined to further specify or adjust the position of ligand binding sites.

Table 1 shows the accuracy considering top 1, 3, 5, 10 largest pockets and all voxels with a small visibility. It is shown that considering top 3 largest pockets improves the accuracy significantly (6.5 – 9.3 %) compared to the results of considering only the largest pocket. However, considering further more pockets, e.g. top 5 or top 10 gave only a marginal gain in the accuracy. These results indicate that most of ligand binding sites of a protein locate at one of the three largest pockets.

Figure 6 is given to clarify that the size of pockets VisGrid identifies is almost in the same size as actual binding sites, which occupy approximately 5% of the total protein surface. Even when three largest pockets are considered for prediction, on average they cover only 15% of the whole surface area of the target protein.

Binding site identification on distorted structures

A further challenging experiment is conducted on predicting binding sites in distorted protein structures by MD simulation (Fig. 7). This is to test robustness of the algorithm to predicted protein structures which usually have unavoidable errors. Figure 7 shows that the specificity stays almost the same but the sensitivity shows approximately a decrease of 0.06 on distorted structures of an RMSD of 1.0 Å, and a decrease of 0.21 on structures of an RMSD of 2.0 Å. These results imply that some binding sites changed their shape as the global structure is distorted and lost a large volume of their cavity, so that VisGrid could not identify them. However, the consistent specificity indicates that false positive predictions will not increase on distorted structures. Here again employing additional physicochemical or sequence characteristics of binding sites will help retaining the sensitivity.

Comparison with the other existing methods

The performance of VisGrid was compared with four existing algorithms, CAST, PASS, SURNET and LIGSITE. First we compared VisGrid with CAST, because both of them predict ligand binding residues so that the comparison can be more straightforward. Table 2A and 2B show the results of the binding site-level accuracy on two datasets of bound and unbound protein pairs. In the same way as Table 1, three methods of selecting voxels with a low visibility are used for VisGrid to detect pockets. Among the three methods, consistent to Table 1, VisGrid performed best when the top 0.8% smallest visible voxels are used (Table 2A, 2B). Overall, VisGrid performed better than CAST in these two datasets in terms of the binding site-based accuracy. In Figure 8, VisGrid and CAST are further compared in the distorted protein structure set. Again in this dataset, VisGrid performed better than CAST. The performance of CAST starts to deteriorate when the RMSD of the distorted structure grows larger than 1.5 Å. In contrast, VisGrid showed more stable performance irrelevant to the RMSD of the distorted structure in this benchmark test. Note here that Figure 7 shows the prediction results of VisGrid in terms of the binding residue-based accuracy, while Figure 8 shows the binding-site based accuracy on the same dataset. Comparing Figures 7 and 8, it can be concluded that although the less binding residues are correctly predicted by VisGrid as structures are more distorted (Fig. 7), it didn’t affect much to the binding-site level accuracy (Fig. 8).

Next, VisGrid is compared with CAST, LIGSITE, PASS, and SURNET in terms of the ligand center-based accuracy. For this comparison, we had to use the ligand center-based accuracy because LIGSITE, PASS, and SURFNET predict the coordinates of the center of pockets but not binding residues. In the data set of 48 bound and unbound protein pairs (Table 3A), VisGrid showed the second best performance in Top1 prediction of bound and unbound proteins. In the data set of 86 pairs of bound and unbound proteins (Table 3B), the rank of VisGrid was the fourth in the top1 prediction for the bound and unbound proteins. Here note that actual outputs of the predictions by CAST and VisGrid examined in Table 2 and 3 are the same, but the apparent number of correct predictions in the two tables for CAST and VisGrid differs because different metrics of the accuracy are used. This illustrates difficulty of comparing performance of programs in a fair fashion. When the five programs are compared on the data set of the distorted structures in terms of the ligand center-based accuracy (Fig. 9), VisGrid clearly outperformed PASS, SURFNET, and CAST. To conclude this section, although a fair comparison and accurate ranking of the programs are not possible and not the aim of this study, VisGrid showed at least comparable performance with the others compared, and showed superior performance to several of the methods. Especially VisGrid showed the top performance in the distorted structure data set.

Examples of predicted binding sites

Four illustrative examples of predicted ligand binding sites are shown in Figure 10. Fig. 10A is a successful prediction of binding sites in 1ra8. A large pocket in the center of the protein, which includes two binding sites of folate and 2-monophosphoadenosine 5'-diphosphoribose are well detected with a sensitivity of 0.86 and a specificity of 0.71. Fig. 10B is a successful prediction of a binding hollow embedded inside of a protein. The ligand, dilinoleoylphosphatidylcholine is wrapped by a curled β-sheet and two helices of phosphatidylcholine transfer protein (1ln1), and almost invisible from outside. The protein in Fig. 10C is the same as shown in Fig. 10A (1ra8), but distorted by MD simulation to an RMSD of 2.5 Å. The binding pocket of 2-monophosphoadenosine 5'-diphosphoribose is deformed and lost its volume by the MD simulation, but the folate binding site is still detected. Fig. 10D is an example which VisGrid failed to detect a target binding site. Dihydrogenphosphate binds to a shallow surface of phosphate binding periplasmic protein, which is not included among the top three largest pockets in the protein.

Discussion

We have introduced the VisGrid algorithm, which identifies geometric features of protein surfaces using an intuitive concept of the visibility. Unlike existing programs which only focus on identification of ligand binding pockets in proteins, VisGrid also identifies large protrusions or flat regions which characterize a protein surface, i.e. annotation of protein surfaces. Characterizing a protein surface by VisGrid is important for function annotation of proteins from the tertiary structure. Finding characteristic local sites in a protein surface is analogous to protein sequence analyses routinely used for function annotation of protein sequences by identifying various features in sequences, such as local sequence motifs [19,16,80,17], conserved regions, and correlated mutation sites[81]. We particularly emphasize here the importance of identification of protrusions and flat regions: besides predicting ligand binding pockets, of great interest in protein bioinformatics is predicting protein-protein interaction interface of a protein and protein docking. To predict protein interaction interface, identifying flat regions is essential information of geometric features of local regions that can be combined with information of physicochemical properties and residue conservation, because it was repeatedly reported that protein interaction sites are relatively flat and the other properties differ according to the nature of the protein complex[82,83]. Figure 11 shows examples of two protein interaction sites which include the most flat regions of the protein surfaces. On the other hand, in a protein docking prediction where the aim is to predict a docked conformation of the two input protein structures, shape complimentarity of local regions of the two proteins gives the essential information to guide the search of the conformation space[84,85]. In our previous works[68,67], pre-identifying complementary regions, such as pairs of a small pocket in one protein and a protrusion in another, can reduce the search space thus the computational time on average 50% without sacrificing the accuracy. There identification of local protrusions is necessary.

The importance of structure-based function prediction is increasing because the structural genomics projects are producing structure of proteins whose function cannot be predicted by conventional sequence-based methods. Function prediction by local structural signature is not only for totally uncharacterized proteins, but will be also useful for any protein, because it could find additional potential hidden function of the protein which is not investigated before. Function prediction from local surface signature involves the initial step of identifying characteristic sites of proteins and the subsequent step of comparison of that site to known sites in a database. In these steps robustness of the algorithm on finding same functional sites from different structures is essential. If an algorithm is sufficiently robust, the range of application will be larger because the algorithm can be also applied to predicted protein structures. Therefore we have benchmarked VisGrid on identifying ligand binding sites in distorted structures (Fig. 7). This kind of difficult by realistic test has not been done for existing pocket finding programs, as far as we are aware. We found that specificity of VisGrid does not drop even if structures are distorted up to 3.0 Å. This is important because it implies that VisGrid will have very low false positive in binding pocket identification even for structures of slightly different conformation in the range of natural flexibility.

Biology has entered the proteomics era when massive amount of data awaits interpretation. In order to take full advantage of expensive proteomics data, bioinformatics infrastructure of many different kinds should be developed. Compared to sequence-based methods, development of tools for structure analyses and annotation has lagged behind. Due to its intuitiveness and simplicity, the visibility criterion for characterizing local protein surface shapes implemented in VisGrid has a wide range of application in development of structure-based tools for protein annotation.

Acknowledgements

This work was supported in part by the National Institute of General Medical Sciences of the National Institutes of Health (R01 GM075004). DK also acknowledges funding from National Science Foundation (DMS 0604776) and NIH (U24 GM077905). The authors thank Preston Spratt for proofreading this manuscript.

Reference List

1. Joyce AR, Palsson BO. The model organism as a system: integrating 'omics' data sets. Nat Rev Mol Cell Biol 2006;7(3):198-210.

2. Kihara D, Yang DY, Hawkins T. Bioinformatics resources for cancer research with an emphasis on gene function and structure prediction tools. Cancer Informatics 2006;225-35.

3. Auerbach D, Thaminy S, Hottiger MO, Stagljar I. The post-genomic era of interactive proteomics: facts and perspectives. Proteomics 2002;2(6):611-623.

4. Schwikowski B, Uetz P, Fields S. A network of protein-protein interactions in yeast. Nat Biotechnol 2000;18(12):1257-1261.

5. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 2000;403(6770):623-627.

6. Bork P, Jensen LJ, von MC, Ramani AK, Lee I, Marcotte EM. Protein interaction networks from yeast to human. Curr Opin Struct Biol 2004;14(3):292-299.

7. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995;270(5235):467-470.

8. Hoheisel JD. Microarray technology: beyond transcript profiling and genotype analysis. Nat Rev Genet 2006;7(3):200-210.

9. Orchard S, Hermjakob H, Apweiler R. Annotating the human proteome. Mol Cell Proteomics 2005;4(4):435-440.

10. Hawkins T, Kihara D. Function prediction of uncharacterized proteins. J Bioinform Comput Biol 2007;5(1):1-30.

11. Koonin EV, Tatusov RL, Rudd KE. Protein sequence comparison at genome scale. Methods Enzymol 1996;266295-322.

12. Koonin EV, Aravind L, Kondrashov AS. The impact of comparative genomics on our understanding of evolution. Cell 2000;101(6):573-576.

13. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 1988;85(8):2444-8.

14. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215(3):403-10.

15. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25(17):3389-402.

16. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A. Pfam: clans, web tools and services. Nucleic Acids Res 2006;34(Database issue):D247-D251.

17. Bru C, Courcelle E, Carrere S, Beausse Y, Dalmar S, Kahn D. The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acids Res 2005;33(Database issue):D212-D215.

18. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, Bork P, Bucher P, Cerutti L, Copley R, Courcelle E, Das U, Durbin R, Fleischmann W, Gough J, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McDowall J, Mitchell A, Nikolskaya AN, Orchard S, Pagni M, Ponting CP, Quevillon E, Selengut J, Sigrist CJ, Silventoinen V, Studholme DJ, Vaughan R, Wu CH. InterPro, progress and status in 2005. Nucleic Acids Res 2005;33(Database issue):D201-D205.

19. Hulo N, Bairoch A, Bulliard V, Cerutti L, De CE, Langendijk-Genevaux PS, Pagni M, Sigrist CJ. The PROSITE database. Nucleic Acids Res 2006;34(Database issue):D227-D230.

20. Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P. SMART 5: domains in the context of genomes and networks. Nucleic Acids Res 2006;34(Database issue):D257-D260.

21. Snel B, Bork P, Huynen MA. The identification of functional modules from the genomic association of genes. Proc Natl Acad Sci U S A 2002;99(9):5890-5895.

22. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 1999;96(8):4285-4288.

23. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A 1999;96(6):2896-2901.

24. Martin DM, Berriman M, Barton GJ. GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. BMC Bioinformatics 2004;5178-

25. Khan S, Situ G, Decker K, Schmidt CJ. GoFigure: automated Gene Ontology annotation. Bioinformatics 2003;19(18):2484-2485.

26. Zehetner G. OntoBlast function: From sequence similarities directly to potential functional annotations by ontology terms. Nucleic Acids Res 2003;31(13):3799-3803.

27. Hawkins T, Luban S, Kihara D. Enhanced automated function prediction using distantly related sequences and contextual association by PFP. Protein Sci 2006;151550-1556.

28. Pearl F, Todd AE, Bray JE, Martin AC, Salamov AA, Suwa M, Swindells MB, Thornton JM, Orengo CA. Using the CATH domain database to assign structures and functions to the genome sequences. Biochem Soc Trans 2000;28(2):269-275.

29. Kinoshita K, Nakamura H. Protein informatics towards function identification. Curr Opin Struct Biol 2003;13(3):396-400.

30. Skolnick J, Fetrow JS, Kolinski A. Structural genomics and its importance for gene function analysis. Nat Biotechnol 2000;18(3):283-7.

31. Yokoyama S, Matsuo Y, Hirota H, Kigawa T, Shirouzu M, Kuroda Y, Kurumizaka H, Kawaguchi S, Ito Y, Shibata T, Kainosho M, Nishimura Y, Inoue Y, Kuramitsu S. Structural genomics projects in Japan. Prog Biophys Mol Biol 2000;73(5):363-376.

32. Teichmann SA, Murzin AG, Chothia C. Determination of protein function, evolution and interactions by structural genomics. Curr Opin Struct Biol 2001;11(3):354-363.

33. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28(1):235-42.

34. Saqi MA, Wild DL. Expectations from structural genomics revisited: an analysis of structural genomics targets. Am J Pharmacogenomics 2005;5(5):339-342.

35. Wilson CA, Kreychman J, Gerstein M. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J Mol Biol 2000;297(1):233-249.

36. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J 1986;5(4):823-826.

37. Kihara D, Skolnick J. Microbial Genomes have over 72% structure assignment by the threading algorithm PROSPECTOR_Q. Proteins 2004;55464-473.

38. Dietmann S, Holm L. Identification of homology in protein structure classification. Nat Struct Biol 2001;8(11):953-957.

39. Orengo CA, Jones DT, Thornton JM. Protein superfamilies and domain superfolds. Nature 1994;372(6507):631-4.

40. Thornton JM, Todd AE, Milburn D, Borkakoti N, Orengo CA. From structure to function: approaches and limitations. Nat Struct Biol 2000;7 Suppl991-994.

41. Polgar L. The catalytic triad of serine peptidases. Cell Mol Life Sci 2005;62(19-20):2161-2172.

42. Fetrow JS, Godzik A, Skolnick J. Functional analysis of the Escherichia coli genome using the sequence- to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. J Mol Biol 1998;282(4):703-11.

43. Wallace AC, Borkakoti N, Thornton JM. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. Protein Sci 1997;6(11):2308-2323.

44. Mooney SD, Liang MH, DeConde R, Altman RB. Structural characterization of proteins using residue environments. Proteins 2005;61(4):741-747.

45. Shulman-Peleg A, Nussinov R, Wolfson HJ. Recognition of functional sites in protein structures. J Mol Biol 2004;339(3):607-633.

46. Connolly ML. Solvent-accessible surfaces of proteins and nucleic acids. Science 1983;221(4612):709-713.

47. Goldman BB, Wipke WT. QSD quadratic shape descriptors. 2. Molecular docking using quadratic shape descriptors (QSDock). Proteins 2000;38(1):79-94.

48. Duncan BS, Olson AJ. Approximation and characterization of molecular surfaces. Biopolymers 1993;33(2):219-229.

49. Exner TE, Keil M, Brickmann J. Pattern recognition strategies for molecular surfaces. I. Pattern generation using fuzzy set theory. J Comput Chem 2002;23(12):1176-1187.

50. Kinoshita K, Nakamura H. Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. Protein Sci 2003;12(8):1589-1595.

51. Lichtarge O, Sowa ME. Evolutionary predictions of binding surfaces and interactions. Curr Opin Struct Biol 2002;12(1):21-27.

52. La D, Sutch B, Livesay DR. Predicting protein functional sites with phylogenetic motifs. Proteins 2005;58(2):309-320.

53. Armon A, Graur D, Ben-Tal N. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol 2001;307(1):447-463.

54. Liang J, Edelsbrunner H, Woodward C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci 1998;7(9):1884-1897.

55. Laskowski RA, Luscombe NM, Swindells MB, Thornton JM. Protein clefts in molecular recognition and function. Protein Sci 1996;5(12):2438-2452.

56. Hendlich M, Rippmann F, Barnickel G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph Model 1997;15(6):359-63, 389.

57. Levitt DG, Banaszak LJ. POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. J Mol Graph 1992;10(4):229-234.

58. Laskowski RA. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph 1995;13(5):323-328.

59. Wojciechowski M, Skolnick J. Docking of small ligands to low-resolution and theoretically predicted receptor structures. J Comput Chem 2002;23(1):189-97.

60. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. J Mol Biol 1997;267(3):727-748.

61. Morris GM, Goodsell DS, Huey R, Olson AJ. Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. J Comput Aided Mol Des 1996;10(4):293-304.

62. Ewing TJ, Makino S, Skillman AG, Kuntz ID. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. J Comput Aided Mol Des 2001;15(5):411-428.

63. Kramer B, Rarey M, Lengauer T. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. Proteins 1999;37(2):228-241.

64. Ogmen U, Keskin O, Aytuna AS, Nussinov R, Gursoy A. PRISM: protein interactions by structural matching. Nucleic Acids Res 2005;33(Web Server issue):W331-W336.

65. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. J Mol Graph 1996;14(1):33-38.

66. Richmond TJ. Solvent accessible surface area and excluded volume in proteins. Analytical equations for overlapping spheres and implications for the hydrophobic effect. J Mol Biol 1984;178(1):63-89.

67. Turuvekere S. Geometric algorithms and methods for binding site identification and alignment of proteins. Master thesis, Department of Mechanical Engineering, Purdue University 2004;

68. Turuvekere S, Agrawal M, Kihara D, Ramani K. Feature Recognition Based Identification of Potential Binding Sites on the Molecular Surfaces. The Protein Society 18th Symposium 2004;

69. Rossi A, Marti-Renom MA, Sali A. Localization of binding sites in protein structures by optimization of a composite scoring function. Protein Sci 2006;15(10):2366-2380.

70. Laskowski RA, Watson JD, Thornton JM. ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res 2005;33(Web Server issue):W89-W93.

71. An J, Totrov M, Abagyan R. Pocketome via comprehensive identification and classification of ligand binding envelopes. Mol Cell Proteomics 2005;4(6):752-761.

72. Noguchi T, Akiyama Y. PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003. Nucleic Acids Res 2003;31(1):492-493.

73. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH--a hierarchic classification of protein domain structures. Structure 1997;5(8):1093-108.

74. Kolinski A, Betancourt MR, Kihara D, Rotkiewicz P, Skolnick J. Generalized comparative modeling (GENECOMP): A combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement. Proteins 2001;44(2):133-49.

75. Skolnick J, Kihara D. Defrosting the frozen approximation: PROSPECTOR--a new approach to threading. Proteins 2001;42(3):319-31.

76. Skolnick J, Kihara D, Zhang Y. Development and large scale benchmark testing of the PROSPECTOR 3.0 threading algorithm. Proteins 2004;56(3):502-518.

77. Brady GP, Jr., Stouten PF. Fast prediction and visualization of protein binding pockets with PASS. J Comput Aided Mol Des 2000;14(4):383-401.

78. Huang B, Schroeder M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. BMC Struct Biol 2006;619-

79. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11(9):739-47.

80. Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P, Uddin A, Zygouri C. PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res 2003;31(1):400-402.

81. Gobel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins 1994;18(4):309-17.

82. Burgoyne NJ, Jackson RM. Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces. Bioinformatics 2006;22(11):1335-1342.

83. Jones S, Thornton JM. Prediction of protein-protein interaction sites using patch analysis. J Mol Biol 1997;272(1):133-143.

84. Chen R, Weng Z. A novel shape complementarity scoring function for protein-protein docking. Proteins 2003;51(3):397-408.

85. Chen R, Li L, Weng Z. ZDOCK: an initial-stage protein-docking algorithm. Proteins 2003;52(1):80-87.

Table 1. The binding site level accuracy[a)].

	Top 1	Top 3 [b)]	Top 5	Top 10	All
Visibility ≤ 8/98	77.0(%)	83.5	84.3	84.6	88.9
Top 150 Voxels	80.7	88.1	88.8	89.1	93.4
Top 0.8% Voxels	85.7	95.0	95.8	96.1	98.4

The L-P set is used for this test.

a) A predicted binding pocket is considered to be correct if it overlaps more than 50% of the actual binding site.

b) A prediction for a protein is considered to be correct if one of the three largest pockets identified by VisGrid overlaps with the actual binding site by more than 50%.

Table 2. Performance comparison with CAST in terms of the binding site level accuracy.

A. The number of correctly prediction in the dataset of 48 ligand bound and unbound protein structures. [a)]

	48 Bound structures		48 Unbound structures	
	Top 1	Top 3 [b)]	Top1	Top3
CAST	32 (66.7%)	42 (87.5%)	34 (70.8%)	40 (83.3%)
Visibility ≤ 8/98	40 (83.3%)	46 (95.8%)	46	48 (100%)
Top 150 Voxels	46	48	46	48
Top 0.8% Voxels	45 (93.8%)	48	46	48

a) The 48 pairs of bound and unbound structures are taken from Table 4 of Huang & Scheroeder[78].

b) A prediction for a protein is considered to be correct if one of the three largest pockets identified by VisGrid overlaps with the actual binding site by more than 50%.

B. The dataset of 86 ligand bound and unbound protein structures from the L-P & U-P sets.

	86 Bound structures		86 Unbound structures	
	Top 1	Top 3 [b)]	Top1	Top3
CAST	65 (75.6%)	73 (84.9%)	56 (65.1%)	68 (79.1%)
Visibility ≤ 8/98	71 (82.6%)	81 (94.2%)	67 (77.9%)	82 (95.3%)
Top 150 Voxels	76 (88.4%)	86 (100%)	77 (89.5%)	85 (98.8%)
Top 0.8% Voxels	80 (93.0%)	86	76	86

Table 3. Performance comparison with CAST, LIGSITE, PASS, SURFNET in terms of the ligand center-based accuracy.

A. The number of correct prediction in the 48 ligand bound/unbound protein pairs.

	48 Bound structures		48 Unbound structures	
	Top 1	Top 3	Top1	Top3
CAST	27 (65.9%)	38 (79.2%)	31 (64.6%)	37 (77.1%)
LIGSITE	40 (83.3%)	44 (91.7%)	36 (75.0%)	38 (79.2%)
PASS	32 (66.7%)	42 (87.5%)	27 (56.3%)	34 (70.8%)
SURFNET	23 (47.9%)	34 (70.8%)	19 (39.6%)	29 (60.4%)
VisGrid: Top 0.8% Voxels	32 (66.7%)	38 (79.2%)	34 (70.8%)	41 (85.4%)

B. Results of the 86 ligand bound/unbound structure pairs from the L-P & U-P sets.

	86 Bound structures		86 Unbound structures	
	Top 1	Top 3	Top1	Top3
CAST	66 (76.7%)	79 (91.9%)	66 (76.7%)	79 (91.9%)
LIGSITE	65 (75.6%)	75 (87.2%)	69 (80.2%)	77 (89.5%)
PASS	54 (62.8%)	71 (82.6%)	54 (62.8%)	71 (82.6%)
SURFNET	63 (73.3%)	77 (89.5%)	63 (73.3%)	77 (89.5%)
VisGrid: Top 0.8% Voxels	61 (70.9%)	66 (76.7%)	55 (64.0%)	63 (73.3%)

Figure Legends

Figure 1. The visibility of a voxel. The visibility of a target voxel is defined as the number of visible directions from the voxel. There are 26 directions from the target voxel when one surrounding layer is considered. Visible directions from the target voxel (in black) are 9, 2, 18 in **A, B, C,** respectively. A direction is considered to be visible if consecutive n (n=20) voxels is not filled by the protein in that direction.

Figure 2. Examples of identified pockets, protrusions and flat regions in protein surfaces. Atoms in color are: blue, those identified as pockets; red, protrusions; green, flat regions. A residue is identified to be in a pocket if its visibility is among the smallest top 8%. 98 directions from a voxel are considered to define the visibility. Atoms in a protrusion region are ones which have the inverse visibility of 2/98 or less (*i.e.* an identified pocket in the negative image). Atoms in green in each protein are the most flat group of surface residues which fit in a sphere of a radius of 10 Å. **A**, HSP90 molecular chaperone (PDB code: 1am1) which has a bound ADP in the crystal structure. A1 and A2 are views from different directions. All the 18 residues which bind to ADP in the binding pocket are correctly identified. The average distance from atoms in the identified flat region to the fitted plane, d_f, is 1.31 Å. **B**, argininosuccinate synthetase which binds ATP (1kp2). 18 among 19 ATP binding residues are identified as pocket regions. d_f = 1.47 Å. **C**, 3-alpha-hydroxysteroid dehydrogenase, which binds NAD (1fk8). 24 among 27 NAD binding residues are identified as pocket regions. d_f = 1.49 Å. **D**, cis-biphenyl-2,3-dihydrodiol-2,3-dehydrogenase, which binds NAD (1bdb). 31 among 34 NAD binding residues are identified as pocket regions. d_f = 1.49 Å. The flat region (shown in D2) locates at the back side of the pocket region (D1). **E**,

D-glyceraldehyde-3-phosphate dehydrogenase, which binds NAD (1gad). 24 out of 25 NAD binding residues are identified as pocket regions**.** d_f = 1.31 Å. E1 and E2 are views from different directions. **F**, ferredoxin-NADP+ reductase, which binds FAD (1e62). 15 out of 20 FAD binding residues are identified as pocket regions. d_f = 1.42 Å. Views from two different directions are shown (F1 and F2).

Figure 3. The sensitivity and the specificity of predicting ligand binding residues relative to the visibility threshold value used to identify pockets. The average value of **A:** the sensitivity and **B:** the specificity over all the proteins in the L-P set is plotted. Amino acid residues are predicted to be ligand binding residues if they are in pockets detected by using a visibility threshold value shown on the X-axis. Two visibility definitions are used: the fraction of visible directions among 98 directions in total (filled symbols with solid lines) or the fraction of visible directions out of 26 directions (empty symbols with dotted lines). Filled circle (●), residues in the largest pocket detected using the 98 visible directions are considered. Filled triangles (▼), residues in one of the top three largest pockets which has the largest overlap with the actual binding pocket are considered. The 98 directions are used to compute the visibility. Filled squares (■), residues in all the detected pockets using the 98 directions are considered. Empty circles (○), residues in the largest pocket detected using the 26 directions is considered. Empty triangles (▽), residues in one of the top three largest pockets with the largest overlap with the actual binding pocket are considered. The 26 directions are used to compute the visibility. Empty squares (□), residues in all the pockets detected using the 26 directions are considered.

Figure 4. The ROC curve of the ligand binding residue prediction. Three datasets are used: black circles, the full L-P set; gray circles, the L-P-reprdb set; empty circles, L-P-cath set. The visibility threshold to detect ligand binding pockets on a protein surface is varied to have a series of the false positive rate and the sensitivity of ligand binding residue prediction. For each visibility threshold, three largest pockets are detected. Residues in one of the three detected pockets which have the largest overlap to the actual binding site are predicted to be ligand binding residues.

Figure 5. Fraction of correctly predicted binding sites relative to the threshold value to define a predicted pocket is correct. If one of the three largest identified pockets overlaps with the correct binding sites by more than the value on the X-axis, that predicted pocket is considered to be correct. The L-E set is used. Solid line, voxels with a visibility of 8/98 or less are used to identify the largest pocket; dashed line, top 150 voxels with the smallest visibility are used; dotted line, voxels of the top 0.8% smallest visibility are used. Only the largest pocket in a protein is considered.

Figure 6. Distribution of fraction of the surface area of predicted pockets relative to the whole surface area of the protein. The number of voxels is used to count a surface area. Proteins in the L-P set are examined. Filled circle, distribution of the actual binding sites of the proteins; empty circle, the largest pockets identified in a protein by using voxels with the visibility of 8/98 or less; filled triangles, the largest pockets identified by using top 150 less visible voxels; empty triangles, the largest pockets identified by using the top 0.8% voxels with the smallest visibility.

Figure 7. The sensitivity and the specificity of predicted binding residues computed on the distorted structures of the L-P set. The X-axis shows the RMSD of the distorted structures to the native structures. **A**, the sensitivity; **B**, the specificity. Circle, the prediction using voxels with the visibility of 8/98 or less; Triangles, the prediction using top 150 voxels with the smallest visibility; Squares, the prediction using the top 0.8% voxels with the smallest visibility.

Figure 8. Performance comparison of VisGrid and CAST on the dataset of distorted structures. The dataset used is the same as the one used for Figure 7. The binding site level accuracy is computed for VisGrid in the three different methods: to use all the voxels with the visibility of 8/98 or less to identify the largest pockets, to use top 150 voxels with the smallest visibility, and to use the top 0.8% voxels with the smallest visibility. A predicted binding site of a protein is considered to be correct if it overlaps more than 50% of the residues of the actual ligand binding site of the protein. **A**, the largest pocket identified is considered; **B**, Among the three detected largest pockets, the best one which has the largest overlap to the actual binding site is considered.

Figure 9. Performance comparison with CAST, LIGSITE, PASS, and SURFNET on the dataset of distorted structures. As for VisGrid, only the results of using the top 0.8% voxels with the smallest visibility is shown. The performance was evaluated in terms of the ligand center-based accuracy. **A**, the top-scoring predicted ligand-center is considered; **B**, prediction made for a protein is considered to be correct if the actual ligand center is included as one of the three highest-scoring predicted ligand centers. The dataset used is essentially the same as

the one used for Figure 7 & 9, but the number of structures considered in the evaluation is smaller, because some of the programs crashed on several structures. The number of structures at each RMSD level on which all the programs produced meaningful outputs thus used in this evaluation is: 0 Å: 153 (164); 0.5 Å: 137 (148); 1.0 Å: 147 (159); 1.5 Å: 151 (163); 2.0 Å: 142 (156); 2.5 Å: 106 (113); 3.0 Å: 43 (45). In the parentheses, the number of the original dataset is shown.

Figure 10. Examples of binding site predictions. Bound ligands are shown in yellow and predicted binding sites are colored in red (, green and blue in **D**). **A,** dihydrofolate reductase complexed with folate and 2-monophosphoadenosine 5'-diphosphoribose (1ra8). 30 out of 35 ligand binding residues are correctly predicted among predicted 42 residues which constitute the largest pocket. Sensitivity: 30/35 = 0.86, specificity: 30/42 = 0.71. **B,** Phosphatidylcholine transfer protein complexed with dilinoleoylphosphatidylcholine (1ln1). An example of a ligand completely buried in the protein. Sensitivity: 32/33 = 0.97; specificity: 32/50 = 0.64. **C,** a distorted structure of 1ra8 to RMSD of 2.5 Å. The binding pocket for folate is still predicted in this distorted structure. Sensitivity: 18/35 = 0.51; specificity: 18/24 = 0.75. **D,** phosphate binding protein complexed with dihydrogenphosphate (1a54A). An example of wrong prediction, where none of the 25 binding residues are not predicted. Three largest predicted binding sites are shown in red, blue, and green.

Figure 11. Examples of flat protein-protein docking interface. The most flat surface of a radius of 7 Å (the same as Fig. 2) are shown in red, and the whole docking interface is colored in pink. **A**, acetylcholinesterase complexed with fasciculin-II (PDB: 1fss). The number of

residues in the docking interface of acetylcholinesterase is 25, among them 10 residues are included in the most flat region colored in red. The average distance from atoms in the flat region (red) to the fitted plane, d_f, is 0.86 Å and the d_f of the whole docking interface is 0.91 Å. **B**, Serratia metallo protease complexed with an inhibitor. The number of residues in the docking interface of the protease is 30, and among them, 9 residues are included in the most flat region colored in red. d_f of the flat region is 0.78 Å and the d_f of the whole docking interface is 0.71 Å.

Figure 1.

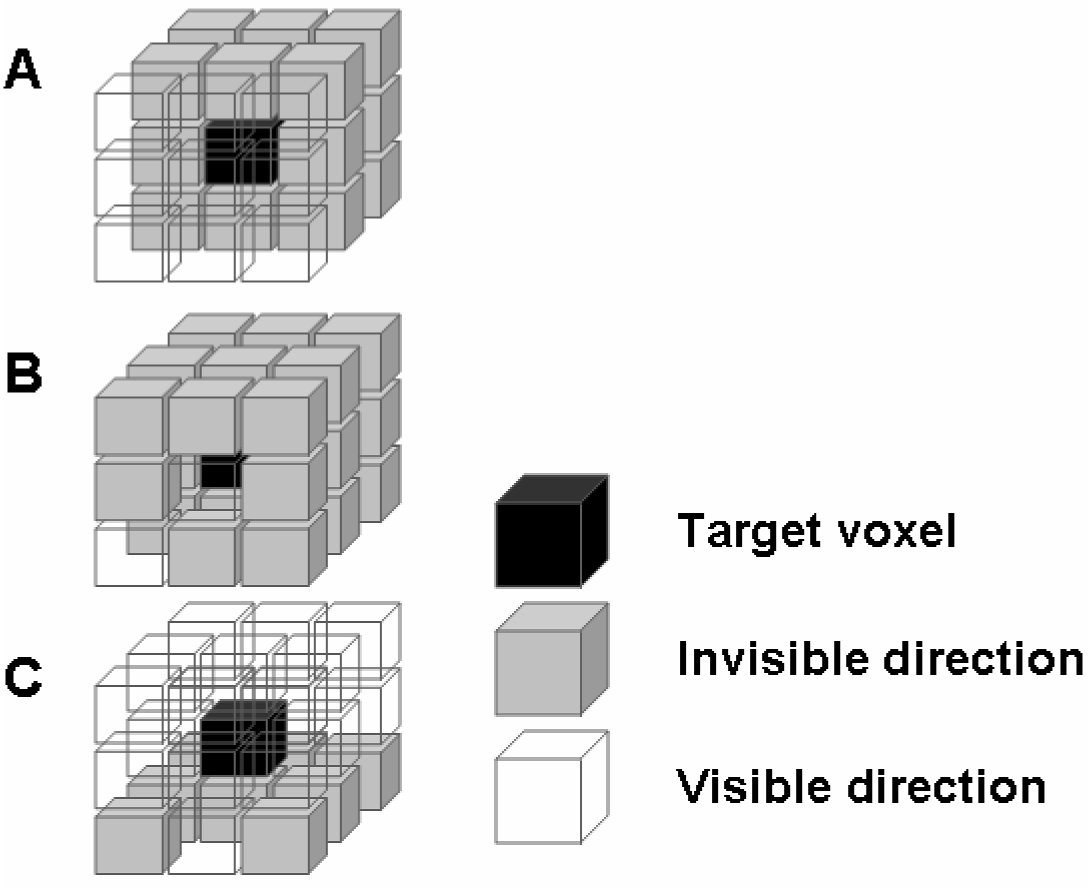


Figure 2

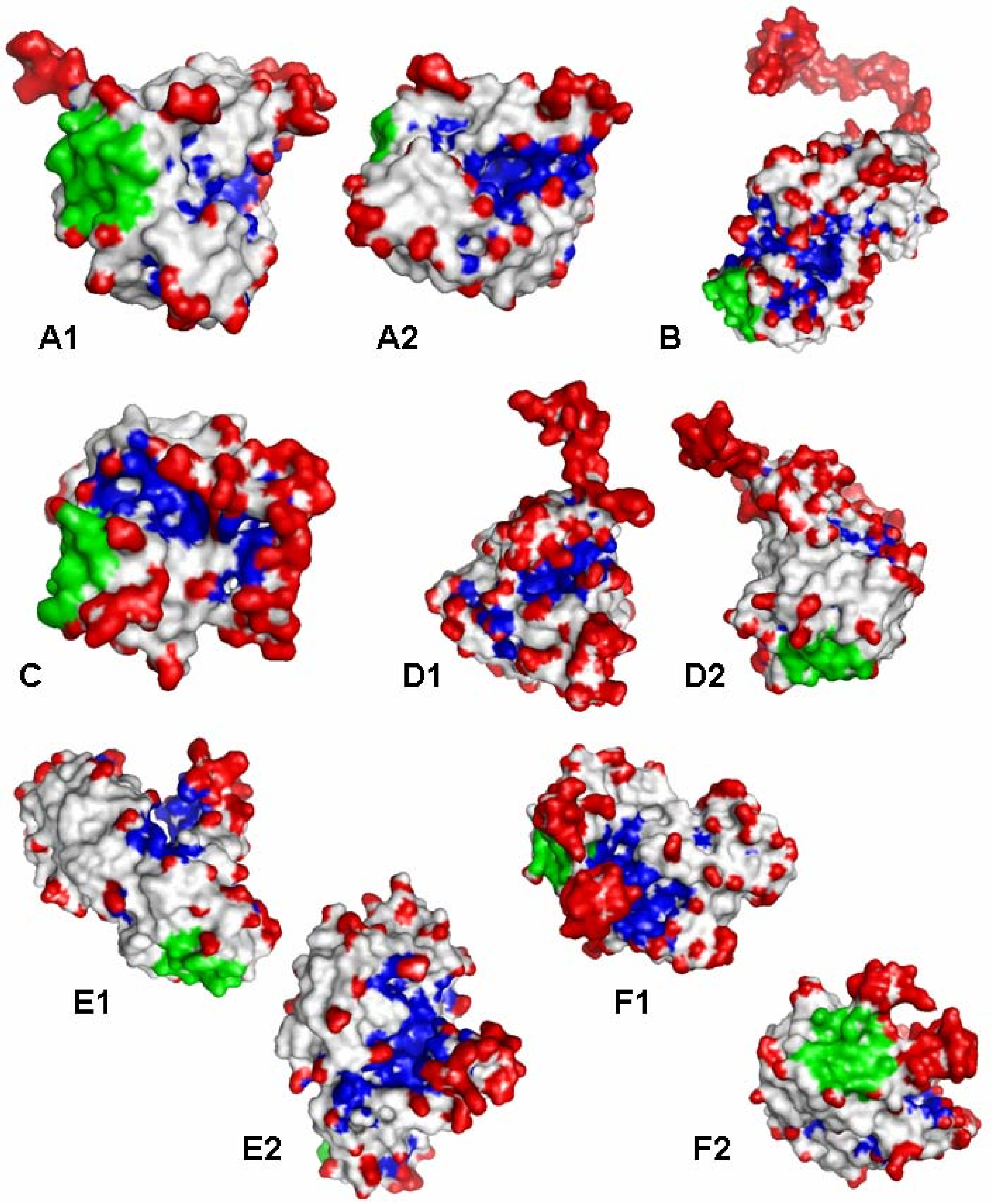
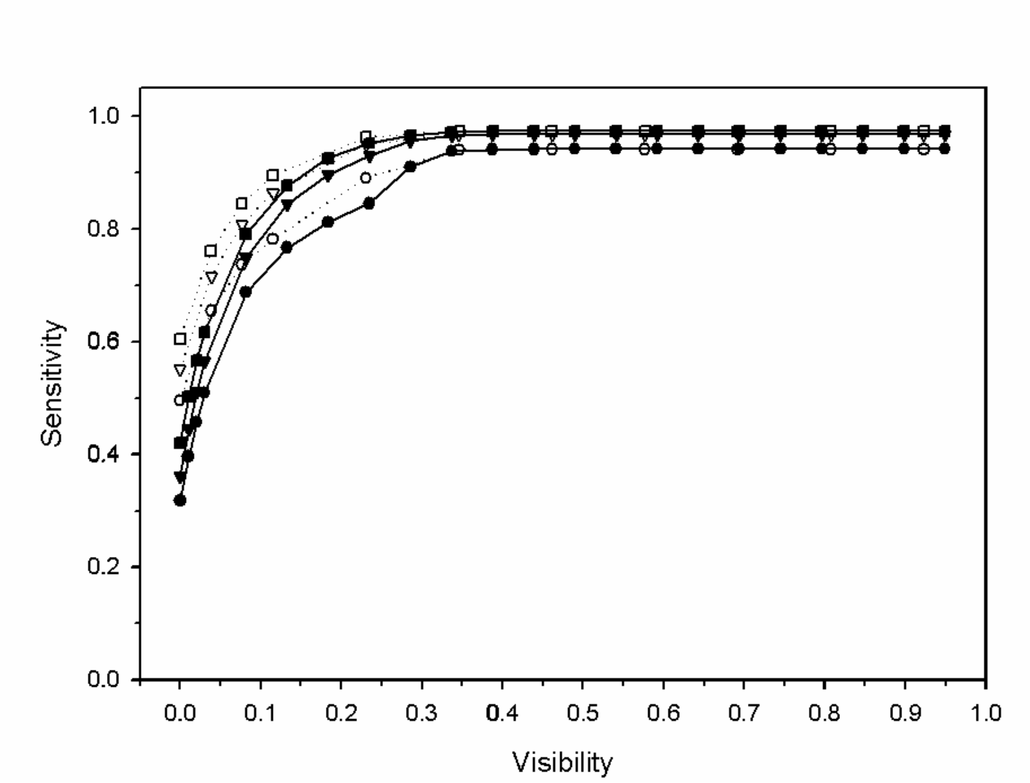


Figure 3
A.


1.0
0.8
0.6
0.4
0.2
0.0
Sensitivity
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Visibility

B.

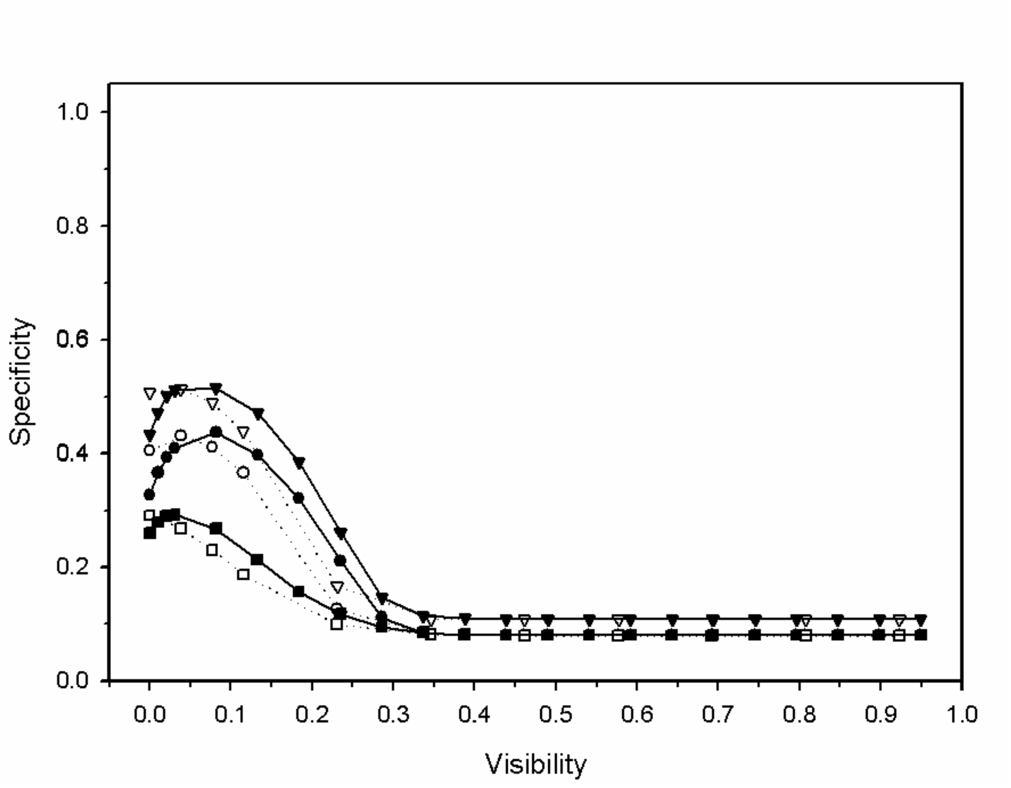

1.0
0.8
0.6
0.4
0.2
0.0
Specificity
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Visibility

Figure 4

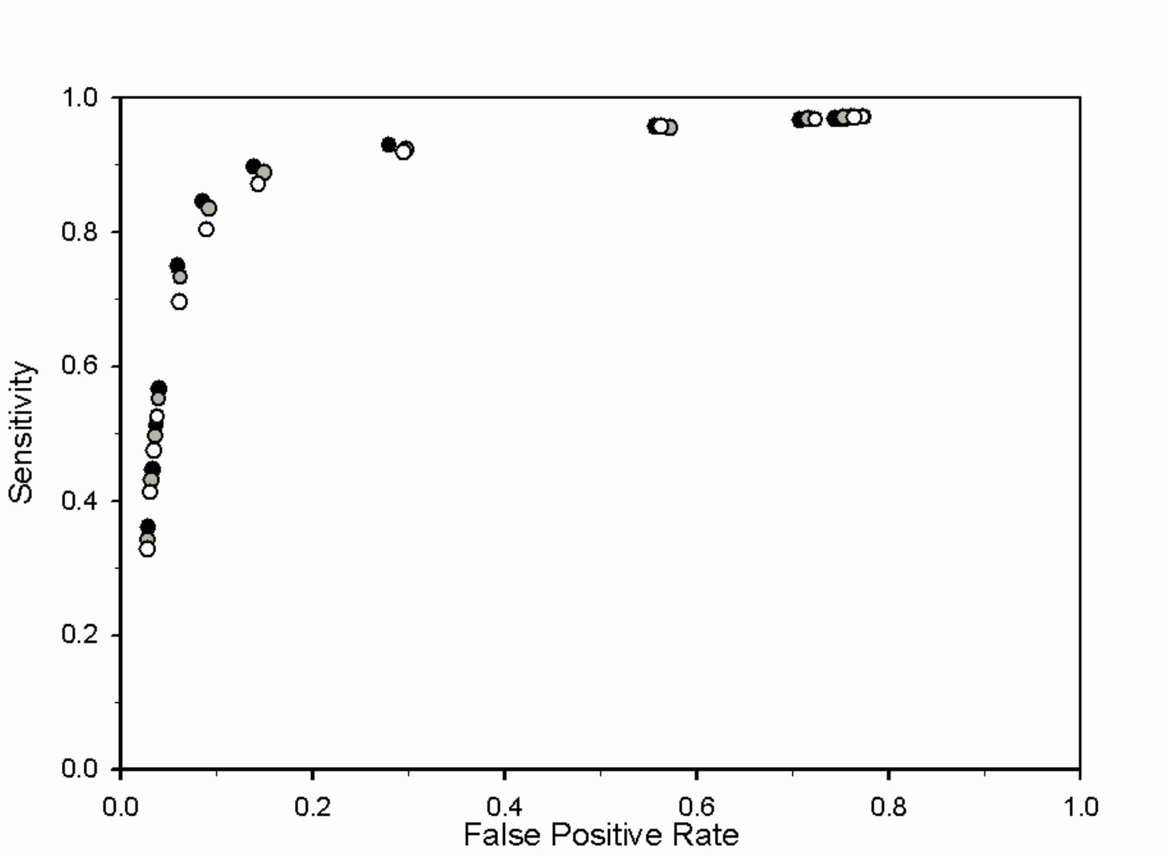
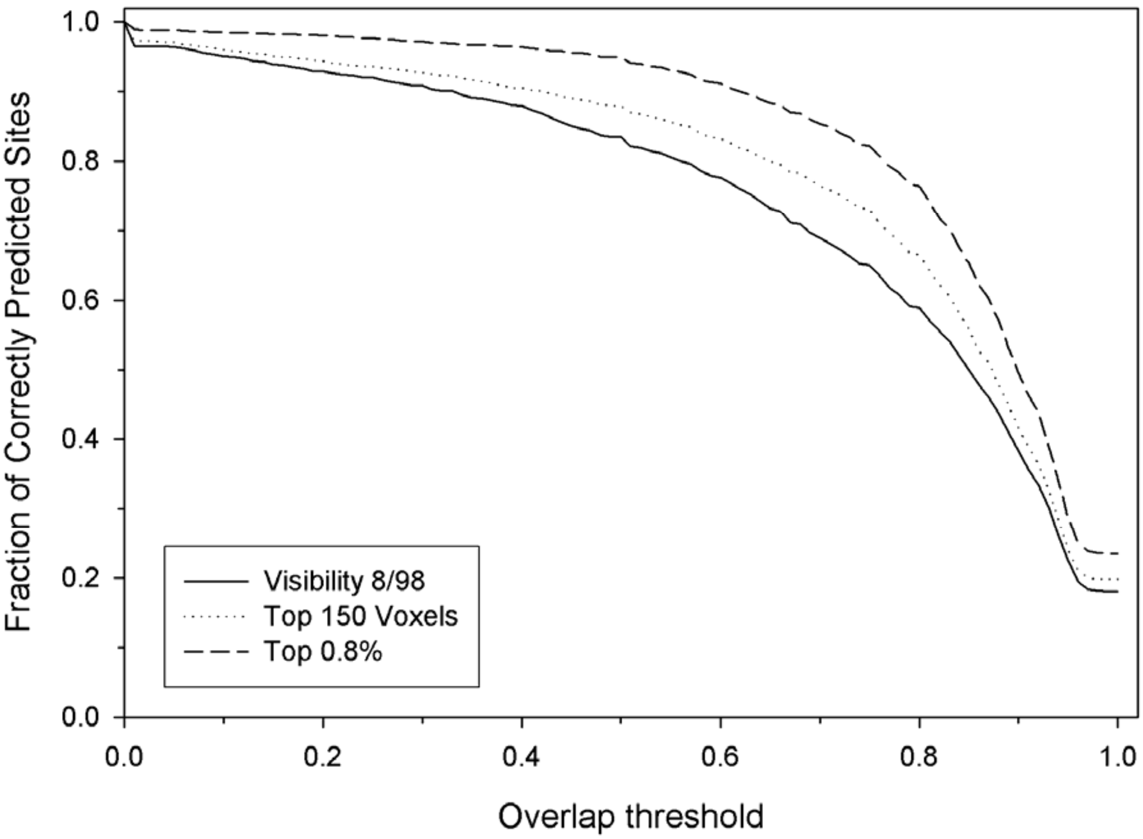

1.0
0.8
0.6
0.4
0.2
0.0
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate

Figure 5



Fraction of Correctly Predicted Sites
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Overlap threshold
Visibility 8/98
Top 150 Voxels
Top 0.8%

Figure 6

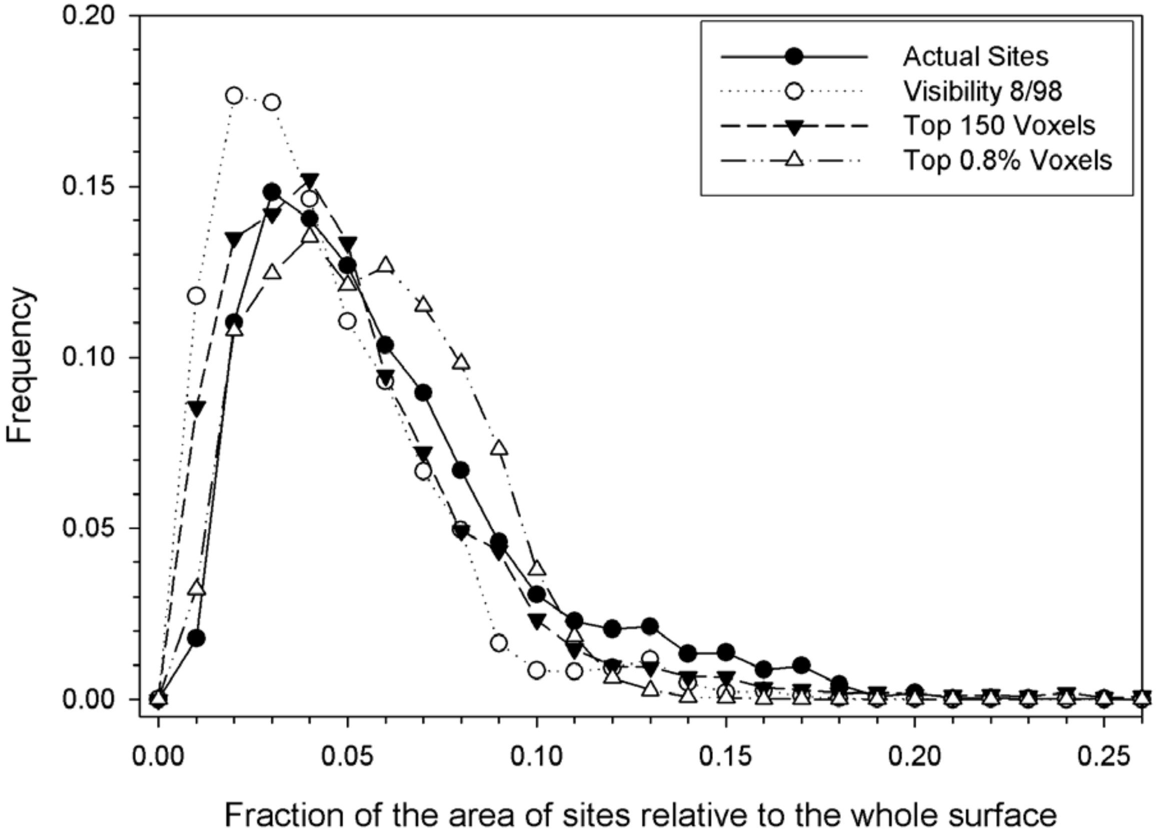

0.20
0.15
0.10
0.05
0.00
Frequency
0.00
0.05
0.10
0.15
0.20
0.25
Fraction of the area of sites relative to the whole surface
Actual Sites
Visibility 8/98
Top 150 Voxels
Top 0.8% Voxels

Figure 7

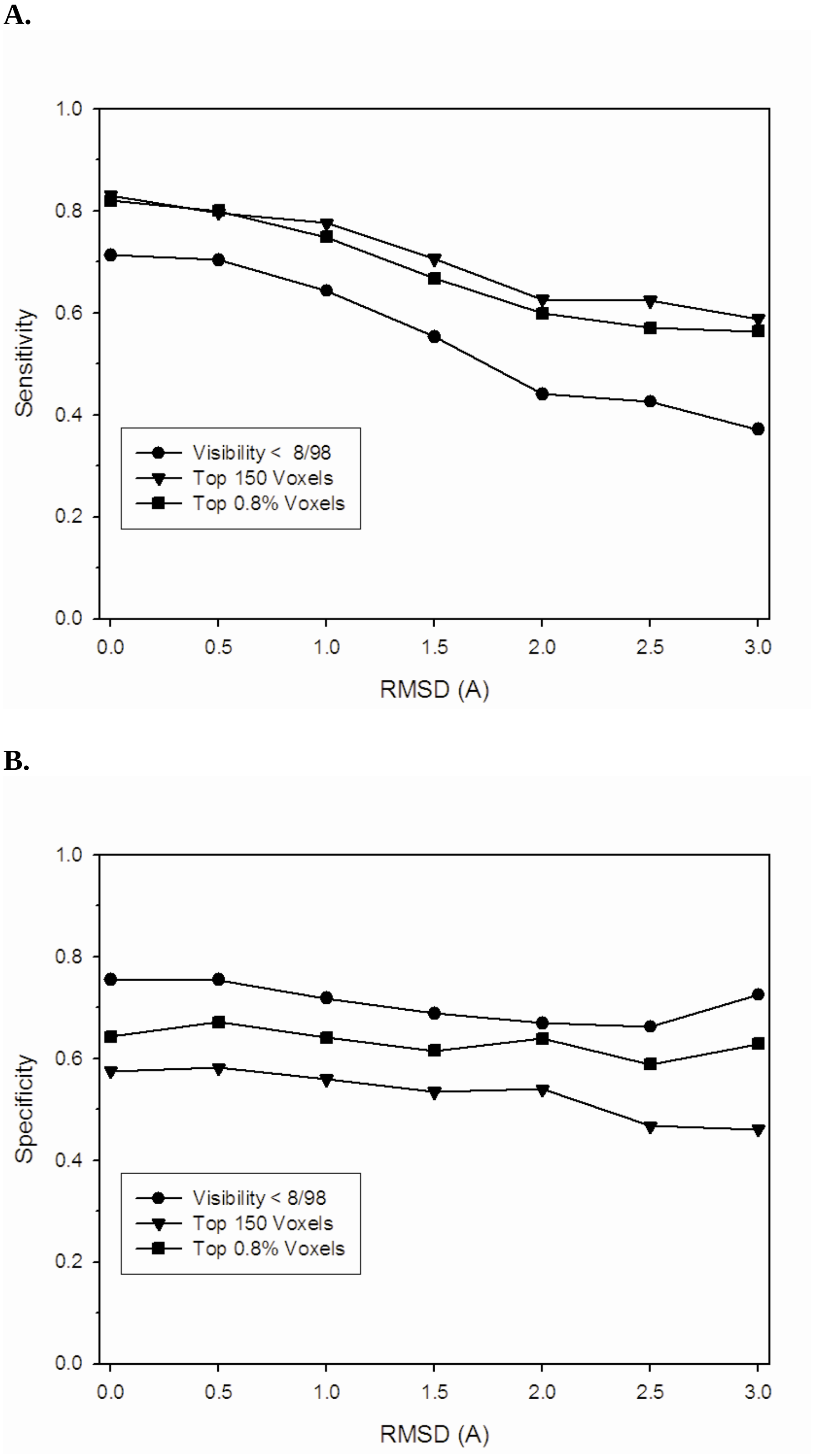
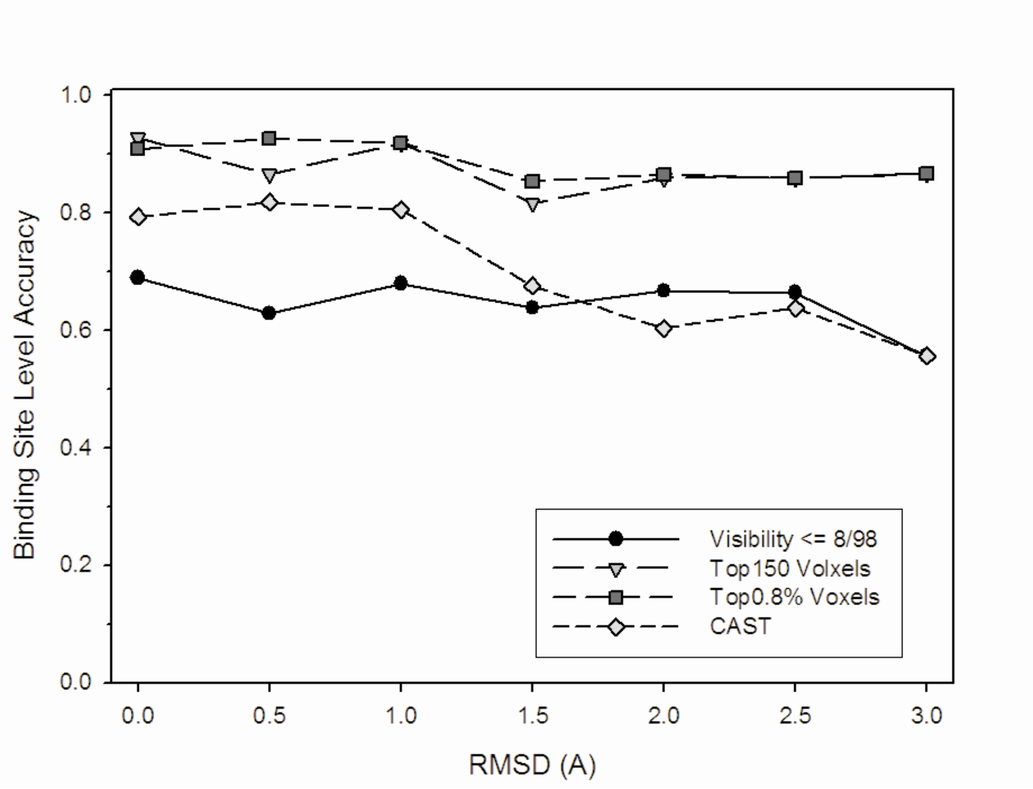

A.
1.0
0.8
0.6
0.4
0.2
0.0
Sensitivity
0.0
0.5
1.0
1.5
2.0
2.5
3.0
RMSD (A)
Visibility < 8/98
Top 150 Voxels
Top 0.8% Voxels
B.
1.0
0.8
0.6
0.4
0.2
0.0
Specificity
0.0
0.5
1.0
1.5
2.0
2.5
3.0
RMSD (A)
Visibility < 8/98
Top 150 Voxels
Top 0.8% Voxels

Figure 8
A.


1.0
0.8
0.6
0.4
0.2
0.0
Binding Site Level Accuracy
0.0
0.5
1.0
1.5
2.0
2.5
3.0
RMSD (A)
Visibility <= 8/98
Top150 Volxels
Top0.8% Voxels
CAST

B.

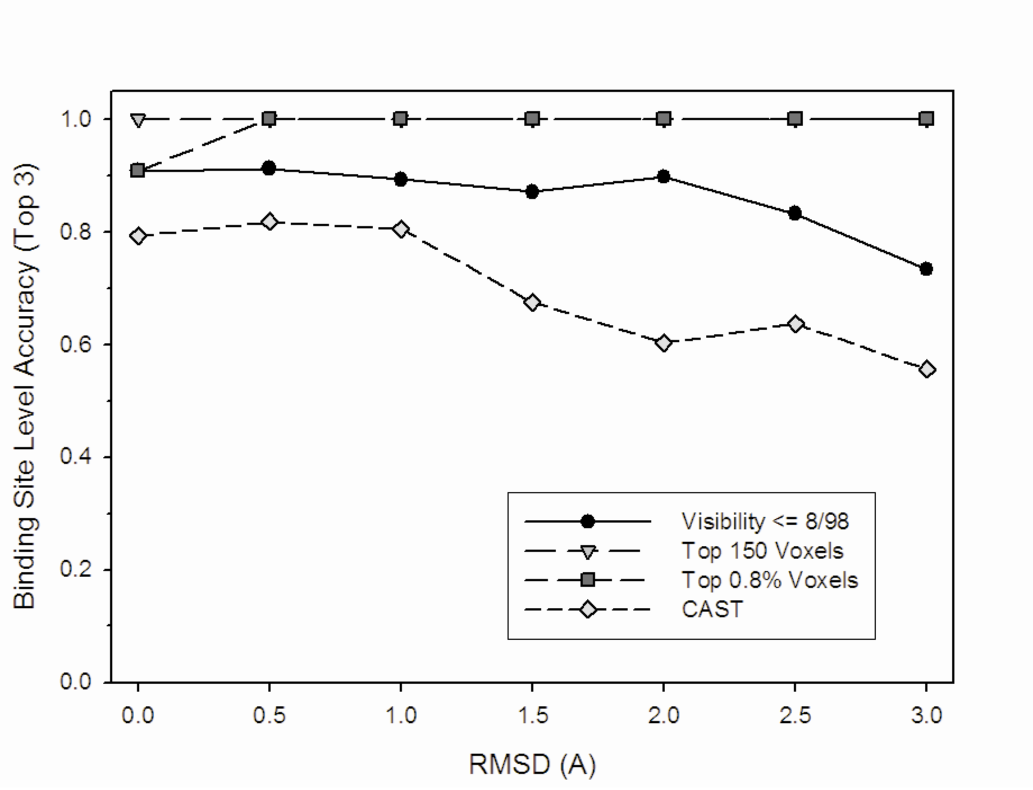
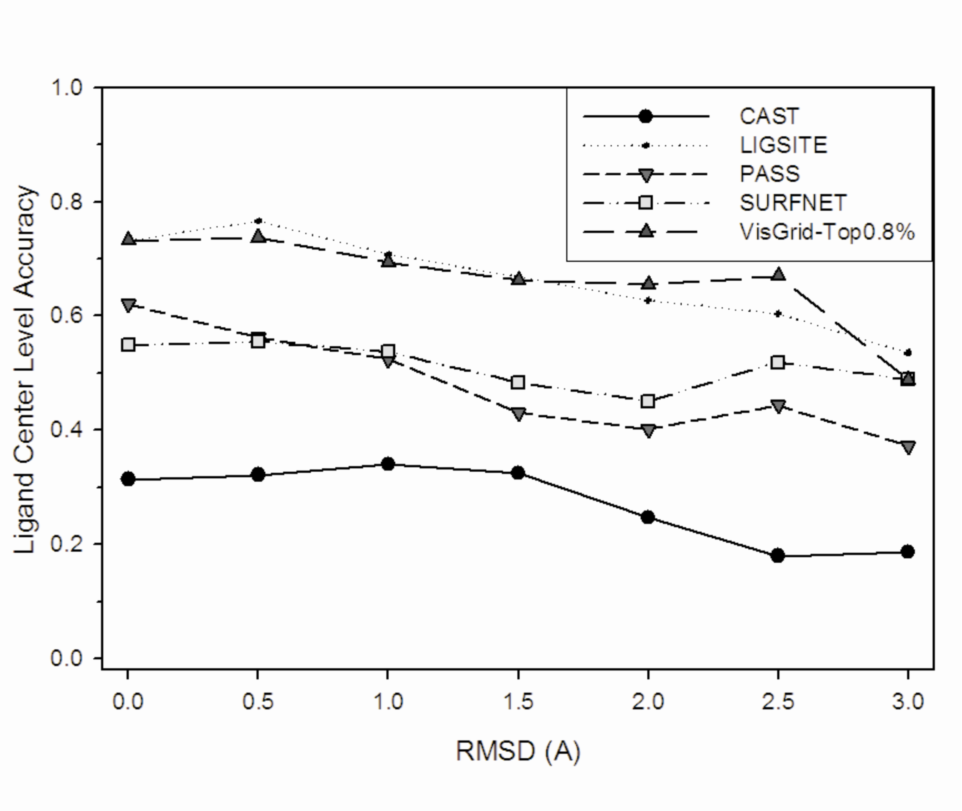

1.0
0.8
0.6
0.4
0.2
0.0
Binding Site Level Accuracy (Top 3)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
RMSD (A)
Visibility <= 8/98
Top 150 Voxels
Top 0.8% Voxels
CAST

Figure 9
A


1.0
0.8
0.6
0.4
0.2
0.0
Ligand Center Level Accuracy
0.0
0.5
1.0
1.5
2.0
2.5
3.0
RMSD (A)
CAST
LIGSITE
PASS
SURFNET
VisGrid-Top0.8%

B

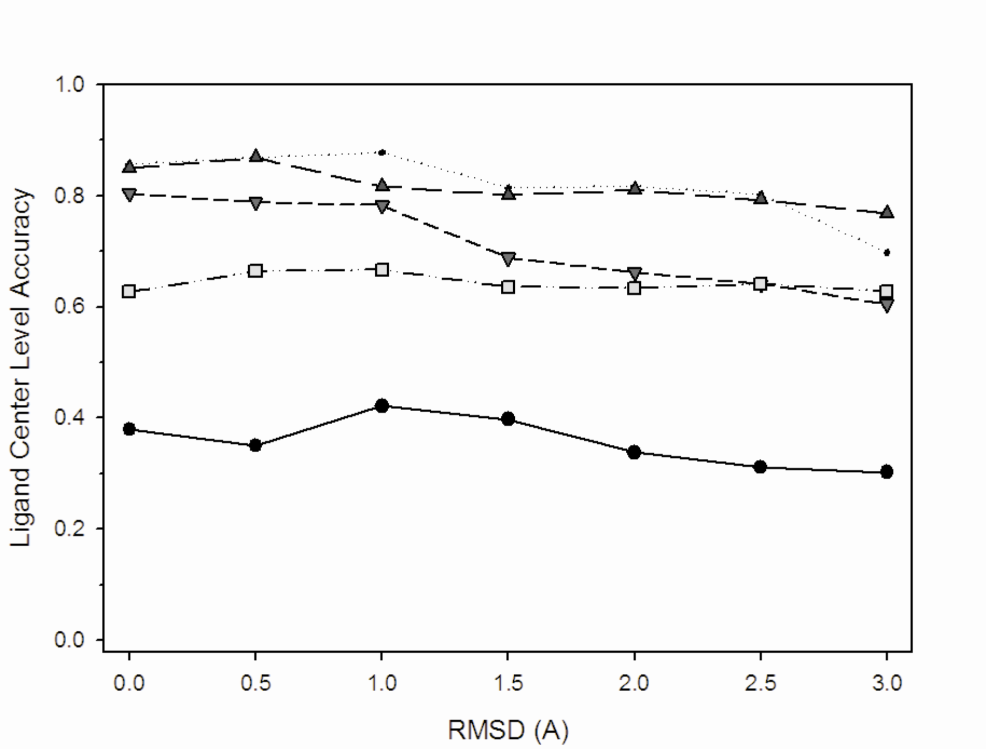

1.0
0.8
0.6
0.4
0.2
0.0
Ligand Center Level Accuracy
0.0
0.5
1.0
1.5
2.0
2.5
3.0
RMSD (A)

Figure 10

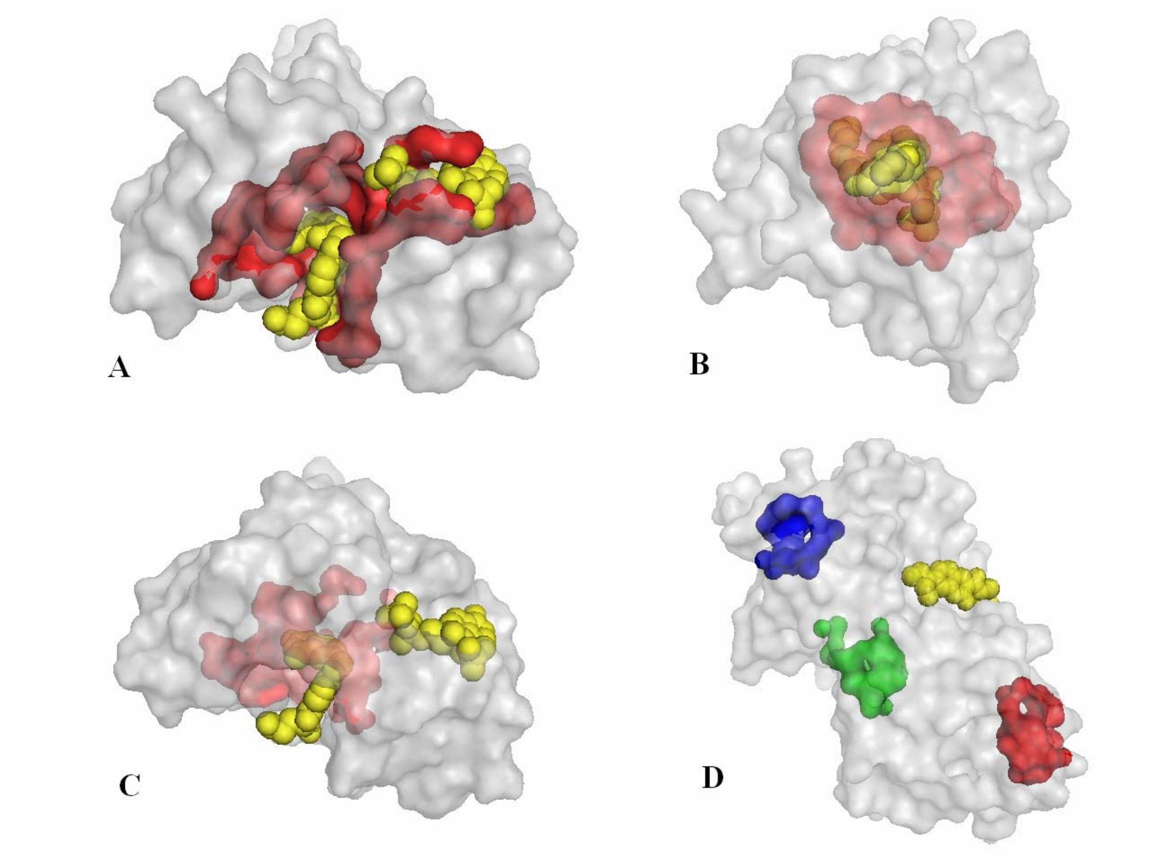


Figure 11