Published in final edited form as:

J Struct Funct Genomics. 2010 March ; 11(1): 21–30. doi:10.1007/s10969-010-9078-7.

Assisted assignment of ligands corresponding to unknown electron density

T. Andrew Binkowski, Marianne Cuff, Boguslaw Nocek, Changsoo Chang, and Andrzej Joachimiak

Midwest Center for Structural Genomics (MCSG), Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439, USA

Abstract

A semi-automated computational procedure to assist in the identification of bound ligands from unknown electron density has been developed. The atomic surface surrounding the density blob is compared to a library of three-dimensional ligand binding surfaces extracted from the Protein Data Bank (PDB). Ligands corresponding to surfaces which share physicochemical texture and geometric shape similarities are considered for assignment. The method is benchmarked against a set of well represented ligands from the PDB, in which we show that we can identify the correct ligand based on the corresponding binding surface. Finally, we apply the method during model building and refinement stages from structural genomics targets in which unknown density blobs were discovered. A semi-automated computational method is described which aims to assist crystallographers with assigning the identity of a ligand corresponding to unknown electron density. Using shape and physicochemical similarity assessments between the protein surface surrounding the density and a database of known ligand binding surfaces, a plausible list of candidate ligands are identified for consideration. The method is validated against highly observed ligands from the Protein Data Bank and results are shown from its use in a high-throughput structural genomics pipeline.

Keywords

Electron density assignment; Function annotation; Ligand identification; Ligand assignment; Protein surfaces

Introduction

During protein crystal structure determination, it is common to uncover that a small molecule ligand has been serendipitously co-crystallized with the macromolecule. Some bound ligands are from the crystallization media and their interpretation is rather straightforward. However, quite often, the ligand is carried out with the protein through purification and crystallization steps. An unambiguous interpretation of the electron density map can be a difficult and time-consuming challenge because low occupancy or partial disorder, yet it may be functionally important.

In the past, the interpretation and fitting of bound ligands relied on the expertise of the crystallographer. However, incorrectly assigned ligands impact structure refinement, hinder accurate depositions and, most importantly, places functionally incorrect structure into the

public domain. In some cases, crystallographers have chosen to exclude ligands from the deposition to prevent misidentification. In either case, improper treatment of the ligand can undermine the quality and functional impact of the newly determined protein structure. In the context of structural genomics programs, identification of bound ligands may unlock valuable clues in the functional characterization in proteins of unknown function.

While methods for ligand fitting are becoming automated and increasingly accurate [1-6], only recently has a method been described to address the precursory problem of identification of unknown bound ligands. Terwilliger et al. [7] measured the correlation between observed unknown density and a library of ligands systematically fitted into the density for ligand identification. This approach was able to identify the top ranking ligand as the correct ligand in 48% of their evaluation for the top 200 most abundant ligands in the PDB. When combined with automatic ligand fitting, assisted ligand identification can help to improve depositions and the overall quality and biological impact of the structures.

Previous studies on protein functional surfaces have shown that binding surfaces are the most conserved features in proteins and provide information about the ligand preference for a given site [8-13]. We have developed an approach, surface assisted interpretation of density (*SurfaceAID*) to leverage this high degree of physicochemical texture and shape conservation of the ligand binding site to assist in identification of bound ligands. Using the local atomic surface surrounding the non-protein electron density as the query, we search a library of known ligand bindings surfaces. Similar surfaces returned from the search provide a list of corresponding ligands to consider, effectively minimizing the search space that a crystallographer must consider when modeling a ligand. In this study, we explore and evaluate the effectiveness of using protein surface similarity for ligand identification by benchmarking retrieval rates for most frequently observed ligands in the PDB. We then provide examples of the application used on structures genomics targets.

Methods

Binding surface definition

Assisted modeling of ligand to unassigned electron density is accomplished through the comparison of a query surface against a database of annotated binding sites. The method requires a predefined surface encompassing the electron-density to be defined as input. Our algorithms are specifically designed to be used at early stages of model building and refinement, so we do not assume that the model is complete or all protein residues have been correctly assigned.

The binding surface, composed of the solvent accessible atoms that are in proximity to the density, can be defined manually by the crystallographer or through simple geometrical rules (i.e. atoms within 5 Å of the density). In cases where the model is complete in the region of unassigned density, auxiliary computational methods can be utilized to automatically define the cavity (e.g. CASTp webserver) [14]. In most cases, automated methods provide the most accurate pocket definitions, as the cavity boundaries adhere to strict criteria (e.g. solvent accessibility, volume characteristics). However, if the model around the density is not complete, automated methods will return inaccurate or meaningless pocket definitions.

Spatial alignment of surfaces

Spatial alignment of surfaces is accomplished by identifying the optimal three-dimensional superposition between the query and the library surfaces. The first step in our alignment is to apply a pose normalization technique using principal component analysis (PCA). To overcome some of the shortcomings of traditional PCA alignments, we implement a step based on the

work of Wei and Yuanjun [15] that replaces the original vertices set with a uniformly generated set of vertices. This step aims to reduce the bias of non-uniform distribution of points in a 3D model (which can be observed in protein surfaces). We apply the algorithm for uniform sampling from the works of Osada et al. [16].

Once the new, uniform surface has been generated, the surface objects are rotated to share a canonical reference frame such that the eigenvector with the largest eigenvalue is aligned along the X -axis (Fig. 3). The second largest and smallest eigenvalues map to the Y and Z -axis, respectively. To further improve our alignments, we refine the positioning by allowing the library surface to wiggle around all three major axes in 1° intervals. At each new position, we rescore the alignment and either continues or cease the wiggle based on a threshold.

Evaluation of surface alignments

Shape complementarity—While the alignment of 3D surfaces is straightforward, scoring the alignments requires careful consideration of both the shape and physicochemical properties of ligand binding surfaces. In our method, shape complementarity is based on maximizing the volume overlaps of the convex hull of the surface, V_H , and the volume of the surface atoms, V_S (Fig. 3).

The convex hull of the surface is identified from a tessellation of the surface atoms (Fig. 3b) and represents a gross approximation of the surface cavity. More precise methods of measuring cavity volume exist, but would require a completed structural model to be built. We accept the limitations of the approximated convex hull volume, as we designed this method under the assumption that the models may not be complete. Surface volumes are calculated using the weighted Delauney triangulation and alpha shape methods [17-19].

The surface overlap volume is defined as the volume difference between the superimposed surfaces and is calculated from the formula:

$$V_{S_{AB}} = V_{S_A} + V_{S_B} - V_{S_{A \cup B}}$$

where V_{S_A} , V_{S_B} , and the posed construct $V_{S_{A \cup B}}$ are the volumes of surface A , B and the superimposed construct AB , respectively. The convex hull overlap volumes are computed similarly:

$$V_{H_{AB}} = V_{H_A} + V_{H_B} - V_{H_{A \cup B}}$$

The overlap volumes are then used to calculate a Tanimoto coefficient, which is a normalized similarity measure between objects [20,21]. By using the self overlap volumes ($V_{S_{AA}}$, $V_{S_{BB}}$), we can define the surface volume overlap Tanimoto:

$$T_{S_{AB}} = \frac{V_{S_{AB}}}{V_{S_{AA}} + V_{S_{BB}} - V_{S_{AB}}}$$

The Tanimoto score identifies the volume overlap of optimally aligned surfaces. The score is bounded between 0 (representing non-overlapping surfaces) and 1.0 (representing identical surfaces). The hull volume overlap Tanimoto is calculated similarly:

$$T_{H_{AB}} = \frac{V_{H_{AB}}}{V_{H_{AA}} + V_{H_{BB}} - V_{H_{AB}}}$$

Physicochemical complementarity—For each given alignment, we evaluate the complementarity of aligned atom types. For each surface atom, we search to find any identical atoms types within 3.6Å, awarding a value of 1 for a atomic match and a 0 to a nonmatch. This distance is chosen to simulate plausible hydrogen bonding distance with an error factor to account for modeling inaccuracies and protein flexibility. Only oxygen, nitrogen, and sulfur atoms are considered in this scoring scheme.

Again, we use a Tanimoto based scoring function, allowing us to describe the atomic complementarity in terms of overall similarity (e.g. a score 0.5 would mean that 50% of the atoms in the query surface have identical counterparts in the library surface). Using the self overlap identities (I_{AA} , I_{BB}), the Tanimoto for the atomic identity, I , between surface A and B is calculated as follows:

$$T_{I_{AB}} = \frac{I_{AB}}{I_{AA} + I_{BB} - I_{AB}}$$

Universal scoring—All the Tanimoto scores are combined to create a single evaluation of the aligned shape and atomic composition complementarity. The surface overlap similarity (SOS) score,

$$\text{SOS} = T_{H_{AB}} + T_{S_{AB}} + T_{I_{AB}}$$

is bounded between 0 (no similarity) and 3 (identical surfaces). This is the default scoring scheme used to identify the most similar surfaces in a library to a given query surface.

Ligand posing

Once a set of similar protein surfaces has been selected from the database, a final step is taken to remove less likely/false positive ligands from consideration. The ligand corresponding to the highly scored library surface is subjected to the same rotation and translation as the surface, effectively positioning it into the query surface. In many ways this is akin to a very crude form of molecular docking. From this pose, we make a simple evaluation of the spatial availability of the ligand to fit in the cavity by measuring any atomic overlaps.

The overlapping volume between the query pocket (V_Q), the ligand (V_L), and the posed construct (V_{QUL}) can be calculated from the formula:

$$V_{QL} = V_Q + V_L - V_{Q \cup L}$$

The overlapping volume is normalized by using the self overlap volumes (V_{QQ} , V_{LL}) to define a more convenient Tanimoto value. We define the surface pose overlap Tanimoto (SPOT):

$$T_{QL} = \frac{V_{QL}}{V_{QQ} + V_{LL} - V_{QL}}$$

The SPOT score identifies amount of clashing between the ligand and the surface, representing how well a ligand fits into a binding pocket. Given this definition, a score of 0 indicates the pocket is sufficiently large enough to contain the entire ligand. A score approaching its upper boundary (of 1), is indicative of a ligand that is too large for the pocket.

In some cases, the SPOT score may not provide any additional insight as ligands occur in many different conformations and the volume of ligand may not correlated well to the pocket volume [22]. Nonetheless, the calculation is non-prohibitive and may provide additional information for the crystallographer.

Searching the Global Protein Surface Survey—The Global Protein Surface Survey (GPSS) is an annotated library of protein surfaces derived from structural data in the PDB. Currently, there are 187,691 ligand binding surfaces in the GPSS (PDB version: July 2009). There are 65,064 unique ligand binding surfaces, if we consider only one ligand-type per structure (e.g. only the first ATP is counted in a structure with three bound ATP molecules). The GPSS is updated weekly to correspond to PDB deposits and can be accessed through a web browser (<http://www.gpss.mcsg.anl.gov>) or via a PyMol (DeLano Scientific, Palo Alto, CA, USA) plugin.

Our method uses the GPSS as its search library and returns a ranked list of the top scoring surfaces and their corresponding ligands. The method returns the top scoring molecules in both PDB and SMILES (simplified molecular input line entry specification) format. The latter are useful for modeling environments such as COOT [23], where an automated ligand fitting procedure [3] can take SMILES strings as input. An overview of our method is diagramed in Fig. 1.

An exhaustive search of the entire GPSS library would take approximately 36 h to conduct on a single processor. The algorithm does take advantage of multi-processor and multi-core computing architecture to dramatically reduce the runtime. More importantly, careful consideration and preparation of the search library has the most dramatic effect of runtime. In practice, we use a subset of the search library that is non-redundant (in terms of surface similarity) and has been filtered to only include natural metabolites and known crystallization artifacts. Using this library subset, a search takes approximately 1 h on a quad-core machine.

Evaluating surface retrieval experiments using receiver operator characteristic curves

Surface retrieval benchmarking experiments are summarized in a receiver operator characteristic (ROC) curve, where the sensitivity is plotted against its specificity at various significance levels of summed probabilities. In the ROC curve, the x -axis represents the false positive rate, or 1-specificity, which is calculated by $1 - \text{TN}/(\text{TN} + \text{FP})$, where TN is the number of true negatives and FP is the number of false positives. The y -axis represents the true positive rate, or sensitivity, and is calculated as $\text{TP}/(\text{TP} + \text{FN})$, where FN is the number of false negatives. An overall performance measure of a classification test can be calculated by the area under the ROC curve (AUC) [24]. Bound between 0 and 1, an AUC of 1 is indicative of a perfectly accurate classification test, in which all true positives are distinguished from false positive. An AUC of 0.5 corresponds to a random classification test (e.g. a coin flip). The AUC is a combined measure of sensitivity and specificity.

For our methodology, the ROC curves measure our ability to accurately identify a surface binding the same ligand from the GPSS. A true-positive data point in our database retrieval experiments is defined when the ligand from the query surface matches the ligand from the corresponding library surface (e.g. retrieving heme binding sites when a heme binding site is used as the query).

Results and discussion

Benchmarking ligand retrieval using surface similarity for the most abundant PDB ligands

The utility of our method relies on its ability to identify binding surfaces sharing similar shapes. To validate this we perform retrieval experiments, in which we search a given binding surface against the GPSS library of surfaces to measure our ability to correctly return surfaces binding the same ligand. To benchmark the method we selected the most 15 most abundant ligands in the PDB (Table 1). The ligand frequency was limited to a single occurrence per structure, so as to remove bias from proteins binding the same ligand in multiple pockets.

For each ligand type, we selected three different binding surfaces representing a high, medium and low resolution structure. Other than the resolution cutoff, the surfaces were randomly selected. Each query surface was searched against the GPSS and the retrieval rates, measured as the AUC, were measured (Table 2). The rank of the first occurrence of the identical ligand was recorded.

From our query ligand set, we observe some trends that affect the overall performance. First, the molecular weight of the ligand exhibits some influence over the AUC retrieval rates. In general, the larger ligands (FAD, NAD, ATP, ADP, HEM) have the best overall AUC retrieval rates. Smaller ligands (GOL, MAN, GLC, MPD, SO4) had overall lowest retrieval rates. Interestingly, the three atom amino (NH₂) and acetyl (ACE) group ligands performed better than most small ligands. This may be due to the fact that the majority of ligands in the PDB are too large to fit into their small surface and are heavily penalized in our scoring system.

The retrieval rate for ATP, a highly ubiquitous and well-represented ligand in the PDB, seems to underperform compared to other ligands. The highest resolution structure (1.2 Å) finds the native ligand in only the fifth position, with other nucleotides (GTP, ADP, GDP) ranking higher. These results are consistent with previously reported results in comparing nucleotide binding sites [11]. This inability to distinguish between nucleotides is also mirrored in some proteins. For example, the highly conserved eukaryotic serine/threonine kinase, casein kinase 2 (CK2), possesses dual-cosubstrate specificity for GTP and ATP [25]. In specialized binding sites such as these, automated methodologies may not be sufficient to uncover subtleties that would provide better specificity.

In general, our method appears to be rather insensitive to the overall structure resolution. For each of the three resolution levels, we do not see a significant difference in the retrieval rates across our query set. The returned rank of the native ligand do not seem to be systematically affected by the resolution. For every ligand type, at least one query returned the identical ligand from another structure as the the first ranked hit. The other queries varied, but were still highly ranked.

The resolution independent performance of our method is not surprising, as the chemical properties of ligand binding surfaces are independent of resolution. Structure resolution would only affect our method to the extent that low resolution data has on the entire PDB, most importantly the completeness of the macromolecular structure. Due to the rather small sample size of this data set, it is unclear if these trends will hold-up against a larger set of ligand types. The authors plan to conduct a more exhaustive analysis of the retrieval rates on a global protein surface scale.

Global benchmarking native ligand identification

To evaluate the robustness of our method against all PDB ligands, we created a data set of 3,120 non-redundant binding surfaces to test. Surfaces included in this set were subject to the following conditions: (1) the protein was non-redundant (<90% sequence similarity) to any

other member in the set, (2) there was no identical ligand from the same structure, and (3) there were at least two instances of the ligand from different structures in the PDB.

Each of the surfaces was then searched against the GPSS library and the rank position of the first occurrence of the identical ligand was recorded. A summary plot for entire set is shown in Fig. 2. For this set, an identical ligand binding surface was the top ranked surface 72% of the time. The correct ligand was found within the top 25 for 98% and top 30 for 99% of the surfaces in our data set. This data suggests that, in the worst case scenario, a crystallographer would be able to identify the correct ligand every time if they would model and refine the top 30 ligands.

Assisted assignment of protein from structural genomics

Structural genomics centers face the challenging task of determining structures of unique proteins with no or low sequence homologs in the PDB. When a protein function is unknown, even fewer clues are present to help decipher incident ligand electron density blobs. We have applied our methodology to unique proteins of unknown function from the Midwest Center for Structural Genomics (MCSG). In all cases, the method was used during early model building stages by expert crystallographers. Three examples are shown below.

Fructose binding surface—A putative sugar isomerase from *A. fulgidus*, contained an unassigned electron density near a small eight residue surface cavity (Fig. 4, white). The density was ambiguously small and consistent with the size of molecules in the crystallization conditions. Attempting to model small ligands present in the crystallization media into the electron density did not improve refinement and resulted in negative density in the $2F_O - F_C$ difference maps, eliminating them from consideration.

The surface around the electron density was queried against the GPSS and returned a highly similar fructose (FRU) binding surface of exo-inulinase from *A. awamori* (PDB id = 1y9g). The library surface was the top ranked hit with a $Surface_{AID}$ score of 2.28. Four residues, Asp¹⁴⁹, Asp¹⁵¹, Gln⁹⁶, and Phe¹⁵², are found to be spatially conserved in the alignment. The alignment of the surfaces is shown in Fig. 4b. Fructose was subsequently modeled into the density (Fig. 4c) and successfully refined for deposition (PDB id = 1zx5).

NADP binding surface—A large unknown density blob was discovered in a putative Δ^1 -pyrroline-5-carboxylate reductase from *N. meningitides* (Fig. 5d). The complete model was built and submitted to the CASTp webserver for automated surface identification. A 27 residue surface was identified which encompassed the density (Fig. 5d, white).

Surface matching revealed striking similarity to many nicotinamide adenine dinucleotide phosphate (NADP) binding sites including the top ranked ($Surface_{AID}$ score = 1.98) from carbonyl reductase sniffer of *D. melanogaster* (PDB id = 1sny, Fig. 5e, light orange). The query surface shares 12 conserved residues with the library surface and exhibits shape similarity to allow the NADP to be easily modeled into the final deposition (PDB id = 2ag8). The NADP molecule is shown fitted to the density in Fig. 5f. The binding site also provided additional insight into the function of this protein and assisted in predicting the target's as a Δ^1 -pyrroline-5-carboxylate reductase.

A candidate set of similar nucleotides—While it is clear that proteins conserve shape and atomic patterns for binding and function, discrimination between surface recognizing closely related ligands can be quite challenging. Many ligands occurring in the PDB are chemically similar and, as expected, their binding surfaces contain equally similar shape and physicochemical properties. This was observed during the search for a ligand corresponding to density discovered in protein HP0184 with unknown function from *H. pylori* (Fig. 6a). A

Surface_{AID} search of the 26 residue surface returns a series of nucleotide binding surfaces with near identical scores (± 0.01) for the top 12 ranked surfaces. Figure 6 shows the binding surface of guanosine-5-triphosphate (GTP) of gamma-tubulin from *H. sapiens* (b, green), guanosine-5-diphosphate (GDP) of mannosylglycerate synthase from *R. marinus* (c, yellow), 2-deoxyguanosine-5-triphosphate (dGTP) of nucleoside diphosphate kinase II from *A. thaliana* (d, blue), and 5'-guanosyl-methylene-triphosphate (GmTP) of oncogene protein from *H. sapiens* (e, cyan). The surfaces are all shown superimposed with protein HP0184 (white) based on the best scoring alignment solution. The corresponding ligands are shown in f, posed by their surface superpositions (ligands are colored to correspond to the color scheme in b–e).

All four candidate ligands were modeled, in turn, into the structure and subjected to a limited number of rounds of refinement and showed similar fit into the electron density. In this case, the crystallographer's judgement drove the final decision and inclusion of the DGT molecule into the structure for final refinement and deposition (PDB id = 2atz). The modeled DGT ligand and its corresponding density are shown in Fig. 6g.

Conclusion

In the absence of any prior knowledge about protein function and a bound ligand, the crystallographer's task transforms from experimental data interpretation to empirical judgement. Faced with a well-defined electron density blob, the crystallographer has some, but limited, options: build dummy atoms, unconfidently assign a ligand, build solvent, or leave the electron density unassigned. All of these options result in a reduction of the quality, accuracy and biological impact of the deposition.

In this study, we introduced a method, *Surface_{AID}*, that can be applied to these difficulty cases. Using *Surface_{AID}*, the crystallographer can explore said difficult ligand space for assistance in identifying ligands corresponding to unknown electron density. Using surface similarity is a highly advantageous approach, as functional sites are conserved beyond any sequence or fold similarity. The method operates independently of any particular crystallography software package and can be easily integrated into existing model building/refinement workflow.

Surface_{AID} presents a ranked list of plausible ligands for fitting into unknown electron density based on geometric shape and physicochemical texture similarity of the atoms surrounding the density blob. The method is designed to work at all stages of the model building and refinement process and shows little degradation of sensitivity for low resolution structure. We showed that this method is highly capable at identifying the native ligand of existing protein–ligand complexes within the top ten returned ligands for 93% of the PDB (non-redundant). Over 99% of native ligands can be identified within the top 30 returned search results. These numbers suggest that a significant reduction in the ligand model building effort of a crystallographer may be possible when using this method.

In general, *Surface_{AID}* performs better at identifying larger ligands, which presumably have more distinctive binding surfaces. The worst performance is observed for identifying small ions, in which our method struggles to obtain the specificity desirable to assist in ligand assignment. These are similar patterns reported in other methods [7]. In some circumstances, *Surface_{AID}* may outperform electron density based methods, where different ligands may fit equally well into observed density [7]. Because our method works independently of the observed electron-density during its initial searching phase, it does not suffer when using low resolution data. It also minimizes any bias that may exist from energetically unfavorable conformations that may be disfavored in methods that build ligands from density.

The obvious limitation of our method is that it is only effective for ligands that already exist in the PDB (i.e. it requires that a protein–ligand complex must already be deposited). This

method would not be suitable in some types of ligand binding studies using novel ligands or where a mixture of ligands is mixed into crystallization conditions. It could, however, be used to systematically exclude known ligands from modeling consideration to provide indirect support for a particular bound ligand. Combining the *SurfaceAID* method with automated ligand fitting routines may improve overall ligand assignment efficiency and accuracy.

Acknowledgments

We would like to thank all members of the Midwest Center for Structural Genomics for helpful discussion and access to crystallographic data. The benchmarking studies were performed on the Argonne Leadership Computing Facilities' BlueGene/P computing system as part of the Department of Energy's INCITE program. This work was supported by National Institutes of Health Grants GM62414 and GM074942 and by the US Department of Energy, Office of Biological and Environmental Research, under contract DE-AC02-06CH11357.

References

1. Diller DJ, Pohl E, Redinbo MR, Hovey BT, Hol WG. *Proteins* 1999;36(4):512–25. [PubMed: 10450093]
2. Oldfield TJ. *Acta Crystallographica Section D* 2001;57(5):696–705.
3. Terwilliger TC. *Acta Crystallographica Section D* 2002;58(11):1937–1940.
4. Zwart PH, Langer GG, Lamzin VS. *Acta Crystallographica Section D* 2004;60(12 Part 1):2230–2239.
5. Aishima J, Russel DS, Guibas LJ, Adams PD, Brunger AT. *Acta Crystallogr D Biol Crystallogr* 2005;61 (Pt 10):1354–1363. [PubMed: 16204887]
6. Wlodek S, Skillman AG, Nicholls A. *Acta Crystallogr D Biol Crystallogr* 2006;62(Pt 7):741–749. [PubMed: 16790930]
7. Terwilliger TC, Adams PD, Moriarty NW, Cohn JD. *Acta Crystallogr D Biol Crystallogr* 2007;63(Pt 1):101–107. [PubMed: 17164532]
8. Laskowski RA, Watson JD, Thornton JM. *J Mol Biol* 2005;351(3):614–626. [PubMed: 16019027]
9. Binkowski TA, Adamian L, Liang J. *J Mol Biol* 2003;332(2):505–526. [PubMed: 12948498]
10. Binkowski TA, Joachimiak A, Liang J. *Protein Sci* 2005;14(12):2972–2981. [PubMed: 16322579]
11. Binkowski TA, Joachimiak A. *BMC Struct Biol* 2008;8:45. [PubMed: 18954462]
12. Stockwell GR, Thornton JM. *J Mol Biol* 2006;356(4):928–944. [PubMed: 16405908]
13. Kahraman A, Morris RJ, Laskowski RA, Thornton JM. *J Mol Biol* 2007;368(1):283–301. [PubMed: 17337005]
14. Binkowski TA, Naghibzadeh S, Liang J. *Nucleic Acids Res* 2003;31(13):3352–3355. [PubMed: 12824325]
15. Wei, L.; He, Y. In: Zha, H.; Pan, Z.; Thwaites, H.; Addison, AC.; Forte, M., editors. *Interactive technologies and sociotechnical systems*, 12th international conference, VSMM 2006; Xi'an, China. 18–20 October 2006; Springer; 2006. p. 263–268.
16. Osada R, Funkhouser T, Chazelle B, Dobkin D. *ACM Trans Graph* 2002;21(4):807–832.
17. Liang J, Edelsbrunner H, Woodward C. *Protein Sci* 1998;7(9):1884–1897. [PubMed: 9761470]
18. Liang J, Edelsbrunner H, Fu P, Sudhakar PV, Subramaniam S. *Proteins* 1998;33(1):1–17. [PubMed: 9741840]
19. Liang J, Edelsbrunner H, Fu P, Sudhakar PV, Subramaniam S. *Proteins* 1998;33(1):18–29. [PubMed: 9741841]
20. Willett P, Barnard J, Downs G. *J Chem Inf Comput Sci* 1998;(38):983–996.
21. Rush T Sr, Grant JA, Mosyak L, Nicholls A. *J Med Chem* 2005;48(5):1489–1495. [PubMed: 15743191]
22. Nayal M, Honig B. *Proteins* 2006;63(4):892–906. [PubMed: 16477622]
23. Emsley P, Cowtan K. *Acta Crystallogr D Biol Crystallogr* 2004;60(Pt 12 Pt 1):2126–2132. [PubMed: 15572765]
24. Sing T, Sander O, Beerenwinkel N, Lengauer T. *Bioinformatics* 2005;21(20):3940–3941. [PubMed: 16096348]

25. Yde CW, Ermakova I, Issinger O-G, Niefind K. J Mol Biol 2005;347(2):399–414. [PubMed: 15740749]

Fig. 1.
The workflow for identifying unknown ligands using *SurfaceAID*

Fig. 2.

The overall performance of *SurfaceAID* is tested against a non-redundant subset of ligand binding surfaces from the PDB. The first ranking of the native ligand from a query surface is shown for the entire dataset. For example, the native ligand was the top ranking ligand for 72% of the query surfaces

Fig. 3.
The ATP binding surface from protein kinase CK2 from *Z. mays* (PDB id=1a6o) shown in canonical reference frame **(a)**. The convex hull of the surface **(b)** provides an approximated volume of the cavity and it constructed from the Delaunay triangulation of the surface **(c)**

Fig. 4. A density blob was found near a surface pocket on putative mannosephosphate isomerase from *A. fulgidus* (a). Fructose was built in to the final model (c) based on similarity to the binding surface from exo-inulinase from *A. awamori* (b, cyan). Electron density is contoured at 1σ

Fig. 5. NADP is modeled into a putative Δ^1 -pyrroline-5-carboxylate reductase (**a**, **b**) based on surface similarity to binding pocket from carbonyl reductase sniffer from *D. melanogaster* (**c**). Electron density is contoured at 1σ

Fig. 6.

A candidate set of potential ligands is suggested by *Surface_{AID}* for a density blob (contoured at 1σ) in protein of unknown function HP0184 from *H. pylori* (a). The binding surfaces of GTP of gamma-tubulin from *H. sapiens* (b), GDP of mannosylglycerate synthase from *R. marinus* (c), dGTP of nucleoside diphosphate kinase II from *A. thaliana* (d), and GmTP of oncogene protein from *H. sapiens* (e) are aligned to the query surface (white). The ligands are all posed into the surface (f) for evaluation. dGTP is ultimately modeled into the final structure (g).

Table 1
The most frequently observed ligands bound to proteins in the Protein Data Bank

Frequency rank	Ligand name	Residue id	Count in PDB	Molecular weight
1	Sulfate ion	SO4	3895	96.06
2	Protoporphyrin IX containing Fe	HEM	1577	616.16
3	<i>n</i> -Acetyl-D-glucosamine	NAG	1523	221.13
4	Glycerol	GOL	1513	92.06
5	Alpha-D-mannose	MAN	622	180.10
6	Flavin-adenine dinucleotide	FAD	556	785.29
7	Adenosine-5'-diphosphate	ADP	484	427.10
8	Nicotinamide-adenine-dinucleotide	NAD	459	663.22
9	Ethylene glycol	EDO	410	62.05
10	Pyridoxal-5'-phosphate	PLP	405	247.06
11	Glucose	GLC	388	180.10
12	Amino group	NH2	377	16.20
13	Acetyl group	ACE	341	43.02
14	(4S)-2-Methyl-2,4-pentanediol	MPD	296	118.12
15	Adenosine-5 <i>p</i> -triphosphate	ATP	289	507.08

Structures with multiples of a single ligand are only counted once

Table 2

The most frequently observed ligands in the Protein Data Bank

Ligand	PDB id	Resolution (Å)	AUC	Native rank
SO4	1gci	0.78	0.55	2
SO4	1ccu	2.25	0.58	3
SO4	2r6e	5.02	0.58	1
HEM	1x8p	0.85	0.79	1
HEM	1dd7	2.25	0.94	1
HEM	3a0h	4.00	0.67	1
NAG	1k5c	0.96	0.70	1
NAG	1z6i	2.50	0.69	3
NAG	1h3y	4.10	0.61	1
GOL	1gci	0.78	0.58	17
GOL	1dm8	2.25	0.56	1
GOL	3dmk	4.19	0.54	4
MAN	2gud	0.94	0.56	1
MAN	1tb6	2.50	0.57	1
MAN	1j89	4.10	0.57	7
FAD	1n4w	0.92	0.97	1
FAD	1ffv	2.25	0.98	1
FAD	1bhy	4.18	0.98	1
ADP	2cnq	1.00	0.88	1
ADP	1r6b	2.25	0.88	1
ADP	2d3c	3.81	0.86	1
NAD	2jhf	1.00	0.96	1
NAD	2b4r	2.25	0.96	1
NAD	1wdm	3.80	0.96	1
EDO	2vb1	0.65	0.59	1
EDO	1m51	2.25	0.54	1
EDO	3fnt	3.30	0.62	1
PLP	1w23	1.08	0.95	1
PLP	1ekv	2.25	0.95	1
PLP	3tat	3.50	0.94	4
GLC	2fvy	0.92	0.53	1
GLC	1qha	2.25	0.51	9
GLC	1acz	3.22	0.43	9
NH2	1hje	0.75	0.85	2
NH2	1b0q	2.25	0.81	1
NH2	2ns3	3.40	0.75	6
ACE	3a11	0.75	0.68	1
ACE	1jcp	2.25	0.73	1
ACE	2i35	3.80	0.69	9
MPD	3a11	0.75	0.52	1

Ligand	PDB id	Resolution (Å)	AUC	Native rank
MPD	ldlu	2.25	0.57	1
MPD	lg1r	3.40	0.56	1
ATP	1xdn	1.20	0.91	5
ATP	1xmi	2.25	0.87	3
ATP	3cjc	3.90	0.88	1