BMC Bioinformatics

RESEARCH ARTICLE Open Access

Knowledge-based annotation of small molecule binding sites in proteins

Ratna R Thangudu, Manoj Tyagi, Benjamin A Shoemaker, Stephen H Bryant, Anna R Panchenko* and Thomas Madej*

Abstract

Background: The study of protein-small molecule interactions is vital for understanding protein function and for practical applications in drug discovery. To benefit from the rapidly increasing structural data, it is essential to improve the tools that enable large scale binding site prediction with greater emphasis on their biological validity.

Results: We have developed a new method for the annotation of protein-small molecule binding sites, using inference by homology, which allows us to extend annotation onto protein sequences without experimental data available. To ensure biological relevance of binding sites, our method clusters similar binding sites found in homologous protein structures based on their sequence and structure conservation. Binding sites which appear evolutionarily conserved among non-redundant sets of homologous proteins are given higher priority. After binding sites are clustered, position specific score matrices (PSSMs) are constructed from the corresponding binding site alignments. Together with other measures, the PSSMs are subsequently used to rank binding sites to assess how well they match the query and to better gauge their biological relevance. The method also facilitates a succinct and informative representation of observed and inferred binding sites from homologs with known three-dimensional structures, thereby providing the means to analyze conservation and diversity of binding modes. Furthermore, the chemical properties of small molecules bound to the inferred binding sites can be used as a starting point in small molecule virtual screening. The method was validated by comparison to other binding site prediction methods and to a collection of manually curated binding site annotations. We show that our method achieves a sensitivity of 72% at predicting biologically relevant binding sites and can accurately discriminate those sites that bind biological small molecules from non-biological ones.

Conclusions: A new algorithm has been developed to predict binding sites with high accuracy in terms of their biological validity. It also provides a common platform for function prediction, knowledge-based docking and for small molecule virtual screening. The method can be applied even for a query sequence without structure. The method is available at http://www.ncbi.nlm.nih.gov/Structure/ibis/ibis.cgi.

Background

The physical interactions between proteins and other molecules in protein crystal structures provide crucial insights into protein function. It is precisely these structures that enable researchers to study interactions in atomic detail, and find out, for example, how a specific mutation in a protein affects its function, or how a few atom modifications in a small molecule might lead to a more effective drug. With the large number of available crystal structures (nearly 60,000 currently in the RCSB Protein Data Bank), it is of great importance to improve the tools available for study of these interactions.

Moreover, a powerful method of inference can be used to predict function and interactions. It is based on the observation that homologous proteins have similar functions and often interact with their small molecules in a similar manner. Thus it is possible to infer protein-small molecule interactions even if there are no crystal structures available for a particular protein of interest, as long as there are structures of sufficiently close homologs. Recent estimates suggest that the majority of Entrez Protein sequences have homologs with a known structure [1,2], thereby providing a reasonable chance to find relevant interactions via structures for protein sequences.

* Correspondence: panch@ncbi.nlm.nih.gov, madej@ncbi.nlm.nih.gov
National Center for Biotechnology Information, 8600 Rockville Pike, Building 38A, Bethesda, MD 20894 USA
Full list of author information is available at the end of the article

Homology inference methods, although powerful, have certain limitations. Common descent does not necessarily imply similarity in function or interactions; and annotations transferred from one protein to a homolog may result in incorrect functional or interolog assignment at larger evolutionary distances [3-6]. To verify and guide annotations, it is often essential to ensure close evolutionary relationships, and at the same time characterize the details of interactions in terms of binding site similarity. Current binding site prediction methods can be subdivided into several major categories: those which use evolutionary conservation of binding site motifs [7-9], those which use information about a structure of a complex [10-12], and docking and other methods [13,14]. Structure-based methods use detailed knowledge of the protein structure to identify binding sites on the basis of the physico-chemical properties of individual residues, their electrostatic contribution, and their location in the 3D structure [15-26].

A number of methods and servers have been developed for predicting protein function by identifying similarities in sequence and structural features of binding pockets in homologous proteins, or evolutionary constraints on residues [27], or by using threading and other approaches [20,28-39]. The main goal of these methods is to provide functional annotation for proteins out to the most distant homology relationships. *FINDSITE* [40], for example, looks for structural templates with bound small molecules for a query protein using threading. The templates are superimposed and the centers of mass of the bound small molecules are clustered to annotate putative binding sites on the query. Threading based methods, although capable of recognizing distant functional relations, are limited by the complexity of model building and low reliability of function transfer associated with distant homology [41,42].

Firestar [31] predicts functionally important residues based on PSI-BLAST [43] alignments between the query sequence and structures with functional information derived from the PDB and the Catalytic Site Atlas [44]. *PHUNCTIONER* [20] uses sequence profiles based on clustered sequences with matching GO [45] terms; potential binding sites are detected from sequence conservation. This method is capable of inferring the location of highly conserved small molecule binding sites, but might be questionable if the conservation of sites is caused by factors other than binding.

Transitive annotation of small molecule binding sites is also possible by detection of functional domains in the query protein sequence through BLAST heuristics and mapping the functionally important residues and/or features from the domain family members [30,46].

There are a few other methods that directly detect small molecule binding sites via geometric analysis of protein structures. These methods include LIGSITE^{csc} [29], CAST [47], PASS [48], SURFNET [49], SCREEN [50], and ConCavity [51]. All of these algorithms attempt to identify solvent-accessible pockets formed by surface residues on the protein, and to rank those pockets (for example by volume), in order to assign the most highly ranked pockets as the predicted/putative small molecule binding sites. LIGSITE^{csc}, SURFNET, and ConCavity use a more complex ranking function that takes into account residue conservation of binding site residues. These geometric methods are reasonably accurate, achieving success rates of 60-70% in correctly identifying small molecule binding sites. In their evaluation of LIGSITE^{csc}, the authors showed that their algorithm outperformed the other three methods on a test set of 48 structures [29]. The SCREEN method identifies binding sites geometrically, and also computes feature vectors that are used by machine learning techniques. SCREEN is included in a suite of powerful modeling tools for functional annotation [52]

Recently we have developed a new database and method called "IBIS" (Inferred Biomolecular Interaction Server [53], http://www.ncbi.nlm.nih.gov/Structure/ibis/ibis.html) which enables researchers to conveniently study biomolecular interactions that have been observed in protein structures and through inference by homology to formulate predictions/hypotheses for biomolecular interactions, even if the data for specific biomolecules is not available. Therefore, IBIS can be considered a resource for functional annotation of proteins that have relevant homologs in the PDB [54]. An input protein sequence may or may not have a structure itself; if not, it is assigned to the most closely related structure(s) using BLAST. IBIS can identify and infer a protein's interaction partners together with the locations of the corresponding binding sites on the protein query. It provides annotations of binding sites for proteins, small molecules (chemicals), nucleic acids, peptides and ions. In this paper we describe the method used in IBIS to annotate protein-small molecule interactions. To ensure biological relevance of binding sites, IBIS clusters similar binding sites found in homologous proteins based on conservation of sequence and structure of the binding site residues. Binding sites which appear evolutionarily conserved among non-redundant sets of homologous proteins are given higher priority. Additionally, binding site clusters are validated by comparing them with available binding site annotations from a manually curated subset of the CDD database [55,56], and sites with non-biological small molecules are excluded. After binding sites are clustered, position specific score matrices (PSSMs) are constructed from the corresponding binding site alignments. Together with other measures, the PSSMs are subsequently used to rank binding sites to assess how well they match the query, and to gauge the biological relevance of binding sites with respect to the query.

A critical difference between our method and others is that IBIS pays particular attention to ensuring the biological relevance of binding sites, and homology between the unknown query sequence and the known structures of protein complexes. Our method might miss some remote similarities which could be detectable, for example by FINDSITE, but in exchange IBIS's top ranked annotations should be considered highly reliable. Unlike other methods, IBIS does not filter out similar structures to speed up the search process, but accounts for all structures so that interesting small molecule binding complexes are easily accessible. Our method derives the actual binding sites from observed structures, and groups them to account for variations in the binding site residues due to differences in small molecule size and conformations. This is essential for proteins which are important drug targets, as they have often been co-crystallized with a great variety of inhibitors. The clustering (grouping) of binding sites by similarity is very important because it identifies the distinct binding modes and allows for an easier interpretation of the results, despite the great growth in the amount of structure data over the last several years. As we have shown, it is possible to do the clustering automatically and in a biologically meaningful way.

Results

Annotation of protein chains with observed and inferred binding sites

There are about 28000 PDB entries with observed small molecule binding sites and about 56000 protein chains. The total number of observed small molecule binding sites is about 91000 and approximately 67000 of these represent biologically relevant small molecules(i.e. around 24000 small molecules represent crystallization agents). Small molecule binding is a specific feature that plays a crucial role in the protein function. About 64% of protein chains in the PDB bind to a single small molecule and 95% bind to no more than four small molecules (Figure 1a). Likewise, the binding site pockets are rather small compared to the size of their functional domains. The binding sites are usually less than 25 residues and 55% of the binding sites in the current study are smaller than 10 residues (Figure 1b). Our algorithm inferred binding sites for 92000 protein chains and the overall average number of binding site clusters inferred per chain is 6.5 (Figure 1c) whereas the average number of biologically relevant binding site clusters inferred per chain is about 4.

One of the important features of this method is that it does not exclude redundant sequences bound to different small molecules. For example, to account for all specific interactions of various drugs targeting the Kinase ATP binding site, it is imperative to consider all the protein sequences even if they are identical.

We validated the IBIS method by comparing the obtained annotations to the manually curated CDD annotations and to other different methods which use geometry of binding pockets and/or sequence conservation of binding sites. It should be mentioned that since the IBIS method is based on different types of structural evidence, the notion of false positives might not be valid in many cases.

Validation of the IBIS method using the Conserved Domain Database

To test the ability of our method to successfully infer the biologically relevant binding sites, a validation procedure was implemented using the manually curated Conserved Domain Database (CDD) [56] alignments and the functional features recorded in it as a standard of truth. Manually curated functional site annotations in CDD have been extracted from the published literature or derived from manual interpretation of individual three-dimensional structures. Altogether 49% of the proteins with observed small molecule binding sites have CDD small molecule binding site conserved annotations whereas over 55% of the proteins with inferred binding sites have at least one site overlapping with CDD annotated binding site annotation.

In our analysis we used the CDD release 2.16 containing 4092 protein chains. We chose representative protein chains purged at the 25% sequence identity level. In this jackknifing experiment, the query protein and its identical homologs are omitted from clustering. Altogether 486 representative chains had at least one structurally similar non-identical homolog which had observed protein-small molecule binding sites. Figure 2a shows how well our method can retrieve the CDD annotated binding sites at the top ranks by calculating the fraction of true positives (sensitivity) or percentage of correctly annotated binding sites (overlap between CDD and IBIS annotated binding sites should be at least 50%). For 207 of these there was only one inferred binding site (cluster) detected, and by default these will always be ranked first. There remain 279 examples which have at least two IBIS binding sites, 209 (75%) of these were ranked first, and 49 were ranked second, so that 258 (92%) were ranked either first or second (Figure 2b).

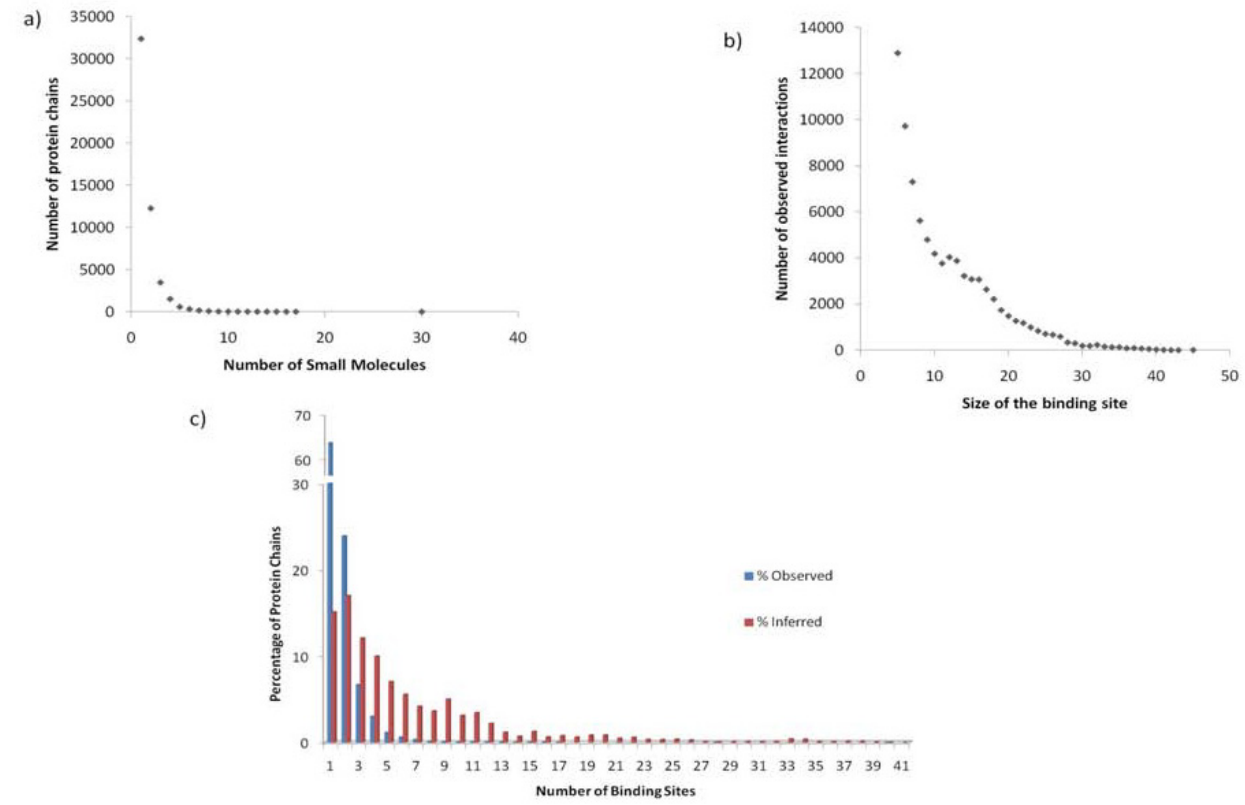


Figure 1 Statistics of small molecules and their binding sites observed in protein structure complexes. a) Number of small molecules and binding sites observed per protein chain, b) size of the observed binding sites, c) histogram showing the number of observed and inferred binding sites with plotted versus the fraction (%) of protein chains having these sites.

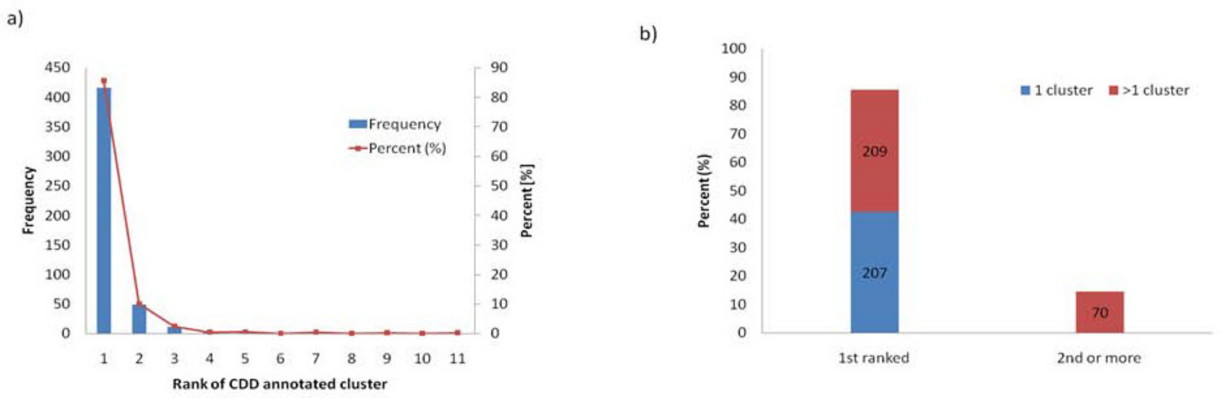


Figure 2 Biological validity of the IBIS inferred binding sites. a) Histogram showing the frequency of protein chains as function of their biological relevancy as suggested by overlap of the inferred binding sites with CDD conserved feature annotation. b) Percentage of proteins with their inferred sites having their 1st and 2nd rank clusters with CD annotations; 165 proteins have only one predicted site.

Since there are a number of proteins which do not have CDD annotations, IBIS inferred binding sites may be biologically relevant in these cases.

Validation of ranking scheme: discriminating between biological and non-biological chemicals

We used the same set of protein queries (604 chains) to evaluate our method using structures which contained both biological and non-biological small molecules (see Additional file 1 Table S1). Our goal is to assess how well our ranking scheme distinguishes between the two groups of binding sites: those containing biological versus non-biological small molecules. If all the bound small molecules in an inferred binding site are non-biological, it is deemed as non-biological. To address this, we applied a linear discriminant analysis which constructs a discriminant function that divides the parameter space into regions so as to separate the groups as distinctly as possible. The method computes the posterior probability of group membership for each observation, and assigns the observation to the group that has the highest probability. As a result, a classification matrix is produced, which gives the fraction of observations correctly assigned to each group by the discriminant function. In our case, a good classification would be quantified by high fractions for both correctly predicted biological binding site clusters and correctly predicted non-biological binding site clusters. We found that our method correctly classifies 87% of biological clusters and 85% of non-biological clusters.

Validation of IBIS method by comparison with - geometric methods

To further validate the prediction ability of our method we compared it with several widely used geometry and energy-based approaches discussed in a recent study [57] which includes LIGSITEcsc [29], PASS [48], Q-SiteFinder [28], Surfnet [49]. We used 44 out of 48 proteins from this paper which have structure homologs with at least 30% sequence identity and also have both small molecule-bound and unbound structures.

Table 1: Prediction sensitivity (%) of the top three predictions by different geometric approaches and their comparison to IBIS.

Method*	Top1	Top2	Top3
$IBIS_{100}$	73	89	89
$IBIS_{90}$	75	91	91
$IBIS_{80}$	72	88	88
LIGSITEcs	71	79	85
PASS	58	67	75
Q-SiteFinder	52	60	75
SURFNET	42	58	62

*$IBIS_{100}$, $IBIS_{90}$, $IBIS_{80}$ dataset contains unbound query proteins for which the average sequence identity between the unbound protein and members of the binding site clusters containing the bound homolog is no more than 100%, 90% and 80% respectively.

For each method tested, the top ranked predicted sites for the unbound structure are compared with the observed binding sites in the bound structure of a protein-small molecule complex of a homolog. Table 1 shows the sensitivity of retrieval of the true observed sites at the top three ranks. To measure the sensitivity of retrieval of bound structures at different levels of similarity between the unbound query and bound structure from the database, we selected from the test set only those unbound-bound pairs which are within a given similarity range (no more than 80, 90, or 100% identity) and denoted them $IBIS_{80}$, $IBIS_{90}$ and $IBIS_{100}$. For example, the $IBIS_{90}$ dataset contains unbound query proteins for which the average sequence identity between the unbound protein and members of the binding site clusters containing the bound homolog is no more than 90%. It is difficult if not impossible to define false positives in our case since there are many binding site clusters which are biologically relevant (for example have a significant overlap with the manually curated CDD functional annotations) but at the same time do not match the binding site of the bound form of the protein from the test set. This happens if, for example, there are multiple binding sites/pockets in the protein which bind different small molecules and have distinct functions. As can be seen from this table IBIS performance is similar to the LIGSITE^{cs} method which uses sequence conservation and reaches about 72% sensitivity. Overall we found that a total of 31 proteins (70%) from the test set have at least one of their IBIS predicted sites overlap with CDD binding site annotation. This suggests that IBIS successfully uses the knowledge of the structure complexes of homologs to predict and rank the relevant sites. The complete details of the prediction results can be seen in Additional file 1 Table S2.

All of these approaches, although they perform reasonably well, are limited by the requirement of differentiating true positives from false positives. Introducing sequence conservation need not necessarily improve the prediction accuracy and could be a source of error, leading to over prediction of the binding site area [58]. IBIS on the other hand predicts only a handful of small molecule binding sites with high probability of being biologically relevant. On average our method predicts 4 'biologically relevant' binding sites per protein chain and over half of all predicted sites map to CDD curator annotations.

Knowledge-based docking using IBIS, an example

To demonstrate the effectiveness of IBIS as a knowledge-based prediction system, we compared our method with an established reverse docking approach. Cai and coworkers [59] employed a reverse docking method to find a potential target protein for a natural product: *N-trans*-caffeoyltyramin in the genome of *Helicobacter pyroli*. Initially all potential binding proteins of *N-trans*-caffeoyltyramin were screened from a database of potential drug targets with known structures from the Protein Data Bank using the reverse docking approach TarFisDock [60]. Only two proteins from the *H.pyroli* genome were found by the TarFisDock method: diaminopimelate decarboxylase (DC) and Peptide deformylase (PDF). After enzymatic validation, only the PDF protein was found to be a probable drug target. The crystal structure complex of *N-trans*-caffeoyltyramin with PDF suggested a highly selective binding in the PDF binding pocket.

We attempted to identify the binding sites of *N-trans*-caffeoyltyramin on the PDF protein sequence. The closest homolog for *H.pyroli* PDF is *P.aeruginosa* PDF which has 45% sequence identity and has been used as a template for inferring the interactions by IBIS. The top ranked and highly conserved inferred binding site of P.aeruginosa PDF when mapped onto H.pyroli PDF is in complete agreement with the native/experimentally determined binding site of the *N-trans*-caffeoyltyramin - PDF complex (Figure 3).

Discussion

A researcher interested in the function of a specific protein will usually be concerned not only with the availability of any functional annotation, but also with the reliability of such information. The most reliable source is experimental data on the protein function but despite the growth of the protein sequence and structure databases, there remains only a small fraction of proteins whose functions have been experimentally characterized. In this paper we present a method which provides the information on protein function annotation through the identification of protein binding sites. The current approach attempts to interlink sequence conservation with structural diversity in deciphering protein function. We specifically focus on protein small molecule binding sites and their biological relevance for protein function. Our method derives the actual binding sites from the structures of all the homologs and groups them based on sequence and structural similarity. For example, to account for all specific interactions of various drugs targeting the Kinase ATP binding site (see Additional file 1 Table S3), it is imperative to consider all the protein sequences even if some of them are identical. Such grouping ensures their biological relevance and at the same

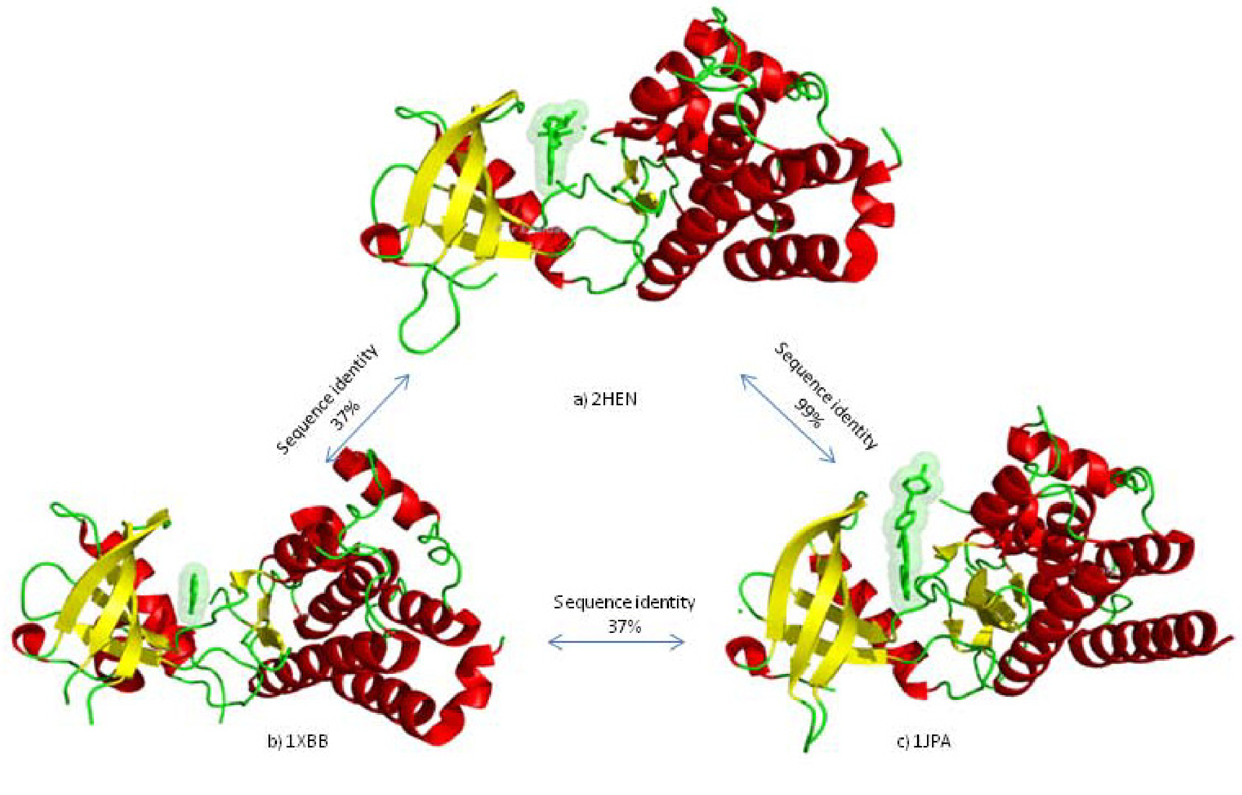
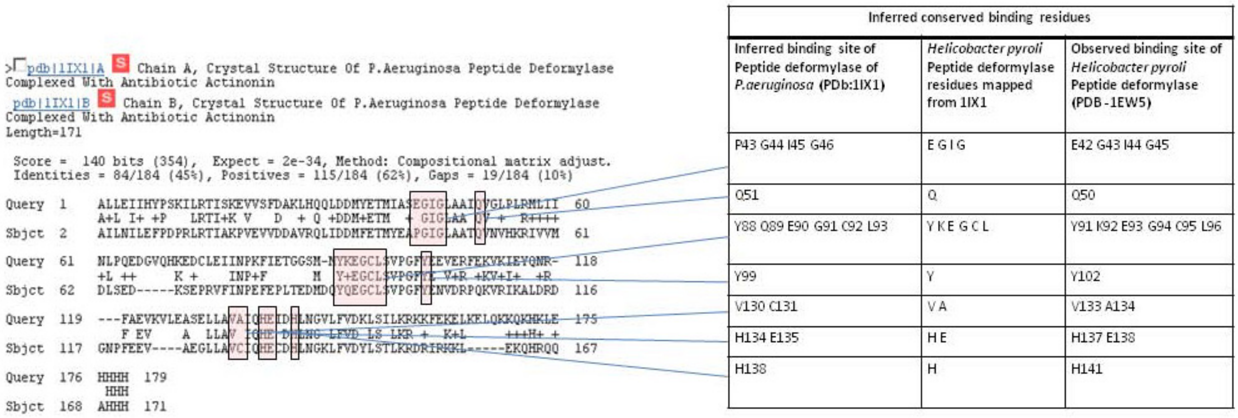


Figure 3 Tyrosine kinase homologoues with varying degrees of sequence conservation with different small molecules in their ATP-binding pocket. (a) Ephb2 Receptor Kinase domain with ADP. (b) Syk Tyrosine Kinase Domain in complex with Gleevec. (c) Ephb2 Receptor Tyrosine Kinase with Adenine.



```
>pdb|1IX1|A S Chain A, Crystal Structure Of P.Aeruginosa Peptide Deformylase
Complexed With Antibiotic Actinonin
 pdb|1IX1|B S Chain B, Crystal Structure Of P.Aeruginosa Peptide Deformylase
Complexed With Antibiotic Actinonin
Length=171

 Score =  140 bits (354),  Expect = 2e-34, Method: Compositional matrix adjust.
 Identities = 84/184 (45%), Positives = 115/184 (62%), Gaps = 19/184 (10%)

Query  1    ALLEIIHYPSKILRTISKEVVSFDAKLHQQLDDMYETMIASEGIGLAAIQVGLPLRMLII  60
            A+L I+ +P   LRTI+K V   D  + Q +DDM+ETM  + GIGLAA QV +  R++++
Sbjct  2    AILNILEFPDPRLRTIAKPVEVVDDAVRQLIDDMFETMYEAPGIGLAATQVNVHKRIVVM  61

Query  61   NLPQEDGVQHKEDCLEIINPKFIETGGSM-NYKEGCLSVPGFYEEVERFEKVKIEYQNR-  118
            +L ++     K +    INP+F      M  Y+EGCLSVPGFYE V+R +KV+I+  +R
Sbjct  62   DLSED-----KSEPRVFINPEFEPLTEDMDQYQEGCLSVPGFYENVDRPQKVRIKALDRD  116

Query  119  ---FAEVKVLEASELLAVAIQHEIDHLNGVLFVDKLSILKRKKFEKELKELQKKQKHKLE  175
               F EV    A  LLAV IQHE DHLNG LFVD LS LKR +  K+L     +++H+ +
Sbjct  117  GNPFEEV----AEGLLAVCIQHECDHLNGKLFVDYLSTLKRDRIRKKL-----EKQHRQQ  167

Query  176  HHHH  179
             HHH
Sbjct  168  AHHH  171
```

Inferred conserved binding residues		
Inferred binding site of Peptide deformylase of *P.aeruginosa* (PDb:1IX1)	*Helicobacter pyroli* Peptide deformylase residues mapped from 1IX1	Observed binding site of *Helicobacter pyroli* Peptide deformylase (PDB -1EW5)
P43 G44 I45 G46	E G I G	E42 G43 I44 G45
Q51	Q	Q50
Y88 Q89 E90 G91 C92 L93	Y K E G C L	Y91 K92 E93 G94 C95 L96
Y99	Y	Y102
V130 C131	V A	V133 A134
H134 E135	H E	H137 E138
H138	H	H141

Figure 4 Mapping of the inferred binding site. Inferred binding site of peptide deformylase *P.aeruginosa* (PDB:1IX1) mapped onto the sequence of *Helicobacter pyroli* and its agreement with the observed binding site in *N-trans*-caffeoyltyramin-PDF complex (PDB: 1EW5). MMDB residue numbering is used which starts from the beginning of the corresponding GenBank protein sequence.

time accounts for variations in the binding site residues due to differences in small molecule sizes and conformations. By using all available structures of close homologs, IBIS provides a great opportunity for analyzing the diversity of binding modes. Figure 4, for example, shows the conserved tyrosine kinase fold with varying degrees of sequence similarity but sharing a highly conserved ATP binding site occupied by different small molecules.

Recently, it was estimated that over two-thirds of all protein sequences in the GenBank database have at least one structure homolog [1,2]. As the on-going structural genomics initiative continues to close the sequence-structure gap, our method might be very useful for annotating proteins with unknown function and structure. Moreover, the location of putative binding sites provides guidance for the protein docking methods for drug design. We have assessed the reliability of our method by direct comparison with the binding site annotations from literature and manual curation and have shown that in the great majority of cases, the method detects and ranks the manually annotated binding site cluster at the first or second rank. This is achievable for a number of reasons, such as using a sufficient level of similarity between the unknown query and its homologs with the known binding sites, accurate clustering of small molecule binding sites using a reasonable similarity measure, and applying a deliberately designed ranking scheme that distinguishes the non-biological from the biologically relevant binding sites.

We have also compared our method with several widely used geometry and energy-based approaches to predict small molecule binding sites. As we have shown, the performance of our prediction method is very similar to popular geometric approaches. Moreover, one of the advantages is that our method can be applied even for a query sequence without structure, which is not the case for those binding site prediction methods which explicitly rely on the specific features of binding pocket geometry.

Using remote homology for functional inference is often based on the general assumption that there is a negative correlation between small molecule binding site similarity and overall sequence similarity. However, small molecule binding site similarity is much more complicated with many examples of strikingly similar binding sites with low (<30%) overall sequence identity and also very weakly similar binding sites with high overall sequence identity [61]. Likewise, the similarities of small molecule binding sites across different protein folds, although providing new insights, leads to new challenges in deciphering the functional relevance. Large-scale automated function prediction methods are often limited by the lack of sufficient understanding of biological function and also by the quality of structure data. Hence, through the IBIS approach, we strive to limit the false positive rate by employing a conservative sequence similarity threshold of at least 30% over the structurally superimposed regions of homologs. It is often possible that the protein-small molecule crystal state may correspond to a global minimum of free energy where biologically relevant interactions are difficult to distinguish from non-specific contacts. For example, a recent estimate suggests some 20% of dimeric structures in PDB may be crystallization artifacts [62]. The elaborate scoring scheme of our method based on recurrence and evolutionary conservation, along with the list of non-biological small molecules, tends to de-emphasize the artifactual interactions and ranks such sites near or at the bottom of the list.

Furthermore, the chemical properties of small molecules bound to the inferred binding sites can be used as a starting step in small molecule virtual screening. The PubChem compound database [63] mapping of IBIS small molecules accomplishes a preliminary step in small molecule virtual screening by clustering the similar chemicals into structurally unique compounds. The functional groups of the small molecules binding in a common binding site of evolutionarily related proteins are likely conserved. Recently it was shown that sequence and structure conservation of the binding site residues contacting these anchor functional groups is significantly higher than those contacting variable regions [40]. IBIS, thus provides a common platform for function prediction, knowledge-based docking and also for small molecule virtual screening.

Conclusions

Finding small molecule binding sites that specify protein function is of great importance in drug development. Here we proposed a method to decipher the function of an unknown protein by interlinking sequence conservation with structural diversity of its homologs. To facilitate validation of the inferred binding sites from homologs, we developed an elaborate scoring scheme that can accurately distinguish biologically relevant sites. The method has been implemented as a web server, IBIS (Inferred Biomolecular Interaction Server - http://www.ncbi.nlm.nih.gov/Structure/ibis/ibis.cgi) to facilitate accurate, efficient and high-throughput function prediction.

Methods

We used the NCBI Molecular Modeling Database (MMDB) [64] as a source of data on protein complexes. The automated MMDB processing of PDB files includes steps such as deposition of the protein sequences into GenBank [65], deposition of small molecules into PubChem [63], addition of corresponding links to these databases in the MMDB records, also links to citations and references in PubMed, and Entrez indexing for quick searching.

Below we describe different steps of processing, including defining observed interactions from structures, inferred interactions from homologs, clustering binding sites and their ranking in terms of biological relevance with respect to the query protein.

Defining observed interactions

In the current release of the Molecular Modeling Database (MMDB) [64], there are about 28000 entries with bound small molecules. The resulting 39000 small molecules are bound to about 56000 protein chains in total. A small molecule is defined as any non-polypeptide, non-nucleic acid molecule in the structure complex or any molecule with a sufficient number of non-standard amino acid/nucleic acids and without an assigned GenBank identifier from NCBI. All the small molecules are standardized in the PubChem database [63] and have valid substance and compound identifiers. In this work we do not consider small molecules that are smaller than 5 heavy atoms or those having molecular weight outside the range of 70-800 Da. Small molecules such as metal ions often play role as crystallization agents, and therefore ions are not considered in this paper.

The filters by atom count and molecular weight only partially remove non-biological small molecules (i.e. buffers, salts, detergents, solvents, and ions added for the purpose of crystallization and/or purification). These non-biological molecules sometimes mimic natural small molecules and tend to bind in functional/active sites of proteins. For validation purposes we used a list of potential non-biological small molecules which has been collected from the literature (see Additional file 1 Table S1) [30,66,67].

We define a protein residue to be in contact with a small molecule if there is at least one (heavy) atom of the residue within 4.0Å of some atom from the small molecule. For most pairs of atoms, this threshold corresponds to the sum of their van der Waals radii plus a tolerance of about 0.5Å to allow for coordinate errors in structure determination. For manual curation of the Conserved Domain Database (CDD) a similar contact definition is used for defining protein-small molecule contacts. We retain only those protein-small molecule complexes which have at least five interacting protein residues. We define *"binding site"* as a set of residues on a given protein chain which are in contact with a given small molecule. Each MMDB entry is analyzed, and all pairs of biomolecules consisting of a protein chain and small molecule in contact with that chain are retained for further analysis.

It should be mentioned that a small molecule can be bound to a single domain or multiple domains which could come from more than one protein chain in the PDB record. Almost half of all the small molecules in the PDB are bound by more than one domain with <75% of all contacts to any single domain [66]. However, using domains as structural units would necessitate automatic domain decomposition methods in many cases [68,69], and the domain boundaries chosen could affect the results. To circumvent the potential technical difficulties in using domains as the structural unit in recording the observed/physical interactions, we use only complete protein chains for defining protein-small molecule interactions. Small molecules binding to multiple protein chains entail even more technical difficulties. For example, simultaneous superposition of multiple chains would need to be checked to ensure similarity of binding sites. Therefore, when multiple chains are involved in a binding

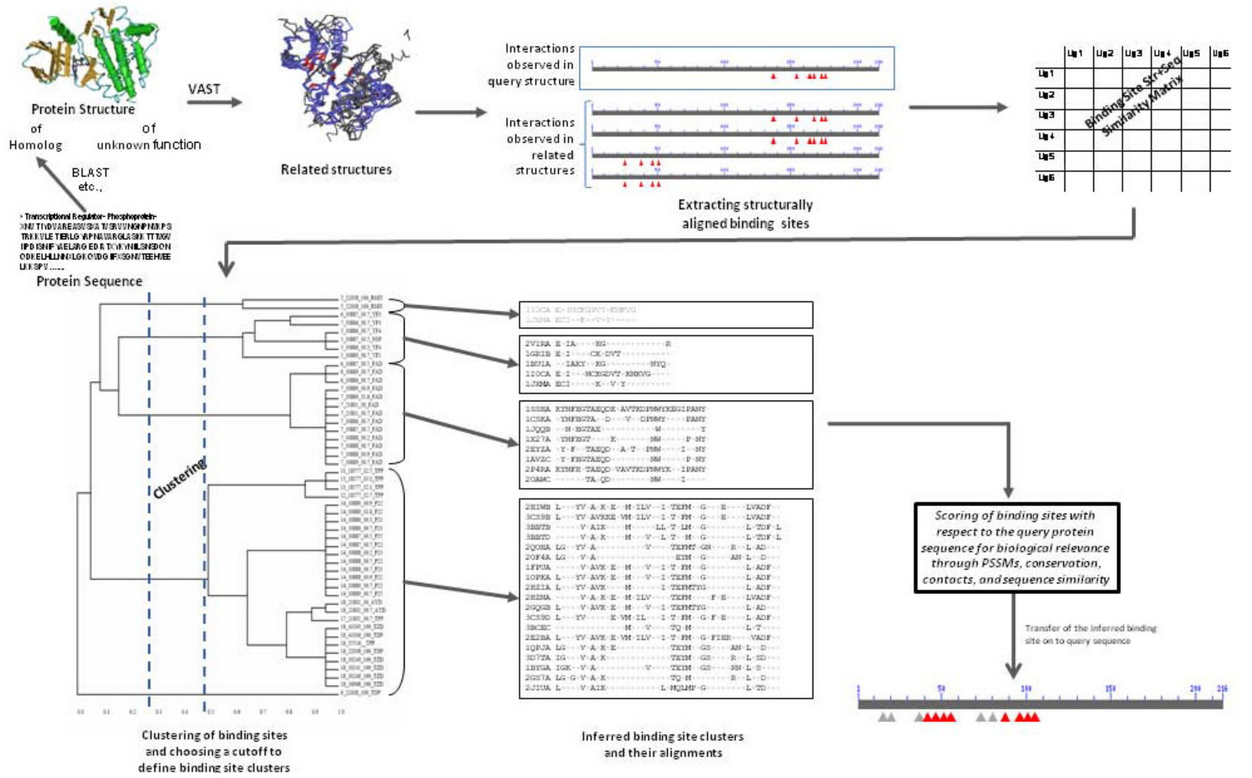

VAST
Protein Structure
of
Homolog
of
unknown function
BLAST
etc.,
> Transcriptional Regulator- Phosphoprotein-
Protein Sequence
Related structures
Interactions
observed in
query structure
Interactions
observed in
related
structures
Extracting structurally
aligned binding sites
Binding Site Str+Seq Similarity Matrix
Clustering
Clustering of binding sites
and choosing a cutoff to
define binding site clusters
Inferred binding site clusters
and their alignments
Scoring of binding sites with
respect to the query protein
sequence for biological relevance
through PSSMs, conservation,
contacts, and sequence similarity
Transfer of the inferred binding
site on to query sequence

Figure 5 Overview of the IBIS binding site annotation procedure.

site, if one of the chains includes 75% or more of the contacts, then we define only one binding site and assign it only to that particular chain. Otherwise, we define separate binding sites on each of the chains. The latter situation is relatively rare as only about 15% of the proteins in the current PDB release have small molecule interactions that fall into this category.

Inferring interactions from homologs

To ensure the biological relevance of binding sites, they are clustered and their sequence and structural similarity is assessed. An overview of the process is shown in Figure 5. Here are the important details concerning the main steps in the processing.

1) Collecting homologs with bound small molecules

To infer interactions based on homology we collect proteins which are structurally similar to a given query protein and have at least 30% sequence identity to the query (we refer to them as "homologous structure neighbors"). Structure neighbors for all PDB/MMDB entries have been pre-computed by the VAST algorithm [70] and stored in the PubVast database. Then we retrieve observed interactions for all structure neighbors (including the query protein). No sequence redundancy filter is applied to remove structures because there are often many structures of the same protein with different bound small molecules, and we may wish to study any of these cases. Since the alignments may contain gaps, we retain only those instances where at least 75% of the residue contacts with the small molecule occur within the structure alignment footprint of the query and neighbor.

2) Measuring binding site similarity

In order to cluster the binding sites of the homologous structure neighbors, it is necessary to construct their alignment and define a similarity measure. We construct the alignment between the structure neighbors A and B by composing the alignment from structure neighbor A to the query, with the alignment from the query to structure neighbor B. It is necessary to construct this alignment by composing through the query, because oftentimes the neighbors A and B will be more closely related to each other than to the query, in which case the "direct" alignment between them will be more extensive than the one through the query, and so it could include binding sites or interface residues that are not relevant to the query. To capture the similarity of the binding sites, the similarity measure includes both the structural equivalency and sequence similarity. The similarity score between two positions i and j of two binding sites is defined as:

$$S_{ij} = H(a_i, a_j)\Delta_{ij} + \phi\Delta_{ij} + w(1 - \Delta_{ij}) \qquad (1)$$

where *H* is the element of the BLOSUM62 matrix corresponding to the aligned amino acids in positions *i* and *j*; Δ_{ij} is equal to 1 if two positions are aligned and 0 otherwise. $\boldsymbol{\theta}$ is an additional weight of "+1" added to each structurally equivalent position. *w* is a gap penalty of "-4", to mimic the most unfavorable substitution score from BLOSUM62 matrix, which showed the best performance in our preliminary studies. The overall similarity score between two binding sites is calculated by summing up S_{ij} over all positions in the gapped alignment. To facilitate comparison of scores from different alignments, the raw score is converted to a bit score with the statistical parameters λ and *K* previously defined in the BLOSUM scoring system.

$$S' = \frac{\lambda S - \ln K}{\ln 2} \qquad (2)$$

The similarity score is then converted into a conservation score *CS* by dividing by the maximum of the bit scores when the binding sites are scored against themselves.

$$\mathrm{CS} = \frac{S(A,B)}{Max\left(S(A,A)' S(B,B) \right)} \qquad (3)$$

3) Clustering of binding sites

Based on the calculated conservation score *CS*, the binding sites of the homologs are clustered using a complete-linkage clustering algorithm, which considers the distance between two clusters to be equal to the maximum distance between their members. A distance cutoff value to define the clusters is chosen using a free energy function defined previously. This function *F* is formulated to maximize the mean similarity of members within a cluster and minimize the complexity of the description provided by cluster membership [71].

$$F = \frac{1}{N} \left\{ \begin{array}{l} \sum_{C} \frac{1}{|C|} \sum_{i,j \in C} S(i,j) \\ +T \sum_{C} |C| \log |C| \\ -TN \log N \end{array} \right\} \qquad (4)$$

where *T* is the temperature factor, *S(i, j)* is the similarity score between binding site *i* and binding site *j* in each cluster, *C* represents a cluster, |*C*| is the number of binding sites in the cluster *C*, and *N* is the total number of binding sites clustered. The temperature *T* is a parameter (constant) that is chosen so as to correctly balance the energy-like and entropy-like terms in the function [71].

Biological relevance of binding sites and their ranking with respect to the query protein

All binding site clusters are ranked in terms of their predicted biological relevance and similarity to the query. First we assess the evolutionary conservation of binding site clusters. Those sites which reoccur in diverse enough protein complexes are ranked higher. Clusters that have only one non-redundant member (after members with more than 90% identity are removed) are considered "singletons" and are not assigned any score (ranked at the bottom of the list). A "conservation score" is computed in order to measure the diversity of cluster members and how well the binding site is conserved across the homologs. To do this, positional conservation in the binding site multiple sequence alignment is calculated using the Shannon entropy measure with the Henikoff-Henikoff sequence weights. Sequence weights are estimated using the complete sequences of neighbors aligned with the query protein.

To account for evolutionary closeness of a given binding site cluster to the query we use the sequence-PSSM score and the average sequence identity between the query and all cluster members calculated over the whole structure-structure alignment (not just binding sites). A position specific score matrix (PSSM) is constructed based on the binding site multiple alignment using the implicit pseudo-count method of Gribskov, McLachlan and Eisenberg [72]. The aligned binding site region of the query protein is then scored against the PSSM and a sequence-PSSM score is calculated. A higher sequence-PSSM score points to a higher probability of this site being a biologically relevant site for the query.

To rank the larger interfaces more highly we also calculate the average number of interfacial contacts which the binding site makes in the complex of the corresponding homolog. All components of the ranking score are then normalized and all clusters are ranked with respect to the Z-scores. Any cluster with all members binding non-biological small molecules is disregarded.

The Z-score is calculated for each of these four corresponding terms (i.e. conservation score, PSSM-score, contact count, and percent sequence identity to query) in the ranking scheme by subtracting the mean value and dividing by the standard deviation obtained from the score distribution of other binding site clusters for a given query protein. The coefficients in front of each term in the ranking score were calculated empirically. The combined score is designed to rank the most biologically relevant sites at the top.

$$Z_{comb} = \left\{ \left(0.4 * Z_{pssm} \right) + \left(0.4 * Z_{conserv} \right) + \left(0.1 * Z_{contact} \right) + \left(0.1 * Z_{pcnt} \right) \right\} \qquad (5)$$

Additional material

Additional file 1 Table S1: The most common non-biological small molecules found in protein structure complexes. Table S2: Summary of the IBIS predictions and CDD annotation validation for the 44 bound and unbound structures used as test set to compare with existing geometric approaches. Table S3: Variety of small molecules binding in the ATP binding pocket of tyrosine kinase homologs.

Authors' contributions

ARP, BAS, SHB, and TM conceived the project. RRT, MT, and BAS implemented the database and analysis programs. RRT, ARP, and TM wrote the paper. All authors contributed to the underlying ideas of the method and the analysis. All authors read and approved the final manuscript.

Acknowledgements

The authors are grateful to Aron Märchler-Bauer, Dachuan Zhang, and Jessica Fong. This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

Author Details

National Center for Biotechnology Information, 8600 Rockville Pike, Building 38A, Bethesda, MD 20894 USA

Received: 16 March 2010 Accepted: 1 July 2010
Published: 1 July 2010

References

1. Wang Y, Addess KJ, Chen J, Geer LY, He J, He S, Lu S, Madej T, Marchler-Bauer A, Thiessen PA, *et al.*: **MMDB: annotating protein sequences with Entrez's 3D-structure database.** *Nucleic Acids Res* 2007, 35(Database issue):D298-300.
2. Fukuchi S, Homma K, Sakamoto S, Sugawara H, Tateno Y, Gojobori T, Nishikawa K: **The GTOP database in 2009: updated content and novel features to expand and deepen insights into protein structures and functions.** *Nucleic Acids Res* 2009, 37(Database issue):D333-337.
3. Bork P, Koonin EV: **Predicting functions from protein sequences--where are the bottlenecks?** *Nat Genet* 1998, **18(4):**313-318.
4. Gerlt JA, Babbitt PC: **Can sequence determine function?** *Genome Biol* 2000, **1(5):**REVIEWS0005.
5. Hegyi H, Gerstein M: **The relationship between protein structure and function: a comprehensive survey with application to the yeast genome.** *J Mol Biol* 1999, **288(1):**147-164.
6. Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JD, Bertin N, Chung S, Vidal M, Gerstein M: **Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs.** *Genome Res* 2004, 14(6):1107-1118.
7. Capra JA, Singh M: **Predicting functionally important residues from sequence conservation.** *Bioinformatics* 2007, **23(15):**1875-1882.
8. Zhang T, Zhang H, Chen K, Shen S, Ruan J, Kurgan L: **Accurate sequence-based prediction of catalytic residues.** *Bioinformatics* 2008, **24(20):**2329-2338.
9. Fischer JD, Mayer CE, Soding J: **Prediction of protein functional residues from sequence by probability density estimation.** *Bioinformatics* 2008, **24(5):**613-620.
10. Burgoyne NJ, Jackson RM: **Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces.** *Bioinformatics* 2006, **22(11):**1335-1342.
11. Ota M, Kinoshita K, Nishikawa K: **Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation.** *J Mol Biol* 2003, **327(5):**1053-1064.
12. Liang S, Zhang C, Liu S, Zhou Y: **Protein binding site prediction using an empirical scoring function.** *Nucleic Acids Res* 2006, **34(13):**3698-3707.
13. Campbell SJ, Gold ND, Jackson RM, Westhead DR: **Ligand binding: functional site location, similarity and docking.** *Curr Opin Struct Biol* 2003, **13(3):**389-395.
14. Thibert B, Bredesen DE, del Rio G: **Improved prediction of critical residues for protein function based on network and phylogenetic analyses.** *BMC Bioinformatics* 2005, **6:**213.
15. Bray T, Chan P, Bougouffa S, Greaves R, Doig AJ, Warwicker J: **SitesIdentify: a protein functional site prediction tool.** *BMC Bioinformatics* 2009, **10(1):**379.
16. Brylinski M, Prymula K, Jurkowski W, Kochanczyk M, Stawowczyk E, Konieczny L, Roterman I: **Prediction of functional sites based on the fuzzy oil drop model.** *PLoS Comput Biol* 2007, **3(5):**e94.
17. Brylinski M, Skolnick J: **A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation.** *Proc Natl Acad Sci USA* 2008, **105(1):**129-134.
18. Jones S, Thornton JM: **Analysis of protein-protein interaction sites using surface patches.** *J Mol Biol* 1997, **272(1):**121-132.
19. Landgraf R, Xenarios I, Eisenberg D: **Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins.** *J Mol Biol* 2001, **307(5):**1487-1502.
20. Pazos F, Sternberg MJ: **Automated prediction of protein function and detection of functional sites from structure.** *Proc Natl Acad Sci USA* 2004, **101(41):**14754-14759.
21. 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.
22. Panchenko AR, Kondrashov F, Bryant S: **Prediction of functional sites by analysis of sequence and structure conservation.** *Protein Sci* 2004, **13(4):**884-892.
23. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM: **Analysis of catalytic residues in enzyme active sites.** *J Mol Biol* 2002, **324(1):**105-121.
24. Bate P, Warwicker J: **Enzyme/non-enzyme discrimination and prediction of enzyme active site location using charge-based methods.** *J Mol Biol* 2004, **340(2):**263-276.
25. Greaves R, Warwicker J: **Active site identification through geometry-based and sequence profile-based calculations: burial of catalytic clefts.** *J Mol Biol* 2005, **349(3):**547-557.
26. Marti-Renom MA, Rossi A, Al-Shahrour F, Davis FP, Pieper U, Dopazo J, Sali A: **The AnnoLite and AnnoLyze programs for comparative annotation of protein structures.** *BMC Bioinformatics* 2007, **8(Suppl 4):**S4.
27. Chelliah V, Chen L, Blundell TL, Lovell SC: **Distinguishing structural and functional restraints in evolution in order to identify interaction sites.** *J Mol Biol* 2004, **342(5):**1487-1504.
28. Laurie AT, Jackson RM: **Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites.** *Bioinformatics* 2005, **21(9):**1908-1916.
29. Huang B, Schroeder M: **LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation.** *BMC Struct Biol* 2006, **6:**19.
30. Snyder KA, Feldman HJ, Dumontier M, Salama JJ, Hogue CW: **Domain-based small molecule binding site annotation.** *BMC Bioinformatics* 2006, **7:**152.
31. Lopez G, Valencia A, Tress ML: **firestar--prediction of functionally important residues using structural templates and alignment reliability.** *Nucleic Acids Res* 2007:W573-577.
32. Qin S, Zhou HX: **meta-PPISP: a meta web server for protein-protein interaction site prediction.** *Bioinformatics* 2007, **23(24):**3386-3387.
33. Skolnick J, Brylinski M: **FINDSITE: a combined evolution/structure-based approach to protein function prediction.** *Brief Bioinform* 2009.
34. Hernandez M, Ghersi D, Sanchez R: **SITEHOUND-web: a server for ligand binding site identification in protein structures.** *Nucleic Acids Res* 2009:W413-416.
35. Talavera D, Laskowski RA, Thornton JM: **WSsas: a web service for the annotation of functional residues through structural homologues.** *Bioinformatics* 2009, **25(9):**1192-1194.
36. Ivanisenko VA, Pintus SS, Grigorovich DA, Kolchanov NA: **PDBSiteScan: a program for searching for active, binding and posttranslational modification sites in the 3 D structures of proteins.** *Nucleic Acids Res* 2004:W549-554.
37. Chang DT, Weng YZ, Lin JH, Hwang MJ, Oyang YJ: **Protemot: prediction of protein binding sites with automatically extracted geometrical templates.** *Nucleic Acids Res* 2006:W303-309.
38. Jambon M, Andrieu O, Combet C, Deleage G, Delfaud F, Geourjon C: **The SuMo server: 3 D search for protein functional sites.** *Bioinformatics* 2005, **21(20):**3929-3930.
39. Shulman-Peleg A, Nussinov R, Wolfson HJ: **SiteEngines: recognition and comparison of binding sites and protein-protein interfaces.** *Nucleic Acids Res* 2005:W337-341.
40. Brylinski M, Skolnick J: **FINDSITE: a threading-based approach to ligand homology modeling.** *PLoS Comput Biol* 2009, **5(6):**e1000405.
41. 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.

42. Rost B: **Enzyme function less conserved than anticipated.** *J Mol Biol* 2002, **318(2)**:595-608.
43. 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)**:3389-3402.
44. Porter CT, Bartlett GJ, Thornton JM: **The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data.** *Nucleic Acids Res* 2004, 32(Database issue):D129-133.
45. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, *et al.*: **The Gene Ontology (GO) database and informatics resource.** *Nucleic Acids Res* 2004, 32(Database issue):D258-261.
46. Marchler-Bauer A, Bryant SH: **CD-Search: protein domain annotations on the fly.** *Nucleic Acids Res* 2004:W327-331.
47. 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.
48. 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.
49. Laskowski RA: **SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions.** *J Mol Graph* 1995, **13(5)**:323-330. 307-328
50. Nayal M, Honig B: **On the nature of cavities on protein surfaces: application to the identification of drug-binding sites.** *Proteins* 2006, **63(4)**:892-906.
51. Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA: **Predicting Protein Ligand Binding Sites by Combining Evolutionary Sequence Conservation and 3 D Structure.** *PLoS Computational Biology* 2009. accepted for publication
52. Petrey D, Fischer M, Honig B: **Structural relationships among proteins with different global topologies and their implications for function annotation strategies.** *Proc Natl Acad Sci USA* 2009, **106(41)**:17377-17382.
53. Shoemaker BA, Zhang D, Thangudu RR, Tyagi M, Fong JH, Marchler-Bauer A, Bryant SH, Madej T, Panchenko AR: **Inferred Biomolecular Interaction Server--a web server to analyze and predict protein interacting partners and binding sites.** *Nucl Acids Res* 2010, 38(Database issue):D518-524.
54. Berman H, Henrick K, Nakamura H, Markley JL: **The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data.** *Nucleic Acids Res* 2007, 35(Database issue):D301-303.
55. Marchler-Bauer A, Anderson JB, DeWeese-Scott C, Fedorova ND, Geer LY, He S, Hurwitz DI, Jackson JD, Jacobs AR, Lanczycki CJ, *et al.*: **CDD: a curated Entrez database of conserved domain alignments.** *Nucleic Acids Res* 2003, **31(1)**:383-387.
56. Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, *et al.*: **CDD: specific functional annotation with the Conserved Domain Database.** *Nucleic Acids Res* 2009, 37(Database issue):D205-210.
57. Huang B: **MetaPocket: a meta approach to improve protein ligand binding site prediction.** *OMICS* 2009, **13(4)**:325-330.
58. Wass MN, Sternberg JEM: **Prediction of ligand binding sites using homologous structures and conservation at CASP8.** *Proteins: Structure, Function, and Bioinformatics* 2009, **77(S9)**:147-151.
59. Cai J, Han C, Hu T, Zhang J, Wu D, Wang F, Liu Y, Ding J, Chen K, Yue J, *et al.*: **Peptide deformylase is a potential target for anti-Helicobacter pylori drugs: reverse docking, enzymatic assay, and X-ray crystallography validation.** *Protein Sci* 2006, **15(9)**:2071-2081.
60. Li H, Gao Z, Kang L, Zhang H, Yang K, Yu K, Luo X, Zhu W, Chen K, Shen J, *et al.*: **TarFisDock: a web server for identifying drug targets with docking approach.** *Nucleic Acids Res* 2006:W219-224.
61. Kinjo AR, Nakamura H: **Comprehensive structural classification of ligand-binding motifs in proteins.** *Structure* 2009, **17(2)**:234-246.
62. Krissinel E: **Crystal contacts as nature's docking solutions.** *J Comput Chem* **31(1)**:133-143.
63. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH: **PubChem: a public information system for analyzing bioactivities of small molecules.** *Nucleic Acids Res* 2009:W623-633.
64. Chen J, Anderson JB, DeWeese-Scott C, Fedorova ND, Geer LY, He S, Hurwitz DI, Jackson JD, Jacobs AR, Lanczycki CJ, *et al.*: **MMDB: Entrez's 3D-structure database.** *Nucleic Acids Res* 2003, **31(1)**:474-477.
65. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, *et al.*: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2009, 37(Database issue):D5-15.
66. Bashton M, Nobeli I, Thornton JM: **Cognate ligand domain mapping for enzymes.** *J Mol Biol* 2006, **364(4)**:836-852.
67. Wang R, Fang X, Lu Y, Yang CY, Wang S: **The PDBbind database: methodologies and updates.** *J Med Chem* 2005, **48(12)**:4111-4119.
68. Koczyk G, Berezovsky IN: **Domain Hierarchy and closed Loops (DHcL): a server for exploring hierarchy of protein domain structure.** *Nucleic Acids Res* 2008:W239-245.
69. Holland TA, Veretnik S, Shindyalov IN, Bourne PE: **Partitioning protein structures into domains: why is it so difficult?** *J Mol Biol* 2006, **361(3)**:562-590.
70. Madej T, Gibrat JF, Bryant SH: **Threading a database of protein cores.** *Proteins* 1995, **23(3)**:356-369.
71. Slonim N, Atwal GS, Tkacik G, Bialek W: **Information-based clustering.** *Proc Natl Acad Sci USA* 2005, **102(51)**:18297-18302.
72. Gribskov M, McLachlan AD, Eisenberg D: **Profile analysis: detection of distantly related proteins.** *Proc Natl Acad Sci USA* 1987, **84(13)**:4355-4358.

doi: 10.1186/1471-2105-11-365
Cite this article as: Thangudu *et al.*, Knowledge-based annotation of small molecule binding sites in proteins *BMC Bioinformatics* 2010, **11**:365