D1096–D1103 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 18 October 2012*
doi:10.1093/nar/gks966

BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions

Jianyi Yang, Ambrish Roy and Yang Zhang*

Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, MI 48109-2218, USA

Received July 12, 2012; Revised July 17, 2012; Accepted September 25, 2012

ABSTRACT

BioLiP (http://zhanglab.ccmb.med.umich.edu/BioLiP/) is a semi-manually curated database for biologically relevant ligand–protein interactions. Establishing interactions between protein and biologically relevant ligands is an important step toward understanding the protein functions. Most ligand-binding sites prediction methods use the protein structures from the Protein Data Bank (PDB) as templates. However, not all ligands present in the PDB are biologically relevant, as small molecules are often used as additives for solving the protein structures. To facilitate template-based ligand–protein docking, virtual ligand screening and protein function annotations, we develop a hierarchical procedure for assessing the biological relevance of ligands present in the PDB structures, which involves a four-step biological feature filtering followed by careful manual verifications. This procedure is used for BioLiP construction. Each entry in BioLiP contains annotations on: ligand-binding residues, ligand-binding affinity, catalytic sites, Enzyme Commission numbers, Gene Ontology terms and cross-links to the other databases. In addition, to facilitate the use of BioLiP for function annotation of uncharacterized proteins, a new consensus-based algorithm COACH is developed to predict ligand-binding sites from protein sequence or using 3D structure. The BioLiP database is updated weekly and the current release contains 204 223 entries.

INTRODUCTION

With the advancement in structural biology and the structural genomics initiatives, the structural repertoire in Protein Data Bank (PDB) (1) is growing rapidly. The total number of solved proteins in the PDB is >80 000, doubling the number of the entries in 2006. Nevertheless, the biological functions for many of these proteins are largely unknown. Since proteins perform their biological functions by interacting with other molecules, establishing the interaction between proteins and ligand molecules is an important step toward understanding the biological functions. In particular, the experimental solutions on the ligand–protein complexes are often used as template to deduce the ligand–protein docking and functional annotation information of other uncharacterized proteins.

Due to the large number of additives used during the procedure of protein purification and/or crystallization, evaluating the biological relevance of ligands present in the PDB structure is a non-trivial problem (2–7). One of the most direct ways to assess the biological relevance of a ligand is by manual verifications, such as reading original literature and examining the annotations across different databases. However, given the growing number of protein entries in the PDB, such manual checking is becoming increasingly infeasible. Efforts have been made to develop automatic procedures to select biologically relevant ligands from the PDB library. For instance, the FireDB database selects ligands based on a mapping between inorganic ligands and Gene Ontology (GO) Annotations (2). The inorganic ligands that are biologically relevant can be missed in FireDB if there are no GO annotations or if there is no mapping for these ligands. LigASite is a ligand-binding site database for benchmarking use, which is very small because it selects ligands using very strict requirements (3), e.g. ligands are selected only when they have >10 heavy atoms and have >70 inter-atomic contacts with the proteins. These requirements may miss true biological ligands (e.g. metal ions). Binding MOAD (5) is a ligand-binding affinity database that selects ligands based on a combination of automated procedure and manual validation. Binding MOAD excludes small DNA/RNA molecules and metal ions, which are in fact important ligand molecules in many proteins (8,9). PDBbind (4) is another ligand-binding affinity database that has less strict requirements than Binding MOAD (e.g. lower structure resolution, inclusion

*To whom correspondence should be addressed. Tel: +1 734 6471549; Fax: +1 734 6156553; Email: zhng@umich.edu of DNA/RNA molecules and peptides). BindingDB (6) is a database that collects binding data directly from scientific literatures. It contains now 620 000 binding data for 5500 proteins and >270 000 drug-like molecules but only 1659 proteins can be unambiguously referenced to the PDB with a 100% sequence identity. For a ligand–protein complex, when no binding affinity data are reported in the literature, the complex is excluded from the PDBbind and BindingDB databases. There are some other databases related to the current study, such as Relibase (10) (http://relibase.rutgers.edu) for ligand–protein interactions, Pocketome (11) (http://pocketome.org/) for druggable binding sites, ProtChemSI (12) (http://pcidb.russelllab.org/) for the network of protein–chemical structural interactions, PepX (13) (http://pepx.switchlab.org/) for protein–peptide interactions and RsiteDB (14) (http://bioinfo3d.cs.tau.ac.il/RsiteDB/) for protein–RNA interactions. Overall, most of the existing databases, although contain very useful information required for specific studies, have missed many biologically relevant ligand–protein interactions that are important for reliable and accurate template-based ligand–protein docking, virtual ligand screening and protein function annotations (7,15–20).

In this article, we aim to construct a comprehensive database of biologically relevant ligand–protein interactions collected from the PDB. To have a precise assessment of the biological relevance of the ligand entries in our database, both computational and manual examinations are performed during the database construction. Each entry in the BioLiP contains a comprehensive list of annotations on: ligand-binding residues, ligand-binding affinity, catalytic site residues, Enzyme Commission numbers, GO terms and cross-links to other popular databases. In addition, to annotate the function of uncharacterized proteins using the BioLiP database, we have developed a new algorithm COACH to predict ligand-binding sites from either protein sequence or 3D structure.

MATERIALS AND METHODS

Procedure for database construction

BioLiP database is constructed using known protein structures in the PDB. The overall procedure of the database construction consists of three major steps:

Step 1: For each entry in the PDB, the 3D structure is downloaded and the modified residues (i.e. residues modified post-translationally, enzymatically, or by design) are translated to standard residues based on the record 'MODRES' in the PDB structure file, which contains the information on the precursor standard residue name for each modified residue. The PDB files and the function annotations in BioLiP are provided with the original PDB residue numbering system since the relative residue positions in the PDB record contain useful information such as disordered fragments/loops, artificial peptide tags, fusion constructs and residue insertions. However, the PDB numbering system is diverse and usually contains gaps/insertions, which makes it difficult to use in computational programming, such as protein fold-recognition, ligand–protein docking and ligand-binding sites prediction experiments. For the convenience of computational approaches, for each entry BioLiP also provides a downloadable version of the renumbered structures with the residues being continuous and starting from 1. For each protein chain (called receptor), the information (if any) to be collected includes the following: (i) ligand-binding affinity from manual survey of the original literature and the existing databases of Binding MOAD (5), PDBbind (4) and BindingDB (6); (ii) catalytic site residues mapped from the Catalytic Site Atlas (21); (iii) annotated EC numbers in the COMPND records; GO terms (22) from the GO Annotation database (23); (iv) UniProt accession code (24) mapped from the SIFTS project (25) and (v) the PubMed abstract of the primary literature citation in the 'JRNL' record. The PubMed abstract is used later to assess the biological relevance of a ligand.

Step 2: Ligands, which are defined as small molecules, are extracted from the PDB file. Three types of ligand molecules are collected in the BioLiP database: the molecules from the 'HETATM' record (excluding water and modified residues), small DNA/RNA and peptides with <30 residues. For molecules from the 'HETATM' record, atoms having identical chain ID and sequence number are put into a group (called HET-group). The name of a HET-group ligand is set to be the residue name at the columns 18–20 of the structure file, which in most cases is a three-letter code (e.g. ATP, ADP, FMB) of the mmCIF format (http://mmcif.rcsb.org/). If a HET-group ligand contains multiple different residue names, it is regarded as a *k*-mer ligand. To avoid conflicts with the existing three-letter code, we name *k*-mer, DNA/RNA and peptide ligands as UUU, NUC and III, respectively. The information can be viewed in the BioLiP webpage by hovering mouse over the corresponding ligand codes.

Metal ions are considered as potential biologically relevant ligands in BioLiP. Most metal ions, such as sodium ion, are first listed as possible artifacts. The PubMed abstract is then used to determine whether they are biologically relevant ligands or crystallization artifacts, since metal ions often play different roles in different protein molecules. For example, the sodium ion in the protein 'human parathyroid hormone 1–34' (PDB ID: 1ET1) is regarded as crystallization artifact, which can be further verified by reading the crystallization paper (PubMed ID: 10837469). However, the sodium ion in the protein 'tetragonal hen egg-white lysozyme' (PDB ID: 193L) is assessed as biologically relevant in BioLiP, with its biological role ('stabilizing the loop Ser60-Leu75') described in the PubMed abstract (PubMed ID: 15299672). Currently, there are ~63 500 BioLiP entries containing metal ions, which are the second largest number of entries in BioLiP (see 'BioLiP in Numbers' section).

Step 3: Each ligand molecule is submitted to a composite automated and manual procedure (Figure 1) to decide its biological relevance. If the ligand molecule is evaluated as biologically relevant, its interaction with the receptor (i.e. binding site residues in the receptor) is deposited into the BioLiP database. Additionally, the ligand-binding affinity, catalytic site residues, EC numbers, GO terms and the cross-links to the PDB, UniProt, PDBsum, PDBe and PubMed databases are also collected and deposited into BioLiP.

Assessment of biological relevance

An accurate annotation of the biological relevance of the ligand entries is essential to the BioLiP data collection. A ligand molecule present in a target protein is considered as biologically relevant if it interacts with the protein and plays certain biological roles, such as inhibitor, activator and substrate analog (3,7). To guarantee the high accuracy and speed, we developed a composite automated and manual procedure as outlined in Figure 1. First, an automated four-step hierarchical procedure is used to verify the biological relevance of a ligand. After the automated procedure is completed, a careful manual check is performed to eliminate possible false positives, which can occur for entries with the commonly used crystallization additives.

To speed up the annotation procedure as well as increase the accuracy, we manually pre-collected a set of 353 suspiciously non-biological ligands, which are frequently used for the protein structure determination (including crystallization additives, non-biological ions, heavy metal and so on.) To generate this list, we first collected all ligands that are observed for >20 times in known protein structures. This list was refined further by analyzing the possible biological role of these ligands, e.g. a ligand is removed from the list if it is found to have biological relevance in the related literature of the structure file or is present in the KEGG database (26). This list is used to help assess the biological relevance of each ligand in PDB automatically (Figure 1) and is available at http://zhanglab.ccmb.med.umich.edu/BioLiP/ligand_list.

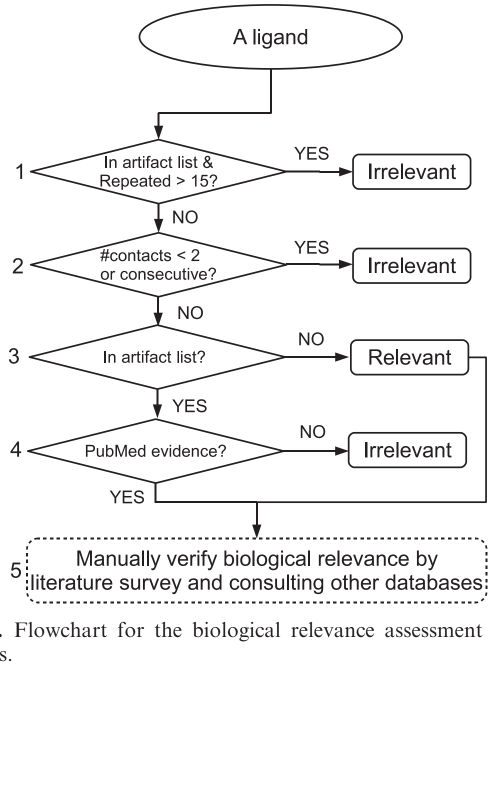

A ligand
1
In artifact list & Repeated > 15?
YES
Irrelevant
NO
2
#contacts < 2 or consecutive?
YES
Irrelevant
NO
3
In artifact list?
NO
Relevant
YES
4
PubMed evidence?
NO
Irrelevant
YES
5
Manually verify biological relevance by literature survey and consulting other databases

Figure 1. Flowchart for the biological relevance assessment of ligand molecules.

The automated filtering procedure consists of four steps:

First, if the candidate ligand is in the artifact list and appears >15 times in the same structure file, then it is likely to be crystallization additive and is considered as biologically irrelevant.

Second, the contacts between the receptor and ligand atoms are computed. The record 'REMARK 350' in the asymmetric unit files is used to exclude crystallization neighbors. This record presents which chains of the structure should be put together and the mathematical transformations (i.e. rotation and translation matrices) operated on each chain to generate biomolecules (i.e. biological unit files). The contacts between two chains are evaluated only when both chains are used to generate a biomolecule. For a receptor residue, if the closest atomic distance between the residue and the ligand is within certain distance cutoff, then the residue is defined as a ligand-binding site residue. The cutoff is set to be 0.5 plus the sum of the Van der Waal's radius of the two atoms under investigation (7). If the number of binding site residues (i.e. number of contacts) is less than two or all the binding site residues are consecutive, it is deemed to be biologically irrelevant because most biological relevant ligands are usually tethered by multiple residues, which are further apart in the sequence space.

Third, if the ligand is not present in the artifact list, then it is considered as biologically relevant and kept in the pipeline for further manual verifications.

Fourth, the PubMed abstract is used to filter out biologically irrelevant ligands. If the ligand is in the artifact list, the simplest way is to treat it as biologically irrelevant and discard it. But this will miss some ligands (false negatives) that are indeed biologically relevant in some cases. For instance, the ligand molecule 'glycerol' (with ligand ID 'GOL') is one of the most frequently used crystallization additives and it is thus regarded as biologically irrelevant by many existing databases. However, this ligand can have a biological role in some proteins. For example, the ligand molecule glycerol binds to the protein 'enzyme diol dehydratase' (PDB ID: 3AUJ) with binding affinity $K_m = 1.2 \pm 0.02\,\text{mM}$ with its biological role described as 'glycerol is bound to the substrate binding site in the $(\beta/\alpha)_8$ or TIM barrel of the diol dehydratase α subunit' in (27). Thus, this ligand is considered as biologically relevant for this protein and added to BioLiP. We found that if a ligand present in a protein has its relevant biological role, it is often mentioned in the PubMed abstract. Based on such observation, we propose to use the PubMed abstract as an additional filter. To this end, the chemical names/synonyms of the ligand (curated from ChEBI, PubChem and PDB databases) are compared with the PubMed abstract. If there is no hit in this comparison procedure, the ligand is deemed to be biologically irrelevant. Otherwise, the ligand is possible to be biologically relevant, which remains to be verified by hand in the next step.

Finally, the manual verification is performed to check for suspicious or ambiguous entries, which are referred to those entries related with the commonly used crystallization additives, such as glycerol, ethanol, methanol, 2-propanol, ethylene glycol, hexylene glycol and polyethylene glycol. Ligands filtered from the above four steps can sometimes still be false positives, which is usually caused by unexpected match between the ligand names/synonyms and the PubMed abstract. In the same example of the ligand ‘glycerol’, it has the synonym ‘glycyl alcohol’, which leads to an unexpected match of the term ‘alcohol’ for the protein ‘arylesterase’ (PDB ID: 3HI4). Therefore, manual verification for ligands that are commonly used as crystallization additives is necessary to ensure the quality of BioLiP. Currently, we do this manual verification mainly by reading the original literatures and consulting other secondary databases. In the current version of BioLiP, manual verifications helped us to remove ∼12 500 entries that were false positives and we added ∼3000 entries that would have been missed by using the automated procedure alone.

RESULTS

BioLiP in numbers

By the time this article was submitted, BioLiP contains 204 223 annotated high-quality ligand–protein interactions, involving 50 621 proteins from the PDB. Among the annotated ligand molecules, 9076 are DNA/RNA ligands, 9849 are peptide ligands, 10 511 are *k*-mer ligands, 63 470 are metal ligands and 111 317 are regular ligands (i.e. the common small-molecule ligands). The pie chart for the ligand distribution in BioLiP database is shown in Figure 2. In total, 20 013 entries have binding affinity data, with 10 445 from Binding MOAD, 13 579 from PDBbind, 7179 from Binding DB and 62 from manual survey of the original literature.

Web interface

The BioLiP database is freely accessible at http://zhanglab.ccmb.med.umich.edu/BioLiP/ with four basic interfaces: BROWSE, SEARCH, DOWNLOAD and COACH. They are introduced below in details.

BROWSE and SEARCH BioLiP

Three different browsing options are provided under the ‘BROWSE’ interfaces in BioLiP: (i) Browse all entries; (ii) Browse all ligands and (iii) Browse all binding affinities. Clicking the ‘Browse all entries’ displays the summary of all entries in BioLiP database in the form of a table, which lists nine major components for each ligand–protein interaction site: BioLiP ID, PDB ID, Binding site number, Ligand ID, EC number, GO terms, UniProt ID, PubMed link and Binding affinity.

The detailed information for each ligand–protein interaction site is available by clicking the corresponding BioLiP ID. A screenshot of the ligand–protein interaction page is given in Figure 3. The information is organized into five sections: Receptor Information, Ligand-Binding/Catalytic Sites, Enzyme Commission, GO and External Links. In the Ligand-Binding/Catalytic Sites section, the ligand-binding information is provided first, including mmCIF formatted ligand identifier, ligand chemical name/synonym, ligand 2D visualization and ligand-binding affinity. Second, the ligand–protein interaction is displayed (in both global and local views) using the Jmol Applet (http://www.jmol.org/). The structures of the

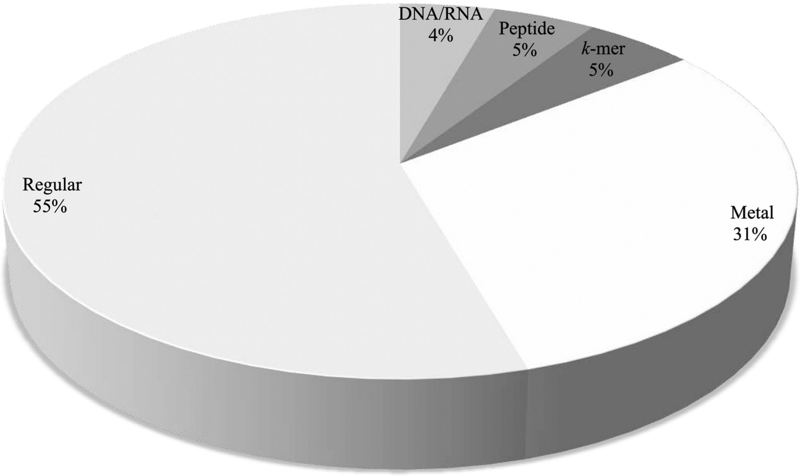
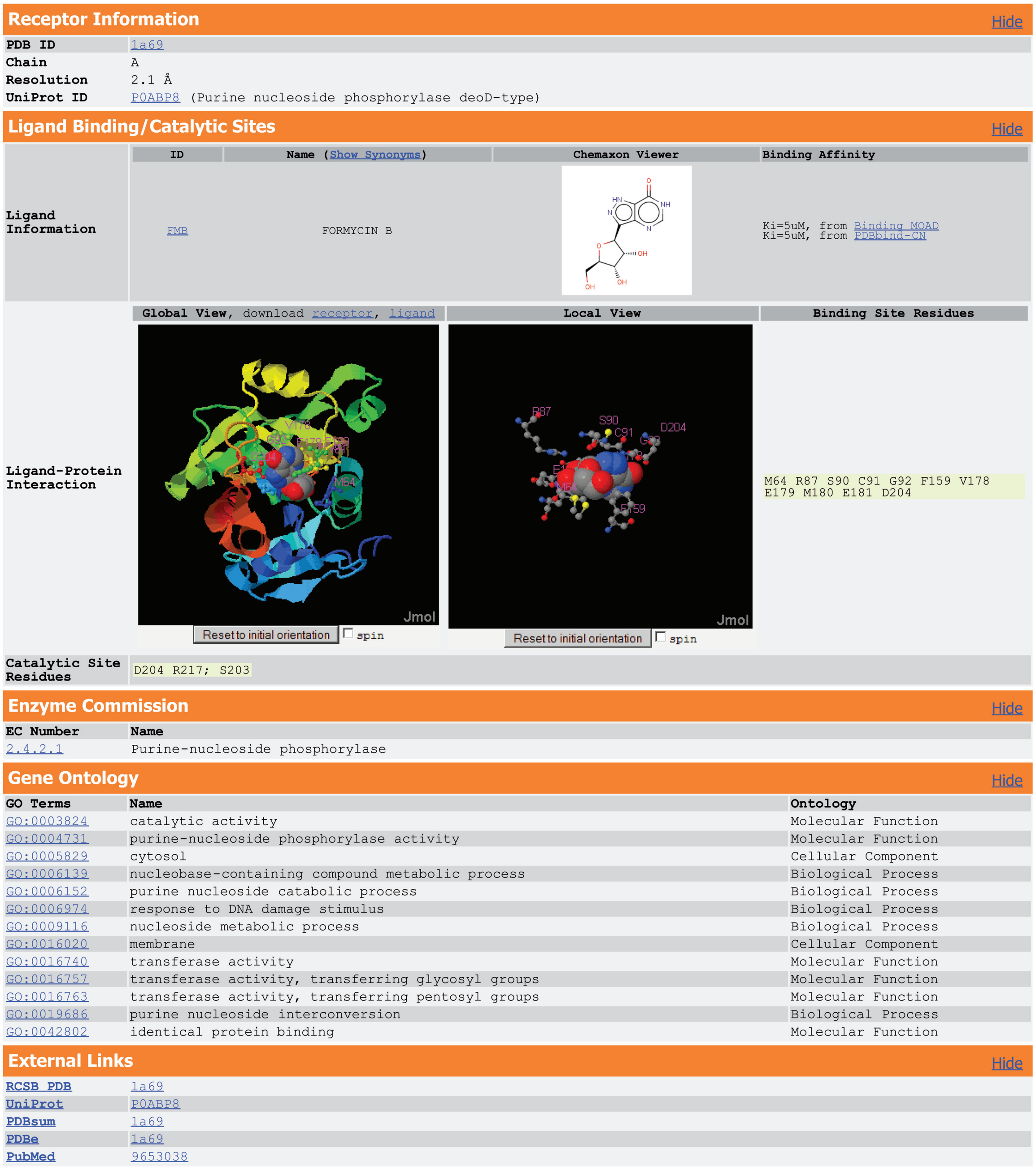


Figure 2. Distribution of ligands in BioLiP. ‘Regular’ represents the common small-molecule ligands except for the DNA/RNA, peptide, *k*-mer and metal ligands.



Receptor Information Hide

PDB ID	1a69
Chain	A
Resolution	2.1 Å
UniProt ID	P0ABP8 (Purine nucleoside phosphorylase deoD-type)

Ligand Binding/Catalytic Sites Hide

	ID	Name (Show Synonyms)	Chemaxon Viewer	Binding Affinity

Ligand Information	FMB	FORMYCIN B		Ki=5uM, from Binding MOAD
				Ki=5uM, from PDBbind-CN

	Global View, download receptor, ligand	Local View	Binding Site Residues
Ligand-Protein Interaction	Reset to initial orientation □ spin	Reset to initial orientation □ spin	M64 R87 S90 C91 G92 F159 V178 E179 M180 E181 D204

Catalytic Site Residues	D204 R217; S203

Enzyme Commission Hide

EC Number	Name
2.4.2.1	Purine-nucleoside phosphorylase

Gene Ontology Hide

GO Terms	Name	Ontology
GO:0003824	catalytic activity	Molecular Function
GO:0004731	purine-nucleoside phosphorylase activity	Molecular Function
GO:0005829	cytosol	Cellular Component
GO:0006139	nucleobase-containing compound metabolic process	Biological Process
GO:0006152	purine nucleoside catabolic process	Biological Process
GO:0006974	response to DNA damage stimulus	Biological Process
GO:0009116	nucleoside metabolic process	Biological Process
GO:0016020	membrane	Cellular Component
GO:0016740	transferase activity	Molecular Function
GO:0016757	transferase activity, transferring glycosyl groups	Molecular Function
GO:0016763	transferase activity, transferring pentosyl groups	Molecular Function
GO:0019686	purine nucleoside interconversion	Biological Process
GO:0042802	identical protein binding	Molecular Function

External Links Hide

RCSB PDB	1a69
UniProt	P0ABP8
PDBsum	1a69
PDBe	1a69
PubMed	9653038

Figure 3. An example of the function annotation in BioLiP. The annotation is for the chain A of the protein ‘purine nucleoside phosphorylase’ (PDB ID: 1A69). As can be seen from the Ligand-Binding/Catalytic Sites section, the ligand ‘Formycin B’ (ligand ID: FMB) binds to this protein with binding affinity $K_i = 5\,\mu$M. The ligand–protein interaction is visualized by the Jmol Applet globally and locally and the 3D structures of the ligand and the protein can be downloaded. The ligand-binding/catalytic site residues are listed in this section as well. The EC number and GO terms together with their names are presented in the Enzyme Commission and Gene Ontology sections. Cross-references to other databases (PDB, UniProt, PDBsum, PDBe and PubMed) are appended in the External Links section.

receptor and ligand are also provided for download in this section. Third, the ligand-binding and catalytic site residues of the receptor are listed.

In the SEARCH interface, eight fields are ready for searching through BioLiP database quickly: PDB ID, BioLiP ID, UniProt ID, EC number, GO term, ligand ID, ligand name and binding affinity. Such search can be completed quickly (<1 s) because all data in BioLiP are organized and processed with MySQL. The search results are presented in similar way as the ‘Browse all entries’ page.

DOWNLOAD BioLiP

The BioLiP database is freely available for download. Two different versions are provided: one is the redundant version that contains all the ligand–protein interaction sites in BioLiP and another is a non-redundant version at a 95% pairwise sequence identity. For single-chain receptor pairs, the sequence identity is defined as the number of identical residues divided by the length of the short chain. For multiple chain receptors, the sequence identity is specified by the pair of chains of the lowest sequence identity, i.e. two complex structures are considered to be redundant only when all the chain pairs have >95% receptor sequence identity. They can be downloaded at http://zhanglab.ccmb.med.umich.edu/BioLiP/download.html. For each version, three sets are available: the 3D structures of the PDB chains that interact with at least one biologically relevant ligand, the 3D structures of corresponding interacting ligands and the detailed information of each ligand-protein interaction site (binding site residues, ligand-binding affinity, catalytic site residues, EC number, GO terms and UniProt ID).

Ligand-binding sites prediction with COACH

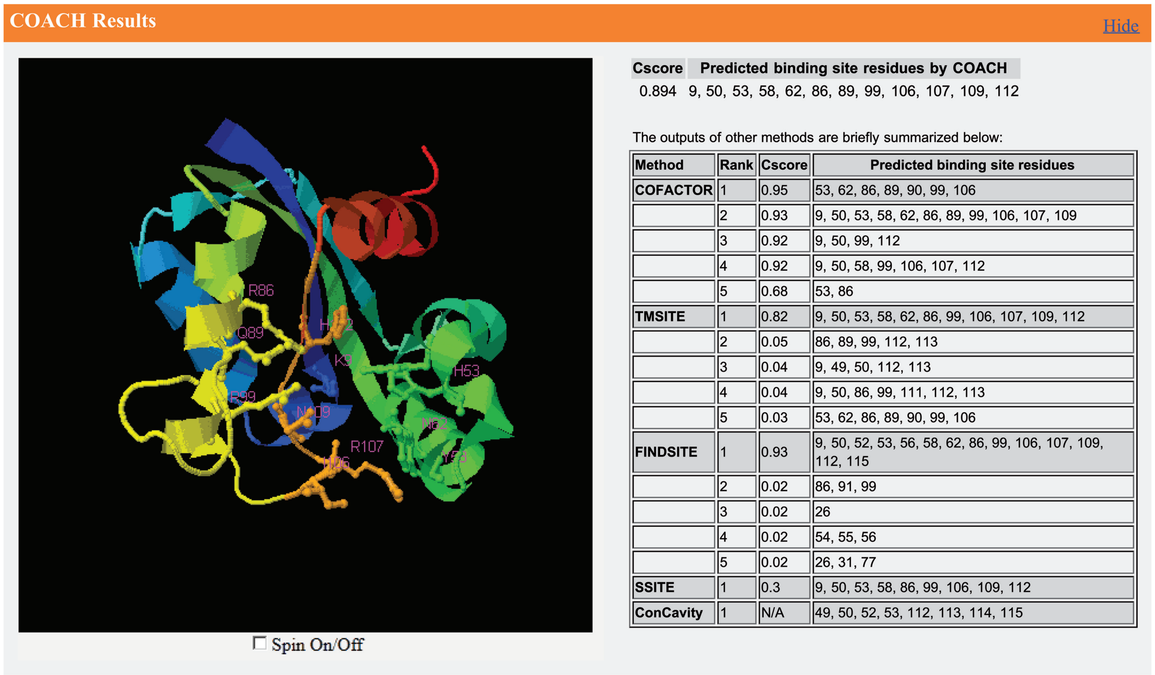
To annotate the function of uncharacterized proteins using the BioLiP database, we have developed a new algorithm COACH to predict ligand-binding sites. COACH is a consensus-based approach for ligand-binding sites prediction that combines the results of five state-of-the-art methods: COFACTOR (17,19), FINDSITE (28), ConCavity (29), TMSITE and SSITE. The first three are published methods and systematically benchmarked in a recent study (19). TMSITE and SSITE are two recently developed methods to predict ligand-binding sites by the complementary structural alignment and sequence profile–profile alignment search, respectively. COACH was found to significantly outperform any of the individual programs in our benchmark test (J. Yang, A. Roy and Y. Zhang, manuscript in preparation). In the output of COACH, the consensus prediction on the ligand-binding sites as well as the top five predictions from COFACTOR (19), FINDSITE (28), ConCavity (29), TMSITE and SSITE are presented.

To use the COACH server, users can provide either amino acid sequence or 3D structure of the target proteins. For the former case, SSITE is used to identify the ligand-binding information based on the sequence profile-to-profile search of the target against the BioLiP library, where the hits of the highest *E*-value is returned. In the latter case, all five methods are used to generate the binding prediction and the consensus hits from multiple searches are selected. A COACH prediction typically takes 0–4 h depending on the size of proteins. After the prediction is completed, an email alert will be sent to the users with the result data kept on our website for 6 months. An example of the COACH prediction is shown in Figure 4, which is also available at http://zhanglab.ccmb.med.umich.edu/BioLiP/BSP000001/.

SUMMARY

We have developed a comprehensive biologically relevant ligand–protein interaction database, BioLiP, for template-based ligand docking, virtual ligand screening and protein function annotation. Although there are already a handful of ligand-binding databases (2,3) in the literature, BioLiP is unique in the following aspects:

(1) A composite automated and manual procedure has been developed to assess the biological relevance of ligands. In order to alleviate the time-consuming task of manual verifications, a four-step hierarchical



Cscore	Predicted binding site residues by COACH
0.894	9, 50, 53, 58, 62, 86, 89, 99, 106, 107, 109, 112

The outputs of other methods are briefly summarized below:

Method	Rank	Cscore	Predicted binding site residues
COFACTOR	1	0.95	53, 62, 86, 89, 90, 99, 106
	2	0.93	9, 50, 53, 58, 62, 86, 89, 99, 106, 107, 109
	3	0.92	9, 50, 99, 112
	4	0.92	9, 50, 58, 99, 106, 107, 112
	5	0.68	53, 86
TMSITE	1	0.82	9, 50, 53, 58, 62, 86, 99, 106, 107, 109, 112
	2	0.05	86, 89, 99, 112, 113
	3	0.04	9, 49, 50, 112, 113
	4	0.04	9, 50, 86, 99, 111, 112, 113
	5	0.03	53, 62, 86, 89, 90, 99, 106
FINDSITE	1	0.93	9, 50, 52, 53, 56, 58, 62, 86, 99, 106, 107, 109, 112, 115
	2	0.02	86, 91, 99
	3	0.02	26
	4	0.02	54, 55, 56
	5	0.02	26, 31, 77
SSITE	1	0.3	9, 50, 53, 58, 86, 99, 106, 109, 112
ConCavity	1	N/A	49, 50, 52, 53, 112, 113, 114, 115

Figure 4. The COACH ligand-binding sites prediction results. The confidence score (Cscore) and the predicted binding site residues by COACH are presented in the first table, which is also visualized by Jmol Applet on the left panel. The top five predictions of other five individual methods are briefly summarized in the second table.

procedure is used to automatically verify the biological relevance of a ligand. After the automated procedure is completed, a careful manual check is performed to correct possible errors. To this end, we manually check and verify possible false-positive entries by reading the original literature and consulting other databases. This manual work ensures the completeness and high quality of BioLiP.

(2) The function annotation in BioLiP is comprehensive. For each ligand–protein interaction, in addition to the ligand-binding interactions, multiple other annotations are also presented, including ligand-binding affinity, the catalytic site residues, EC numbers, GO terms and cross-links to other databases. The completeness of the binding affinity data in BioLiP is unprecedented, which includes not only all high-quality annotations from the Binding MOAD (5), PDBbind (4) and BindingDB (6) databases but also data obtained by manual survey of the original literature. These data will facilitate functional annotations for most of uncharacterized proteins, when close ligand-binding templates are available.

(3) A new reliable algorithm COACH is developed to predict ligand-binding sites using the BioLiP database. COACH combines the results of five state-of-the-art ligand-binding sites prediction methods. COACH was found to significantly outperform any of the individual programs in our benchmark test.

(4) All data in BioLiP database are freely available for download. Two versions of database are provided, one is for the whole data set and the other is a non-redundant version at 95% sequence identity cutoff. These data sets can be very useful for template-based protein-ligand docking (15,18), virtual ligand screening (16) and protein function annotations (17,19).

These features are expected to have impacts on the research for studying protein–ligand interaction, protein function and structure-based drug design. Especially, we believe BioLiP would be a valuable resource for studying protein structure–function relationship. Recently, BioLiP has been successfully used for protein-ligand docking (18) and protein function prediction by COFACTOR (17,19), which was evaluated as the best ligand-binding prediction method in the Function Prediction Section in the CASP9 experiment (7). BioLiP was also used by COACH and COFACTOR in the CASP10 experiment for the function predictions, where exciting progress in this line is anticipated.

FUNDING

Funding for open access charge: NSF Career Award [DBI 0746198, in part]; National Institute of General Medical Sciences [GM083107, GM084222, in part].

Conflict of interest statement. None declared.

REFERENCES

1. Rose,P.W., Beran,B., Bi,C., Bluhm,W.F., Dimitropoulos,D., Goodsell,D.S., Prlić,A., Quesada,M., Quinn,G.B., Westbrook,J.D. *et al.* (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.*, **39**, D392–D401.
2. Lopez,G., Valencia,A. and Tress,M. (2007) FireDB–a database of functionally important residues from proteins of known structure. *Nucleic Acids Res.*, **35**, D219–D223.
3. Dessailly,B.H., Lensink,M.F., Orengo,C.A. and Wodak,S.J. (2008) LigASite—a database of biologically relevant binding sites in proteins with known apo-structures. *Nucleic Acids Res.*, **36**, D667–D673.
4. Wang,R., Fang,X., Lu,Y. and Wang,S. (2004) The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.*, **47**, 2977–2980.
5. Benson,M.L., Smith,R.D., Khazanov,N.A., Dimcheff,B., Beaver,J., Dresslar,P., Nerothin,J. and Carlson,H.A. (2008) Binding MOAD, a high-quality protein–ligand database. *Nucleic Acids Res.*, **36**, D674–D678.
6. Liu,T., Lin,Y., Wen,X., Jorissen,R.N. and Gilson,M.K. (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.*, **35**, D198–D201.
7. Schmidt,T., Haas,J., Cassarino,T.G. and Schwede,T. (2011) Assessment of ligand-binding residue predictions in CASP9. *Proteins*, **79**, 126–136.
8. Pabo,C.O. and Sauer,R.T. (1984) Protein-DNA recognition. *Annu. Rev. Biochem.*, **53**, 293–321.
9. Yamashita,M.M., Wesson,L., Eisenman,G. and Eisenberg,D. (1990) Where metal ions bind in proteins. *Proc. Natl Acad. Sci. USA*, **87**, 5648–5652.
10. Hendlich,M., Bergner,A., Günther,J. and Klebe,G. (2003) Relibase: design and development of a database for comprehensive analysis of protein–ligand interactions. *J. Mol. Biol.*, **326**, 607–620.
11. An,J., Totrov,M. and Abagyan,R. (2005) Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol. Cell Proteomics*, **4**, 752–761.
12. Kalinina,O.V., Wichmann,O., Apic,G. and Russell,R.B. (2012) ProtChemSI: a network of protein–chemical structural interactions. *Nucleic Acids Res.*, **40**, D549–D553.
13. Vanhee,P., Reumers,J., Stricher,F., Baeten,L., Serrano,L., Schymkowitz,J. and Rousseau,F. (2010) PepX: a structural database of non-redundant protein–peptide complexes. *Nucleic Acids Res.*, **38**, D545–D551.
14. Shulman-Peleg,A., Nussinov,R. and Wolfson,H.J. (2009) RsiteDB: a database of protein binding pockets that interact with RNA nucleotide bases. *Nucleic Acids Res.*, **37**, 369–D373.
15. Brylinski,M. and Skolnick,J. (2009) FINDSITE^{LHM}: a threading-based approach to ligand homology modeling. *PLoS Comput. Biol.*, **5**, e1000405.
16. Zhou,H. and Skolnick,J. (2012) FINDSITEX: a structure based, small molecule virtual screening approach with application to all identified human GPCRs. *Mol. Pharm.*, **9**, 1775–1784.
17. Roy,A. and Zhang,Y. (2012) Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement. *Structure*, **20**, 987–997.
18. Lee,H.S. and Zhang,Y. (2012) BSP-SLIM: a blind low-resolution ligand-protein docking approach using predicted protein structures. *Proteins*, **80**, 93–110.
19. Roy,A., Yang,J. and Zhang,Y. (2012) COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Res.*, **40**, W471–W477.
20. Roche,D., Tetchner,S. and McGuffin,L. (2011) FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins. *BMC Bioinformatics*, **12**, 160.
21. Porter,C.T., Bartlett,G.J. and Thornton,J.M. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, **32**, D129–D133.

22. The Gene Ontology Consortium. (2012) The Gene Ontology: enhancements for 2011. *Nucleic Acids Res.*, **40**, D559–D564.
23. Dimmer,E.C., Huntley,R.P., Alam-Faruque,Y., Sawford,T., O'Donovan,C., Martin,M.J., Bely,B., Browne,P., Chan,W.M., Eberhardt,R. *et al.* (2012) The UniProt-GO Annotation database in 2011. *Nucleic Acids Res.*, **40**, D565–D570.
24. The UniProt Consortium. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.
25. Velankar,S., McNeil,P., Mittard-Runte,V., Suarez,A., Barrell,D., Apweiler,R. and Henrick,K. (2005) E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res.*, **33**, D262–D265.
26. Kanehisa,M., Goto,S., Sato,Y., Furumichi,M. and Tanabe,M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.
27. Yamanishi,M., Kinoshita,K., Fukuoka,M., Saito,T., Tanokuchi,A., Ikeda,Y., Obayashi,H., Mori,K., Shibata,N., Tobimatsu,T. *et al.* (2012) Redesign of coenzyme B12 dependent diol dehydratase to be resistant to the mechanism-based inactivation by glycerol and act on longer chain 1, 2-diols. *FEBS J.*, **279**, 793–804.
28. Brylinski,M. and Skolnick,J. (2008) A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc. Natl Acad. Sci. USA*, **105**, 129–134.
29. Capra,J.A., Laskowski,R.A., Thornton,J.M., Singh,M. and Funkhouser,T.A. (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput. Biol.*, **5**, e1000585.