BIOINFORMATICS ORIGINAL PAPER
Vol. 31 no. 3 2015, pages 405–412
doi:10.1093/bioinformatics/btu626

Databases and ontologies
Advance Access publication October 9, 2014

PDB-wide collection of binding data: current status of the PDBbind database

Zhihai Liu[1], Yan Li[1], Li Han[1], Jie Li[1], Jie Liu[1], Zhixiong Zhao[1], Wei Nie[1], Yuchen Liu[1] and Renxiao Wang[1,2,*]

[1]State Key Laboratory of Bioorganic and Natural Products Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, 345 Lingling Road, Shanghai 200032 and [2]State Key Laboratory of Quality Research in Chinese Medicine, Macau Institute for Applied Research in Medicine and Health, Macau University of Science and Technology, Macau, People's Republic of China

Associate Editor: Janet Kelso

ABSTRACT

Motivation: Molecular recognition between biological macromolecules and organic small molecules plays an important role in various life processes. Both structural information and binding data of biomolecular complexes are indispensable for depicting the underlying mechanism in such an event. The PDBbind database was created to collect experimentally measured binding data for the biomolecular complexes throughout the Protein Data Bank (PDB). It thus provides the linkage between structural information and energetic properties of biomolecular complexes, which is especially desirable for computational studies or statistical analyses.

Results: Since its first public release in 2004, the PDBbind database has been updated on an annual basis. The latest release (version 2013) provides experimental binding affinity data for 10 776 biomolecular complexes in PDB, including 8302 protein–ligand complexes and 2474 other types of complexes. In this article, we will describe the current methods used for compiling PDBbind and the updated status of this database. We will also review some typical applications of PDBbind published in the scientific literature.

Availability and implementation: All contents of this database are freely accessible at the PDBbind-CN Web server at http://www.pdbbind-cn.org/.

Contact: wangrx@mail.sioc.ac.cn.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 23, 2014; revised on August 1, 2014; accepted on September 12, 2014

1 INTRODUCTION

Biological macromolecules (proteins and nucleic acids) often execute their functions via molecular recognition with other biological macromolecules or organic small molecules. Characterization of the interactions between molecular complexes is thus helpful for interpreting the mechanism of various life processes. Modern X-ray crystallography and multidimensional nuclear magnetic resonance (NMR) techniques have produced a wealth of structural information of molecular complexes. The worldwide Protein Data Bank (PDB) holds the largest open resource for experimentally determined biomolecular structures (Berman *et al.*, 2003). It has been growing rapidly during the past 20 years or so. By June 2014, over 100 000 structures had already been deposited into PDB. A number of derivative databases based on PDB have been developed with their own scopes and emphases, such as PDBsum (Laskowski, 2001), Relibase + (Hendlich *et al.*, 2003), Het-PDB Navi (Yamaguchi *et al.*, 2004) and scPDB (Kellenberger *et al.*, 2006). These databases provide annotated protein–ligand complex structures or their special features, such as binding pockets and ligand molecules. Collectively, the structural information provided by these databases allows visual examination of interactions between a wide range of molecular complexes.

On the other hand, energetic properties are also indispensable for obtaining an in-depth understanding of such events. Here, the major problem is that energetic data, such as binding data, are often scattered in the huge pool of literature so that they are difficult to access. To tackle this problem, some existing databases started to incorporate binding data; for example, ChEMBL (Gaulton *et al.*, 2012) and PubChem (Bolton *et al.*, 2008; Wang *et al.*, 2014) are perhaps the two most comprehensive databases in this category. They currently provide a large amount of experimental binding data of diverse bioactive compounds as well as relevant information on molecular targets, experimental conditions, etc. Other databases were created with a focus on binding data, e.g. BindingDB (Liu *et al.*, 2007) and PDSP Ki (Roth *et al.*, 2000). Obviously, the rich knowledge of the interactions between bioactive small molecules and their potential targets provided by these databases is widely welcomed by chemical biologists and medicinal chemists.

Nevertheless, the link between structural information and energetic properties is desirable especially for theoretical and computational studies on molecular recognition. This link is missing, or at least not the primary focus of the databases mentioned above. With more and more successful applications of computational methods to modern drug discovery (Manetti *et al.*, 2008; Schames *et al.*, 2004; Wlodawer, 2002), the requirement for a combined knowledge of structures and binding data has become increasingly urgent. To fill the gap between structural information and energetic properties of biomolecular complexes, a practical strategy is to collect experimentally measured binding data for the biomolecular complexes deposited

*To whom correspondence should be addressed.

in the PDB. Currently, PDBbind (Wang, *et al.*, 2004, 2005) and binding MOAD (Benson *et al.* 2008; Hu *et al.* 2005) are two outstanding databases created for this purpose. Both of them are based on a systematic sampling of the entire PDB. A few early databases also belong to this category, including LPDB (Roche *et al.* 2001), PLD (Puvanendrampillai *et al*, 2003) and AffiDB (Block *et al.*, 2006), but they are much smaller in size and have undergone virtually no updates since original release. A summary of current databases collecting structural information and/or binding data of protein–ligand complexes in all three categories is given in the Supplementary Table S1.

The PDBbind database was originally developed by Prof. Shaomeng Wang's group at the University of Michigan and was first released to the public on May, 2004. Since 2007, this database has been maintained and further developed by our group at the Shanghai Institute of Organic Chemistry in China under a mutual agreement with the University of Michigan. Since then, the database has been released through the PDBbind-CN Web server at http://www.pdbbind-cn.org/ and http://www.pdbbind.org.cn/. Over the years, we have regularly updated the PDBbind database to keep pace with the growth of PDB itself. The methods for compiling the database have also been improved along the way. The latest release of PDBbind, i.e. version 2013, provides binding data for >10 700 molecular complexes in the PDB, making it the largest collection of its kind to date. In all, >2400 registered users from some 40 countries access this database regularly. There are also a large number of anonymous users around the world. In this article, we will describe the current methods used for compiling the PDBbind and its updated status. We will also review some typical applications of PDBbind published in the scientific literature to illustrate the significant value of this database.

2 METHODS FOR COMPILING THE PDBBIND DATABASE

2.1 Classification of biomolecular complexes

The first step of our entire workflow is to classify the valid biomolecular complexes deposited in the PDB. This task is fairly complicated, as PDB itself does not provide a clear and systematic classification on whether a PDB structure is a 'complex'. Our early experience was that keyword-based analysis of the text annotations in PDB files missed too many true complexes. Thus, we have developed a classification scheme completely based on the structural information given in the PDB file. This scheme is automated by a set of in-house programs, and has been continuously refined over the past years. The basic flowchart of our current scheme is shown in Figure 1. The input is a standard PDB-format structure file. Our program will then classify the given structure as a non-complex (pure protein or nucleic acid molecule), a complex but not in our interests or a valid complex. Valid complexes include four major categories, i.e. (i) complexes formed between protein and small-molecule ligand, (ii) complexes formed between nucleic acid and small-molecule ligand, (iii) complexes formed between two protein molecules and (iv) complexes formed between protein and nucleic acid.

The flowchart shown in Figure 1 is basically self-explanatory. Yet, a few issues need to be explained in more detail. First, our scheme relies on a dictionary to distinguish 'valid' organic ligand molecules from biological cofactors/coenzymes, inorganic ions, buffer components, organic solvent molecules and other miscellaneous 'heterogens' included in PDB files. This dictionary is compiled based on the standard 'Chemical Component Dictionary' provided by the PDB. It has been updated annually along with the PDBbind database itself. Second, the definition of a 'valid' protein–protein complex is also a bit complicated. In our scheme, a valid protein–protein complex must be formed by at least two peptide chains from two different protein molecules. A homodimer or multimer of the same protein molecule is not counted. Each peptide chain in the complex should have at least 20 amino acid residues. This cutoff is chosen because Trp-cage, which has 20 residues, is arguably the smallest functional protein. Our program also examines the number of interacting residues on the binding interface of two peptide chains. Here, we follow Tsai's method (Tsai *et al.*, 1996) to define interacting residues. A valid protein–protein complex must have at least 10 interacting residues at each side. Otherwise, the binding interface is probably the result of a crystal packing rather than a meaningful molecular recognition process.

As of January 1, 2013, there were 87 085 experimentally determined structures released by the PDB. Classification of these structures by our scheme is given in Table 1. Four categories of valid complexes, including 29 008 protein–ligand complexes, 5341 protein–protein complexes, 3852 protein–nucleic acid complexes and 717 nucleic acid–ligand complexes, were considered in the following steps.

2.2 Collection of binding data

The second step of our workflow is to collect binding affinity data from the scientific literature for biomolecular complexes classified as 'valid'. Our focus is the 'primary reference' provided in each relevant PDB structure file. This strategy offers a good chance to obtain the desired binding data by examining a single article and has been proven to be effective over the years. The real challenge at this step is that binding affinity data as well as relevant information (such as binding assay method and experimental settings) have to be retrieved from the given reference manually. Virtually no journals offer such data in a standard format. To reduce labor, we have developed a set of computer scripts to conduct a keyword-based search through the full text of a given article. Only the articles containing clues of binding data will be examined manually. Each such article is then examined independently by two persons. Binding data will not be formally recorded unless both persons obtain a consistent result. This quality-control practice effectively reduces human mistakes in our data to a low level.

It needs to be reemphasized that every single binding data in the PDBbind was retrieved from an original reference rather than copied from other data resources. Up to now, we have processed >24 000 scientific publications in total. We actually mark where the binding data were found in the original

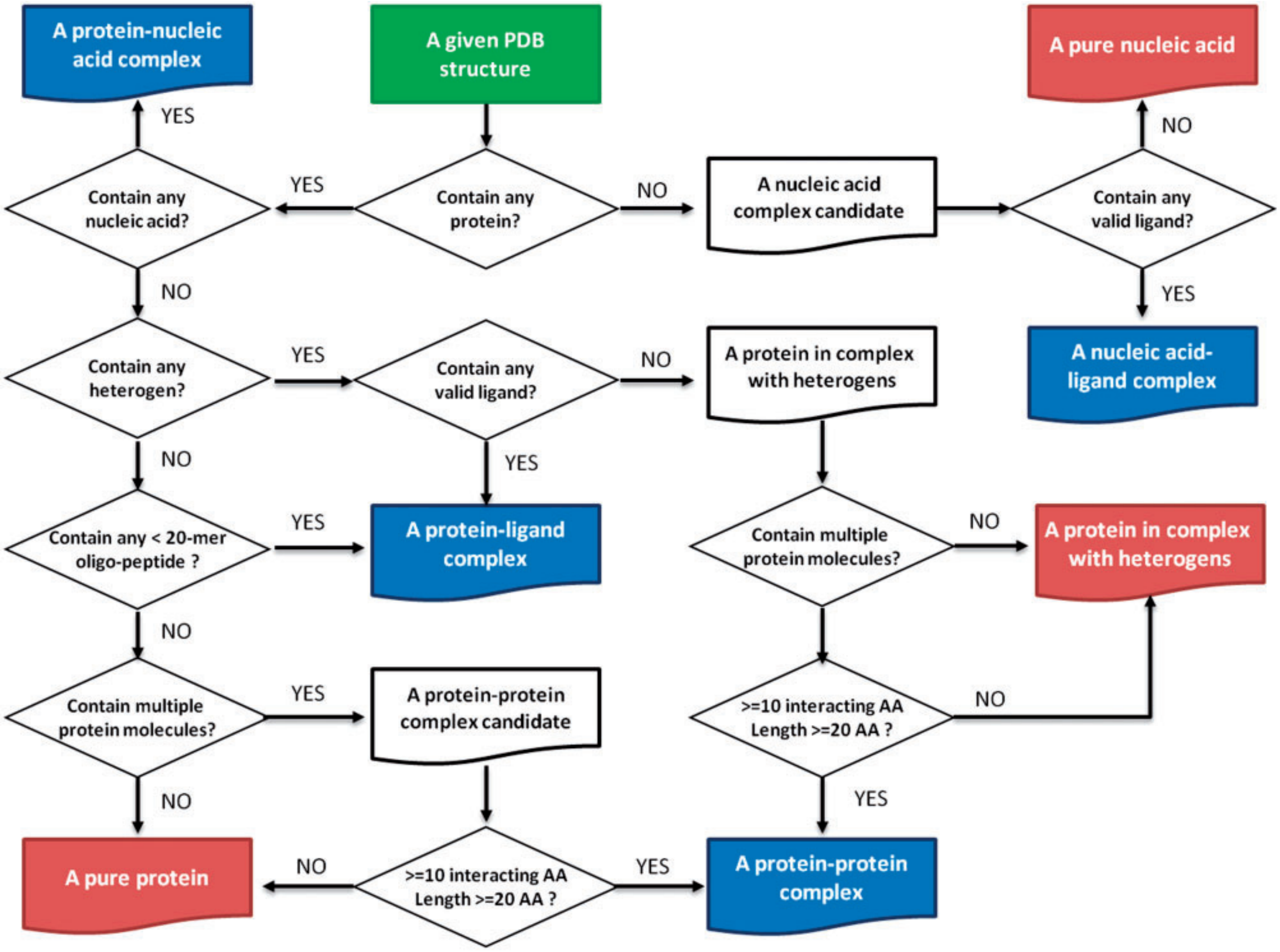

A protein-nucleic acid complex
A given PDB structure
A pure nucleic acid
YES
NO
Contain any nucleic acid?
YES
Contain any protein?
NO
A nucleic acid complex candidate
Contain any valid ligand?
NO
YES
Contain any heterogen?
YES
Contain any valid ligand?
NO
A protein in complex with heterogens
A nucleic acid-ligand complex
NO
YES
Contain any < 20-mer oligo-peptide ?
YES
A protein-ligand complex
Contain multiple protein molecules?
NO
A protein in complex with heterogens
NO
Contain multiple protein molecules?
YES
A protein-protein complex candidate
>=10 interacting AA Length >=20 AA ?
NO
NO
YES
A pure protein
NO
>=10 interacting AA Length >=20 AA ?
YES
A protein-protein complex

Fig. 1. Flowchart for classification of the biomolecular structures in the PDB

Table 1. Classification of the entire PDB as released on January 1, 2013

Category	Number	Percentage in the entire PDB
Proteins in complex with valid organic ligands.	29 008	33.31
Proteins in complex with proteins.	5341	6.13
Proteins in complex with nucleic acids.	3852	4.42
Nucleic acids in complex with organic ligands.	717	<1
Apo-proteins and proteins in complex with 'junk' molecules.	37 148	42.66
Proteins in complex with biological cofactors/coenzymes.	8586	9.86
Nucleic acids.	1746	2.00
Other oligomers.	687	<1
Total	87 085	

reference. Thus, if someone is suspicious of any data in our database, we can easily find its origin in our archives and make a correction when necessary.

We record three major forms of binding data: the dissociation constant (K_d), the inhibition constant (K_i) and the concentration at 50% inhibition (IC_{50}). If multiple forms of binding data of a given complex are available, we actually record all of them but present the end user with only the most preferred data with a priority order of $K_d > K_i > IC_{50}$. Similarly, if the binding data of a given complex is measured under different experimental settings, we present the end user with only the data measured at room temperature and neutral pH or a condition closest to this setting.

The overall 'yield ratio' at this step is 25–30%, i.e. binding data can be found for 25–30% of the valid complexes identified at the previous step. In addition to binding data, we also record the basic structural and biological/chemical information of each complex, such as release year, resolution and *R*-factor of the crystal structure, protein name, protein sequence, ligand name, ligand structure, enzyme commission (EC) number and so on. Although such information is also available from other databases, such as PDB, integration of such information helps the user to find the desired information more conveniently.

2.3 Further processing of datasets

A major motivation for creating the PDBbind database is to provide high-quality datasets for the development of molecular docking and scoring methods for structure-based drug design. The protein–ligand complexes with known binding data compiled through the previous steps, which we call the 'general set', are not recommended to serve directly as such a dataset. This is because many complexes still may have a problem either with their structures or binding data, and thus, are not 'healthy' enough for docking/scoring studies. The size of the general set is also too large for this purpose.

Therefore, as an additional feature of PDBbind, we provide users a set of protein–ligand complexes selected out of the general set since the first public release (version 2004). We call this dataset the 'refined set'. A fairly complicated set of rules is applied to the selection of the refined set with considerations of (i) the quality of the complex structures, (ii) the quality of the binding data and (iii) the biological/chemical nature of the complex. This set of rules has also evolved over the years to become more and more stringent. The current rules have been described in detail in our recent publication (Li *et al.*, 2014a). In PDBbind version 2013, 2959 protein–ligand complexes are included in the 'refined set'. In addition, to make the refined set readily readable by most molecular modeling software, the original PDB structural file of each complex is split into a protein molecule and a ligand molecule. Both partners in each complex are processed properly and saved in standard formats (such as PDB, Mol2 and SDF). All these processed structural files together with the binding data can be downloaded as a package from our PDBbind-CN Web site.

Furthermore, a subset of protein–ligand complexes, which we call the 'core set', is selected out of the refined set as a standard benchmark for evaluating docking/scoring methods. This is because there is considerable sample redundancy in the refined set. For example, nearly 10% of the protein–ligand complexes in the refined set are formed by HIV-1 protease. Our opinion is that a standard benchmark needs to control sample redundancy, otherwise the evaluation results could be biased. Since PDBbind version 2007, we have started to compile the core set through systematic sampling of the refined set. The rules for selection of the core set have also evolved during the past years. The current rules have been described in our recent publication (Li *et al.*, 2014a). The core set in PDBbind version 2013 contains 195 protein–ligand complexes in 65 families. The core set can also be downloaded as a package from our PDBbind-CN Web site.

Protein–protein complexes and protein–nucleic acid complexes have been added to PDBbind since version 2008 (Table 2). In version 2013, there are already 1804 protein–protein complexes and 587 protein–nucleic acid complexes with known experimental binding data. Computational studies on the molecular recognition between biological macromolecules need high-quality datasets as well. We are now working on defining the rules for selecting these two types of complexes. Refined sets for these two types of complexes will be provided in PDBbind in the near future.

2.4 Design of Web site interface

The PDBbind database is now accessible at the PDBbind-CN Web server (http://www.pdbbind-cn.org and http://www.pdbbind.org.cn/). Users can browse, analyze and search the contents of PDBbind on this Web interface. The basic interface is the 'BROWSE' page, where users can view all entries in PDBbind one by one. This page provides binding data and other basic information of each complex. Users can view the 3D structure of the complex, or only the binding pocket, the protein molecule or the ligand molecule with various display options powered by a Java plug-in 'MarvinView' released by ChemAxon (Fig. 2A). On this page, hyperlinks to some external databases, including PDB, PDBsum and PubChem, are also provided so that users can get other information about the complex conveniently. Under mutual agreement, users can also find the binding data provided by PDBbind on the Web pages of PDB, PDBsum and PubChem.

A text-based search function is provided on the 'DATA' page, which enables three search methods. A number of queries can be used, including complex subset type, PDB code, protein name, ligand name, EC number, release year, resolution, binding data and so on. Users can save outcomes as an Adobe PDF or Microsoft Excel table, and download the related structure files (Fig. 2B). Substructure or similarity search of the organic ligand molecules is provided on the 'STRUCTURE' page. A user can input a SMILES text string, or draw a structure with the MarvinSketch plug-in to conduct the structure search (Fig. 2C). This search is carried out over all organic ligand molecules in the PDB, not limited to the scope of the PDBbind. The latest release of PDBbind contains >11 000 organic ligand molecules from the PDB. A sequence-based search of protein or nucleic acid molecules is provided on the 'SEQUENCE' page. Users can carry out a BLAST similarity search among all protein and nucleic acid sequences throughout the PDB. A user can input a sequence in the FASTA format to conduct the search. The outcomes can be viewed online or can be downloaded (Fig. 2D).

3 UPDATES OF PDBbind AND FACTS OF THE CURRENT RELEASE

As the PDBbind database was released at our PDBbind-CN Web server in 2007, we have been able to update it on an annual basis to keep up with the growth of the PDB. To conduct an update, we download the entire PDB in the first week of each year, and then complete the classification, data collection, data processing and Web design steps in the next few months. The newly released version is therefore named after the release year. Table 2 gives a summary of the basic information of each version after 2007. One can see that the binding data in PDBbind has increased around 15% each year. Compared with version 2007, the binding data in the current release (version 2013) have almost tripled. Originally, PDBbind only included complexes formed between protein and organic ligand molecules because this type of complex is of direct interest for drug discovery. An ambitious expansion since version 2008 also includes complexes formed between biological macromolecules, i.e. protein–protein complexes and protein–nucleic acid complexes. Thereafter,

Table 2. A comparison of current and past versions of PDBbind[a]

Version	Entries in PDB	Biomolecular complexes	Complexes with binding data				
			Protein–ligand complexes	Nucleic acid–ligand complexes	Protein–protein complexes	Protein–nucleic acid complexes	Total
2007	40 876	11 822	3124	0	0	0	3124
2008	48 092	18 211	3539	40	471	250	4300
2009	55 118	23 284	4277	44	1053	304	5678
2010	62 387	26 434	5075	55	1281	361	6772
2011	70 224	30 259	6051	66	1441	428	7986
2012	78 235	34 180	7121	79	1597	511	9308
2013	87 085	38 918	8302	83	1804	587	10 776
2014	96 592	44 569	~10 650	~90	~1600	~660	~13 000[b]

[a] This table does not include early versions of PDBbind before 2007.

[b] This is an estimated number, as version 2014 was under compilation at the time of writing.

PDBbind has become a true PDB-wide collection of binding data.

The latest release of PDBbind is version 2013, which provides the experimental binding data and structures of 10 776 molecular complexes in the PDB. Among them are 8302 protein–ligand complexes, 1804 protein–protein complexes, 587 protein–nucleic acid complexes and 83 nucleic acid–ligand complexes (Table 2). The binding data in this version include 4682 K_d values, 3028 K_i values and 3066 IC_{50} values, spanning over 15 orders of magnitude from 0.29 fM to 400 mM. The distribution of these binding data for four categories of complexes is shown in Figure 3. One can see that all three types of binding data are significantly populated for protein–ligand complexes, whereas the binding data for the other three categories of complexes are primarily reported as K_d values in the literature.

4 TYPICAL APPLICATIONS OF THE PDBbind DATABASE

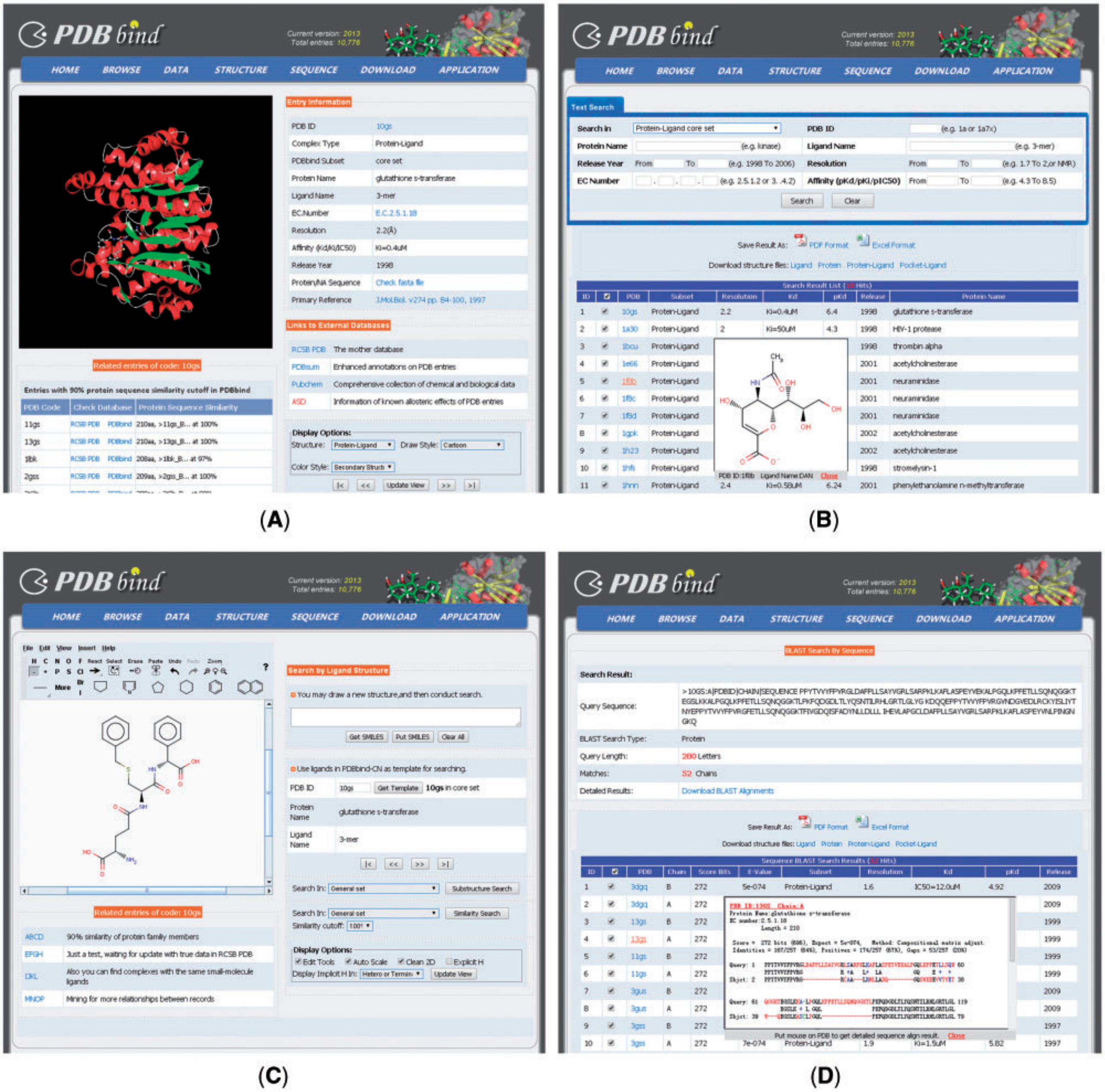
Since the public release of PDBbind in 2004, it has been used by a growing number of researchers worldwide. We have noticed that the registered PDBbind users have quite different backgrounds. Sometimes they use the PDBbind data in ways beyond our original expectation. According to our survey, >120 studies using PDBbind data have already been published in the literature. We will briefly review some typical applications below. Hopefully, our descriptions will inspire the readers to invent even more ingenious ways of applying the PDBbind database.

4.1 Development and validation of docking/scoring methods

A primary motivation for creating the PDBbind database is to provide high-quality datasets for developing and validating scoring functions and docking methods. Indeed, a fair number of research groups use our database in this way. For example, Olson *et al.* developed AutoDock Vina, a new program for molecular docking and virtual screening (Trott and Olson, 2010) with ~100-fold speedup and improved binding pose predictions as compared with the previous AutoDock 4. The scoring function in the AutoDock Vina was calibrated on the PDBbind refined set. Other recently published scoring functions, which used protein–ligand complex datasets from PDBbind include NNScore (Durrant and McCammon, 2010), DSX (Neudert and Klebe, 2011), PHOENIX (Tang *et al.* 2011), LISA (Zheng *et al.* 2011), B2BScore (Liu *et al.* 2013), SFCscore (Zilian and Sotriffer, 2013), ID-Score (Li *et al.*, 2013a), KECSA (Zheng *et al.* 2013) and many others.

Our group has also applied PDBbind extensively in this area. For example, we have developed the CASF benchmark, i.e. comparative assessment of scoring functions, over the past years. The first published study was CASF-2007 (Cheng *et al.*, 2009). An updated study, i.e. CASF-2013, was recently published (Li *et al.*, 2014a, b). This benchmark aims at providing an objective and systematic evaluation of scoring function performance, which can be shared among the community. The PDBbind core set served as the primary test set in this benchmark. Another example was a test of the MM-PB/SA method on a set of 24 protein–ligand complexes selected from the PDBbind database (Li *et al.* 2010). All of these complexes have multiple conformations determined by an NMR technique so that the standard conformational sampling method in MM-PB/SA can be evaluated.

Here, we want to mention in particular that the PDBbind refined set as well as the core set is compiled by a set of standards reflecting the commonsense in docking/scoring studies, which we expect to serve the majority in this community. The users of course have the freedom to make selections among the rough data provided by PDBbind based on their own standards. For example, if one is particularly interested in complexes in which a metal ion mediates the interactions between protein and ligand, he/she can compile a subset of such protein–ligand complexes out of the general set or the refined set in PDBbind. The users are welcome to do so, and we hope that they are willing to share their selections with others. However, for obvious reasons, we cannot respond to every request of this type from the users.



(A)

Entry Information

PDB ID	10gs
Complex Type	Protein-Ligand
PDBbind Subset	core set
Protein Name	glutathione s-transferase
Ligand Name	3-mer
EC.Number	E.C.2.5.1.18
Resolution	2.2(Å)
Affinity (Kd/Ki/IC50)	Ki=0.4uM
Release Year	1998
Protein/NA Sequence	Check fasta file
Primary Reference	J.Mol.Biol. v274 pp. 84-100, 1997

Links to External Databases

RCSB PDB	The mother database
PDBsum	Enhanced annotations on PDB entries
Pubchem	Comprehensive collection of chemical and biological data
ASD	Information of known allosteric effects of PDB entries

Related entries of code: 10gs

Entries with 90% protein sequence similarity cutoff in PDBbind

PDB Code	Check Database	Protein Sequence Similarity
11gs	RCSB PDB PDBbind	210aa, >11gs_B... at 100%
13gs	RCSB PDB PDBbind	210aa, >13gs_B... at 100%
1bk	RCSB PDB PDBbind	208aa, >1bk_B... at 97%
2gss	RCSB PDB PDBbind	209aa, >2gss_B... at 100%

Display Options:
Structure: Protein-Ligand Draw Style: Cartoon
Color Style: Secondary Struct
|< << Update View >> >|

(B)

Text Search

Search in	Protein-Ligand core set	PDB ID	(e.g. 1a or 1a7x)
Protein Name	(e.g. kinase)	Ligand Name	(e.g. 3-mer)
Release Year	From To (e.g. 1998 To 2006)	Resolution	From To (e.g. 1.7 To 2,or NMR)
EC Number	. . . (e.g. 2.5.1.2 or 3. .4.2)	Affinity (pKd/pKi/pIC50)	From To (e.g. 4.3 To 8.5)

Search Clear

Save Result As: PDF Format Excel Format
Download structure files: Ligand Protein Protein-Ligand Pocket-Ligand

Search Result List (10 Hits)

ID	☑	PDB	Subset	Resolution	Kd	pKd	Release	Protein Name
1	☑	10gs	Protein-Ligand	2.2	Ki=0.4uM	6.4	1998	glutathione s-transferase
2	☑	1a30	Protein-Ligand	2	Ki=50uM	4.3	1998	HIV-1 protease
3	☑	1bcu	Protein-Ligand				1998	thrombin alpha
4	☑	1e66	Protein-Ligand				2001	acetylcholinesterase
5	☑	1f8b	Protein-Ligand				2001	neuraminidase
6	☑	1f8c	Protein-Ligand				2001	neuraminidase
7	☑	1f8d	Protein-Ligand				2001	neuraminidase
8	☑	1gpk	Protein-Ligand				2002	acetylcholinesterase
9	☑	1h23	Protein-Ligand				2002	acetylcholinesterase
10	☑	1hfs	Protein-Ligand				1998	stromelysin-1
11	☑	1hnn	Protein-Ligand	2.4	Ki=0.58uM	6.24	2001	phenylethanolamine n-methyltransferase

PDB ID:1f8b Ligand Name:DAN Close

(C)

File Edit View Insert Help
H C N O F React Select Erase Paste Undo Redo Zoom ?
P S Cl Br I More

Search by Ligand Structure

You may draw a new structure,and then conduct search.

Get SMILES Put SMILES Clear All

Use ligands in PDBbind-CN as template for searching.

PDB ID	10gs Get Template **10gs** in core set
Protein Name	glutathione s-transferase
Ligand Name	3-mer

|< << >> >|

Search In: General set Substructure Search
Search In: General set Similarity Search
Similarity cutoff: 100%

Display Options:
☑ Edit Tools ☑ Auto Scale ☑ Clean 2D □ Explicit H
Display Implicit H In: Hetero or Termin Update View

Related entries of code: 10gs

ABCD	90% similarity of protein family members
EFGH	Just a test, waiting for update with true data in RCSB PDB
IJKL	Also you can find complexes with the same small-molecule ligands
MNOP	Mining for more relationships between records

(D)

BLAST Search By Sequence

Search Result:

Query Sequence:	>10GS:A|PDBID|CHAIN|SEQUENCE PPYTVVYFPVRGLDAFPLLSAYVGRLSARPKLKAFLASPEYVEKALPGQLKPFETLLSQNQGGKT EGSLKKALPGQLKPFETLLSQNQGGKTLPKFQDGDLTLYQSNTILRHLGRTLGLYG KDQQEPPYTVVYFPVRGVNDGVEDLRCKYISLIYT NYEPPYTVVYFPVRGFETLLSQNQGGKTFIVGDQISFADYNLLDLLL IHEVLAPGCLDAFPLLSAYVGRLSARPKLKAFLASPEYVNLPINGN GKQ
BLAST Search Type:	Protein
Query Length:	280 Letters
Matches:	52 Chains
Detailed Results:	Download BLAST Alignments

Save Result As: PDF Format Excel Format
Download structure files: Ligand Protein Protein-Ligand Pocket-Ligand

Sequence BLAST Search Results (52 Hits)

ID	☑	PDB	Chain	Score Bits	E-Value	Subset	Resolution	Kd	pKd	Release
1	☑	3dgq	B	272	5e-074	Protein-Ligand	1.6	IC50=12.0uM	4.92	2009
2	☑	3dgq	A	272						2009
3	☑	13gs	B	272						1999
4	☑	13gs	A	272						1999
5	☑	11gs	B	272						1999
6	☑	11gs	A	272						1999
7	☑	3gus	B	272						2009
8	☑	3gus	A	272						2009
9	☑	3gss	B	272						1997
10	☑	3gss	A	272	7e-074	Protein-Ligand	1.9	Ki=1.5uM	5.82	1997

```
PDB ID:13GS  Chain:A
Protein Name:glutathione s-transferase
EC number:2.5.1.18
        Length = 210

 Score =  272 bits (696), Expect = 5e-074,   Method: Compositional matrix adjust.
 Identities = 167/257 (64%), Positives = 174/257 (67%), Gaps = 53/257 (20%)

Query: 1   PPYTVVYFPVRG[illegible] 60
            PPYTVVYFPVRG [illegible]
Sbjct: 2   PPYTVVYFPVRG[illegible] 38

Query: 61  [illegible] 119
Sbjct: 39  [illegible] 79
```

Put mouse on PDB to get detailed sequence align result. Close

Fig. 2. Web interfaces of the PDBbind database: (**A**) Basic information of each complex; (**B**) text-based search; (**C**) ligand structure search; (**D**) sequence-based search

4.2 Analysis of drug–target interactions and structure-based drug design

The protein–ligand complexes included in PDBbind cover a wide range of validated and potential drug targets and many bioactive organic molecules. Such information can be used to study the complex interactions between drugs and their targets. One example is the application of network pharmacology to predict drug effect and toxicity resulting from multi-target interactions. Hsin *et al.* (2013) presented an interesting computational approach that combines two machine learning systems and multiple molecular docking methods to assess binding potentials of a given compound against multiple target proteins in a complex molecular network. Both systems were built and validated by using the PDBbind refined set (version 2007).

Ligands with known binding data in PDBbind were also used to build a statistical model for discriminating drugs and non-drugs by Garcia-Sosa *et al.* (2012). In their study, a probability-based approach was introduced for quantitative classification of compounds as drugs and non-drugs. Disease and organ-specificity was also taken into account. In addition to demonstrating good predictive power, their classification model is chemically

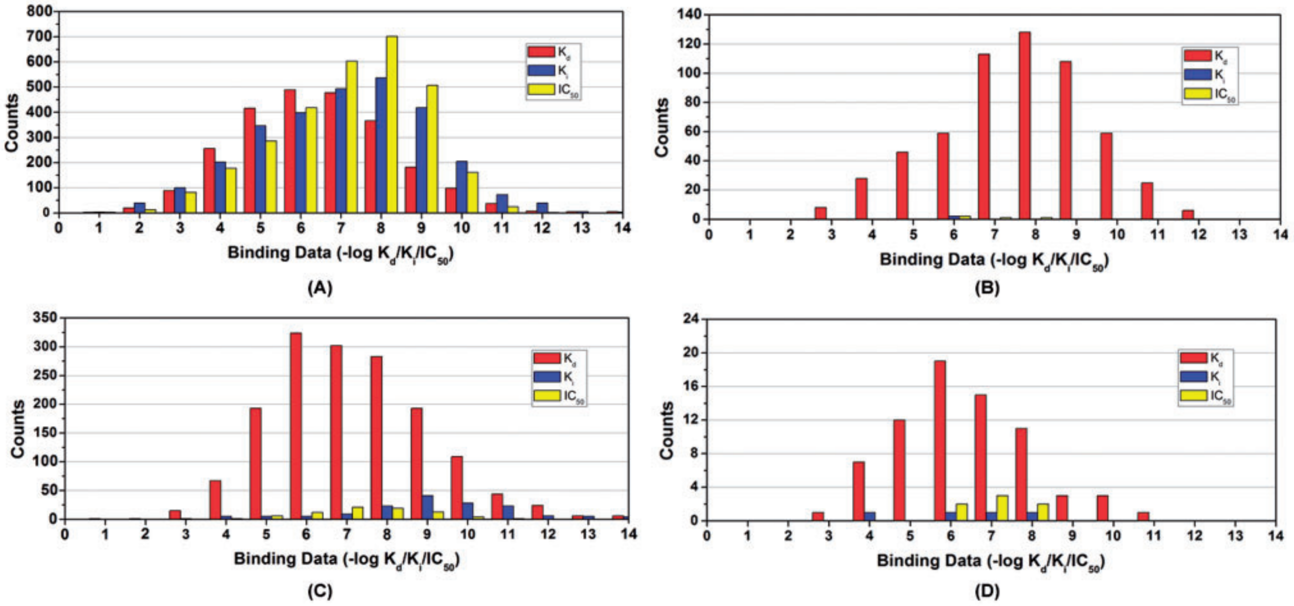


Fig. 3. Distribution of binding data in the PDBbind version 2013: (**A**) Protein–ligand complexes; (**B**) Protein–nucleic acid complexes; (**C**) Protein–protein complexes; (**D**) Nucleic acid–ligand complexes. Three major types of binding data are colored in red (K_d), blue (K_i) and yellow (IC_{50}), respectively

interpretable. This model is useful as an improved filter before complicated computations or experimental work is carried out.

Liu *et al*. (2009) presented a quantitative model regarding the relationship between the intrinsic disorder in protein structure and its biological function. Their predictions were validated by genome-wide surveys on both the level of disorder in protein functions as defined by the GO ontology (The Gene Ontology Consortium, 2000) and binding data from PDBbind. Their results indicate that both catalytic and low-affinity proteins prefer ordered structures; while only high-affinity proteins can have disordered structures. Their results also suggest that binding affinity can be tuned by increasing structural disorder to maximize the specificity of promiscuous interactions. This study provides a new point of view for understanding the mechanism of protein–ligand interactions.

Recently, we conducted a statistical survey of the so-called characteristic interaction patterns (CIPs) on protein–protein binding interfaces from PDBbind (Li *et al*., 2013b). Common CIPs shared by different protein–protein binding interfaces were analyzed with the annotated functions from the GO ontology. We concluded that protein–protein interfaces having common CIPs with high conservation scores are usually associated with identical or similar biological functions. Based on the detection of CIPs at a protein–protein binding interface, we also proposed a fragment-based strategy for designing small-molecule inhibitors of protein–protein interactions by using the large library of protein–ligand complex structures in PDBbind. By comparing the CIPs at protein–protein interfaces with those at protein–ligand interfaces, appropriate fragments were truncated from known ligand molecules and then assembled to form complete ligand molecules. We applied this strategy to the *de novo* design of Bcl-2 family proteins, a familiar protein–protein interaction target (Ding *et al*., 2013). A number of designed molecules were synthesized and tested, and some of them proved to be effective binders of the Mcl-1 protein.

5 SUMMARY

The PDBbind database aims to provide a PDB-wide collection of binding data for various biomolecular complexes. Since its first public release in 2004, the PDBbind database has been updated on an annual basis. The latest release (version 2013) provides experimental binding data for 10 776 biomolecular complexes in PDB, including 8302 protein–ligand complexes and 2474 other types of complexes. All contents are freely accessible at the PDBbind-CN Web server (http://www.pdbbind-cn.org/). This database offers a valuable knowledge basis for computational studies of molecular recognition, structure-based drug design, statistical analysis of drug–target interaction network and many other applications. The PDBbind database is expected to reach a new level in terms of size and data quality in the coming years.

ACKNOWLEDGEMENTS

The authors are grateful to Dr. Ying-Duo Gao and Dr. M. Katharine Holloway from Merck for their careful proofreading.

Funding: This work was supported by the Chinese National Natural Science Foundation [Grant 81172984, 21072213 to Prof. R. Wang and Grant 21102168 to Dr. Y. Li], the Chinese Ministry of Science and Technology [863 High-Tech Grant 2012AA020308 to Prof. R. Wang], and the Science and Technology Development Fund of Macao SAR [Grant 055/2013/A2 to Prof. R. Wang]. This work was also supported by a postdoctoral fellowship from MSD China to Dr. Y. Li.

Conflict of interest: none declared.

REFERENCES

Benson,M.L. *et al*. (2008) Binding MOAD, A high-quality protein-ligand database. *Nucleic Acids Res*., **36**, D674–D678.

Berman,H.M. *et al*. (2003) Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol*., **10**, 980.

Block,P. *et al*. (2006) AffinDB: a freely accessible database of affinities for protein-ligand complexes from the PDB. *Nucleic Acids Res*., **34**, D522–D526.

Bolton,E. *et al*. (2008) PubChem: integrated platform of small molecules and biological activities. In: *Annual Reports in Computational Chemistry*, Chapter 12, Vol. 4. Elsevier, Oxford, pp. 217–240.

Cheng,T. *et al*. (2009) Comparative assessment of scoring functions on a diverse test set. *J. Chem. Inf. Model*., **49**, 1079–1093.

Ding,X. *et al*. (2013) De novo design, synthesis and evaluation of benzylpiperazine derivatives as highly selective binders of Mcl-1. *ChemMedChem*, **8**, 1986–2014.

Durrant,J.D. and McCammon,J.A. (2010) NNScore: a neural-network-based scoring function for the characterization of protein-ligand complexes. *J. Chem. Inf. Model*., **50**, 1865–1871.

Garcia-Sosa,A.T. *et al*. (2012) DrugLogit: Logistic discrimination between drugs and nondrugs including disease-specificity by assigning probabilities based on molecular properties. *J. Chem. Inf. Model*., **52**, 2165–2180.

Gaulton,A. *et al*. (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res*., **40**, D1100–D1107.

Hendlich,M. *et al*. (2003) Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. *J. Mol. Biol*., **326**, 607–620.

Hsin,K.-Y. *et al*. (2013) Combining machine learning systems and multiple docking simulation packages to improve docking prediction reliability for network pharmacology. *PLoS One*, **8**, e83922.

Hu,L. *et al*. (2005) Binding MOAD (mother of all databases). *Proteins*, **60**, 333–340.

Kellenberger,E. *et al*. (2006) sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank. *J. Chem. Inf. Model*., **46**, 717–727.

Laskowski,R.A. (2001) PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res*., **29**, 221–222.

Li,G.B. *et al*. (2013a) ID-Score: a new empirical scoring function based on a comprehensive set of descriptors related to protein-ligand interactions. *J. Chem. Inf. Model*., **53**, 592–600.

Li,Y. *et al*. (2013b) Mining the characteristic interaction patterns on protein-protein binding interfaces. *J. Chem. Inf. Model*., **53**, 2437–2447.

Li,Y. *et al*. (2010) Test MM-PB/SA on true conformational ensembles of protein-ligand complexes. *J. Chem. Inf. Model*., **50**, 1682–1692.

Li,Y. *et al*. (2014a) Comparative assessment of scoring functions on an updated benchmark: I. Compilation of the test set. *J. Chem. Inf. Model*., **54**, 1700–1716.

Li,Y. *et al*. (2014b) Comparative assessment of scoring functions on an updated benchmark: II. Evaluation methods and general results. *J. Chem. Inf. Model*., **54**, 1717–1736.

Liu,J. *et al*. (2009) Toward a quantitative theory of intrinsically disordered proteins and their function. *Proc. Natl Acad. Sci. USA*, **106**, 19819–19823.

Liu,Q. *et al*. (2013) Binding affinity prediction for protein-ligand complexes based on β contacts and B factor. *J. Chem. Inf. Model*., **53**, 3076–3085.

Liu,T. *et al*. (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res*., **35**, D198–D201.

Manetti,F. *et al*. (2008) N-(thiazol-2-yl)-2-thiophene carboxamide derivatives as Abl inhibitors identified by a pharmacophore-based database screening of commercially available compounds. *Bioorg. Med. Chem. Lett*., **18**, 4328–4331.

Neudert,G. and Klebe,G. (2011) DSX: a knowledge-based scoring function for the assessment of protein-ligand complexes. *J. Chem. Inf. Model*., **51**, 2731–2745.

Puvanendrampillai,D. and Mitchell,J.B.O. (2003) Protein ligand database (PLD): additional understanding of the nature and specificity of protein-ligand complexes. *Bioinformatics*, **19**, 1856–1857.

Roche,O. *et al*. (2001) Ligand-protein database: Linking protein-ligand complex structures to binding data. *J. Med. Chem*., **44**, 3592–3598.

Roth,B. *et al*. (2000) The multiplicity of serotonin receptors: uselessly diverse molecules or an embarrasment of riches? *Neuroscientist*, **6**, 252–262.

Schames,J.R. *et al*. (2004) Discovery of a novel binding trench in HIV integrase. *J. Med. Chem*., **47**, 1879–1881.

Tang,Y.T. *et al*. (2011) PHOENIX: a scoring function for affinity prediction derived using high-resolution crystal structures and calorimetry measurement. *J. Chem. Inf. Model*., **51**, 214–228.

Tsai,C.-J. *et al*. (1996) A data set of protein-protein interfaces generated with sequence-order-independent comparison technique. *J. Mol. Biol*., **260**, 604–620.

The Gene Ontology Consortium. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet*., **25**, 25–29.

Trott,O. and Olson,A.J. (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem*., **31**, 455–461.

Wang,R. *et al*. (2002) Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput. Aided. Mol. Des*., **16**, 11–26.

Wang,R. *et al*. (2004) The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem*., **47**, 2977–2980.

Wang,R. *et al*. (2005) The PDBbind database: methodologies and updates. *J. Med. Chem*., **48**, 4111–4119.

Wang,Y. *et al*. (2014) PubChem bioassay: 2014 update. *Nucleic Acids Res*., **42**, D1075–D1082.

Wlodawer,A. (2002) Rational approach to AIDS drug design through structural biology. *Annu. Rev. Med*., **53**, 595–614.

Yamaguchi,A. *et al*. (2004) Het-PDB Navi.: a database for protein-small molecule interactions. *J. Biochem*., **135**, 79–84.

Zheng,Z. *et al*. (2011) Ligand identification scoring algorithm (LISA). *J. Chem. Inf. Model*., **51**, 1296–1306.

Zheng,Z. *et al*. (2013) Development of the knowledge-based and empirical combined scoring algorithm (KECSA) to score protein-ligand interactions. *J. Chem. Inf. Model*., **53**, 1073–1083.

Zilian,D. and Sotriffer,C.A. (2013) SFCscoreRF: a random forest-based scoring function for improved affinity prediction of protein-ligand complexes. *J. Chem. Inf. Model*., **53**, 1923–1933.