Beginner's Guide to the PDBbind Database (v.2017)

The PDBbind database provides a comprehensive collection of experimentally measured binding affinity data for the biomolecular complexes in the Protein Data Bank (PDB). This type of knowledge is the much needed basis for many computational and statistical studies on molecular recognition. PDBbind was first released to the public in May 2004. Over 4,500 users from over 70 countries have already registered to use this database. The PDBbind database is now updated annually to keep up with the growth of PDB. The current release is **version 2017**.

What information does PDBbind provide?

- **Binding affinity data:** Originally, PDBbind only considered the complexes formed between proteins and small-molecule ligands. Other types of biomolecular complexes in PDB have been covered by PDBbind as well since 2008. This release contains binding data (K_d, K_i & IC_{50} values) for protein-ligand (14,761), protein-protein (2,181), protein-nucleic acid (837), and nucleic acid-ligand (121) complexes. All binding data are curated by ourselves from over 32,000 original references.

- **Processed structural files for download:** PDBbind also provides processed "clean" structural files for most of the protein-ligand complexes in this release. In brief, the biological unit of each complex is split into a protein molecule (in PDB format) and a ligand molecule (in Mol2 and SDF format). Atom/bond types on the ligand molecule are assigned as appropriate and examined manually. These structural files can be readily utilized by most molecular modeling software, which are wrapped in a data package for download.

- **Web-based display and analysis tools:** The user can access PDBbind through a web-based portal at http://www.pdbbind-cn.org/. Registration is free for academic as well as industrial users. On the PDBbind-CN web site, basic information of each complex is summarized on a single page. Text-based and structure-based search among the contents of PDBbind is also enabled. This web site actually provides structural information for all valid protein-ligand complexes in the Protein Data Bank, not limited to those with known binding data.

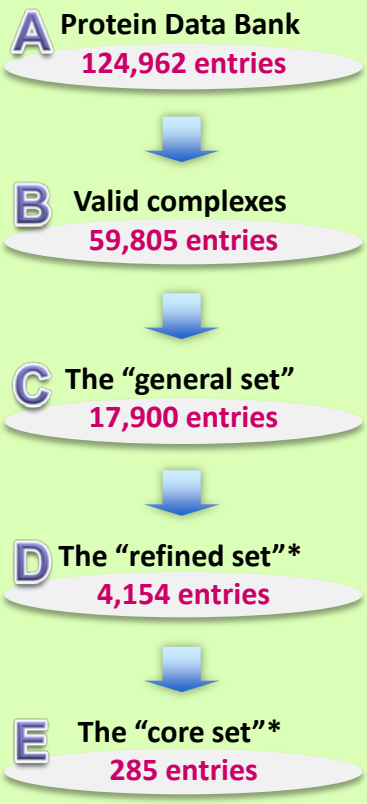
Basic Information of the PDBbind Database*

Version	Entries In PDB	Complex with binding data	Protein-ligand complex	Protein-protein complex	Protein-nucleic acid complex	Nucleic acid-ligand complex
2004	28,991	6,847	2,276	2,276	N.A.	N.A.
...	...	...	...	...	...	
2013	87,085	10,776	8,302	1,804	587	83
2014	96,952	12,995	10,656	1,592	660	87
2015	105,183	14,620	11,987	1,807	717	109
2016	114,344	16,179	13,308	1,976	777	118
2017	124,962	17,900	14,761	2,181	837	121

*: Information of some earlier versions (v.2005 – v.2012) are not included in this table due to space limit.

Basic structure of the PDBbind data set

PDBbind is compiled through a stepwise process. It has a hierarchical structure as follows.



* Only complexed formed by proteins and small-molecule ligands are considered in this data set.

(A) The PDBbind v.2017 is based on the contents of PDB officially released at the first week of 2017, which contained totally **124,962** experimentally determined structures. Theoretical models are not considered.

(B) The entire PDB was screened by a set of computer programs to identify four major categories of molecular complexes, including protein-small ligand, nucleic acid-small ligand, protein-nucleic acid and protein-protein complexes. This step identified a total of **59,805** entries as valid complexes.

(C) The primary reference of each complex was examined to collect experimentally determined binding affinity data (K_d, K_i and IC_{50}) of the given complex. Binding data for **17,900** complexes were collected in this way. They are the main body of the PDBbind database, which is referred to as the "**general set**".

(D) As an additional feature, a "**refined set**" was compiled to select the protein-ligand complexes with better quality out of the general set. A number of filters regarding binding data, crystal structures, as well as the nature of the complexes were applied in selection (see ref.3 below for details). The refined set in this release consists of **4,154** protein-ligand complexes.

(E) A "**core set**" is also included in previous releases of PDBbind. Compilation of the "core set" aims at providing a relatively small set of high-quality protein-ligand complexes for docking/scoring studies. In particular, the "core set" has served as the primary test set in our Comparative Assessment of Scoring Functions (CASF) benchmark. The "core set" is a representative, non-redundant subset of the "refined set". This data set is not updated annually as the PDBbind database itself. Besides, it is much more than a list of protein-ligand complexes but with a huge amount of derivative data. Researchers may obtain the core set by downloading the CASF data package from our PDBbind-CN web site (http://www.pdbbind-cn.org/casf.asp).

References and notes

The PDBbind database is currently maintained by Prof. Renxiao Wang's group at the Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences. To cite the PDBbind database, please refer to the following references:

(1) Liu, Z.H. et al. *Acc. Chem. Res.* 2017, *50*, 302-309. (PDBbind v.2016)
(2) Liu, Z.H. et al. *Bioinformatics*, 2015, *31*, 405-412. (PDBbind v.2014)
(3) Yan, L.; et al. *J. Chem. Inf. Model.*, 2014, *54*, 1700-1716. (PDBbind v.2013)
(4) Cheng, T. J.; et al. *J. Chem. Inf. Model.*, 2009, *49*, 1079-1093. (PDBbind v.2007)
(5) Wang, R. X.; et al. *J. Med. Chem.* 2005, *48*, 4111-4119; *J. Med. Chem.* 2004, *47*, 2977-2980. (early versions)