Beginner's Guide to the PDBbind Database (v.2020)

The PDBbind database provides a comprehensive collection of experimental binding affinity data for the biomolecular complexes in the Protein Data Bank (PDB). This type of information is much needed by various computational and statistical studies on molecular recognition. A prototype of PDBbind was first released to the public in May 2004. Since 2007, this database has been updated regularly on an annual base to keep up with the growth of PDB. The current release is **version 2020**.

What does PDBbind provide?

- **Binding data:** The main value of PDBbind is the collection of experimentally measured binding affinity data (in form of K_d, K_i or IC_{50} values) that match the biomolecular complex structures in PDB. Originally, PDBbind only included the complexes formed between proteins and small-molecule ligands. Other types of complexes have been covered by PDBbind since 2008. This release contains binding data for protein-ligand, protein-protein, protein-nucleic acid, and nucleic acid-ligand complexes (see the table below for details). All binding data are curated by ourselves from original references rather than being copied from some third-party resources. So far, a total of ~40,500 references have been checked for this purpose.
- **Processed structural files:** As an additional value, PDBbind also provides processed "clean" structural files for all of the protein-ligand complexes included in this release, which can be readily utilized by most molecular modeling software. In brief, the biological unit of each complex is split into a protein molecule (saved in the PDB format) and a ligand molecule (saved in the Mol2 and SDF format). Atom/bond types on the ligand molecule are assigned by a special computer program and then examined and corrected manually. All processed structural files of protein-ligand complexes are wrapped in a data package that can be downloaded from the PDBbind-CN web site.
- **Web-based display and analysis tools:** The user can access PDBbind through a web portal at http://www.pdbbind-cn.org/. Registration is free for both academic and commercial users. On this web site, basic information of each complex is summarized on a single page. Text-based and structure-based search among the contents of PDBbind are 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.

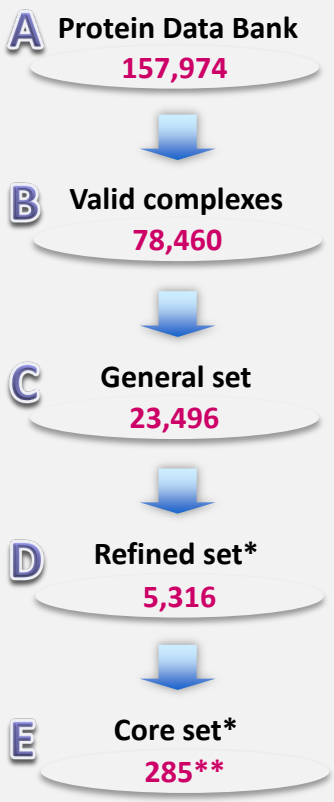
Basic Information of the PDBbind Database*

Version	Entries In PDB	All complex with binding data	Protein-ligand complex	Protein-protein complex	Protein-nucleic acid complex	Nucleic acid-ligand complex
2004	28,991	2,276	2,276	N.A.	N.A.	N.A.
...	...	...	...	...	...	...
2017	124,962	17,900	14,761	2,181	837	2017
2018	135,859	19,588	16,151	2,416	896	2018
2019	146,836	21,382	17,679	2,594	973	136
2020	157,974	23,496	19,443	2,852	1,052	149

*: Some earlier versions (v.2005 – v.2016) are not included in this table due to space limit.

The hierarchical structure of the PDBbind data sets

The data sets in PDBbind are compiled through a stepwise process as follows.



* This data set contains only complexes formed between proteins and small-molecule ligands.
** Number for core set v.2016.

(A) The PDBbind v.2020 is based on the contents of PDB officially released at the first week of year 2020, which contained a total of **157,974** experimentally determined structures. Theoretical models are not considered by us.

(B) The entire PDB is 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 identifies a total of **78,460** PDB entries as valid complexes.

(C) The primary reference of each complex is examined to collect experimentally determined binding affinity data (K_d, K_i or IC_{50}) of the given complex. Binding data for **23,496** complexes have been collected in this way. They form the main contents of the PDBbind database, which is referred to as the "**general set**".

(D) An additional "**refined set**" is 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 are applied to selection (see ref.2 below for details). The refined set in this release consists of **5,316** protein-ligand complexes.

(E) The "**core set**" is also a derivative data set based on the contents of PDBbind. Compilation of the PDBbind core set aims at providing a relatively small set of high-quality protein-ligand complexes for validating docking/scoring methods. In particular, this data set has served as the primary test set in the popular Comparative Assessment of Scoring Functions (CASF) benchmark developed by our group. The PDBbind core set is not included in the PDBbind data package any more because it is not updated annually as PDBbind itself. Researchers can obtain the PDBbind core set by downloading the CASF data package at http://www.pdbbind-cn.org/casf.asp. The latest available version of the PDBbind core set is included in CASF-2016.

References and notes

The PDBbind database is currently maintained by Prof. Renxiao Wang's group at the School of Pharmacy, Fudan University. To cite the PDBbind database, please refer to the following publications:

(1) Su, M.Y. et al. *J. Chem. Inf. Model.* 2019, *59*, 895-913. (CASF-2016)
(2) Liu, Z.H. et al. *Acc. Chem. Res.* 2017, *50*, 302-309. (PDBbind v.2016)
(3) Liu, Z.H. et al. *Bioinformatics*, 2015, *31*, 405-412. (PDBbind v.2014)
(4) Yan, L.; et al. *J. Chem. Inf. Model.*, 2014, *54*, 1700-1716. (PDBbind v.2013 & CASF-2013)
(5) Cheng, T. J.; et al. *J. Chem. Inf. Model.*, 2009, *49*, 1079-1093. (PDBbind v.2007 & CASF-2007)
(6) Wang, R. X.; et al. *J. Med. Chem*. 2005, *48*, 4111-4119; *J. Med. Chem.* 2004, *47*, 2977-2980. (proto-type)