Beginner's Guide to the PDBbind Database (v.2015)

The PDBbind database provides a comprehensive collection of experimentally measured binding affinity data for the biomolecular complexes in the Protein Data Bank (PDB). It provides the much needed knowledge basis for many computational and statistical studies on molecular recognition. It was first released to the public in May 2004. More than 2,800 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 2015**.

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 (11,987), protein-protein (1,807), protein-nucleic acid (717), and nucleic acid-ligand (109) complexes. All binding data are curated by ourselves from over 27,000 original references rather than copied from other data sources.

- **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, and they are included 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/. 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 (>12,000 unique small-molecule ligands), not limited to those with known binding data.

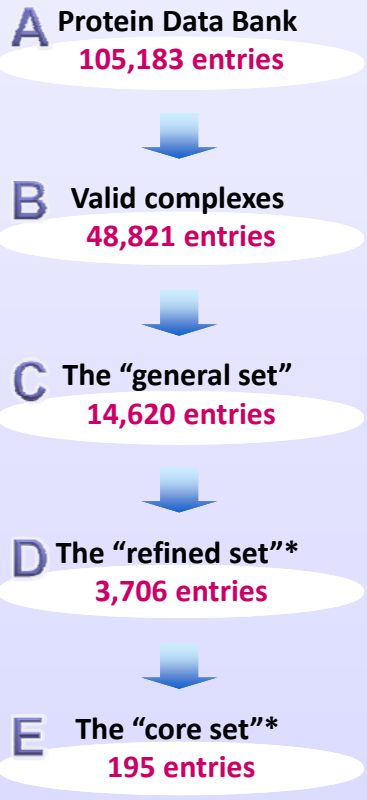
History of the PDBbind Database *

Version	Entries in PDB	Complexes considered	Complexes with binding data		
			General Set	Refined Set	Core Set
2004	28991	6,847	2,276	1,091	231
…	…	…	…	…	…
2011	70,224	30,259	7,986	2,455	216
2012	78,235	34,180	9,308	2,897	201
2013	87,085	38,918	10,776	2,959	195
2014	96,952	44,569	12,995	3,446	195
2015	105,183	48,821	14,260	3,706	195

*: Information of some earlier versions (v.2005 – v.2010) 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.2015 is based on the contents of PDB officially released at the first week of 2015, which contained totally **105,183** 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 **48,821** 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 **14,620** 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 references below for details). The refined set in this release consists of **3,706** protein-ligand complexes.

(E) A "**core set**" was further complied as a high-quality benchmark for evaluating various docking/scoring methods. It was compiled through a systematic, non-redundant sampling of the refined set with emphasis on the diversity in structures and binding data. In short, the refined set was clustered by protein sequence similarity using a cutoff of 90%. As for each cluster containing at least five members, the one with the highest binding constant, the one with the lowest binding constant, and the one with a medium binding constant were selected as the representatives of this cluster. Note that the core set in this release has not been updated. It is still the same as the one in v.2013, which consists of a total of **195** protein-ligand complexes in 65 clusters.

References and Notes

The PDBbind database was originally developed by Prof. Shaomeng Wang's group at the University of Michigan. It is currently maintained by Prof. Renxiao Wang's group at the Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences under a mutual agreement. To cite the PDBbind database, please refer to the following references:

(1) Liu, Z.H. et al. *Bioinformatics*, **2015**, *31*, 405-412.
(2) Yan, L.; et al. *J. Chem. Inf. Model.*, **2014**, *54*, 1700-1716.
(3) Cheng, T. J.; et al. *J. Chem. Inf. Model.*, **2009**, *49*, 1079-1093.
(4) Wang, R. X.; et al. *J. Med. Chem.* **2005**, *48*, 4111-4119; *J. Med. Chem.* **2004**, *47*, 2977-2980.