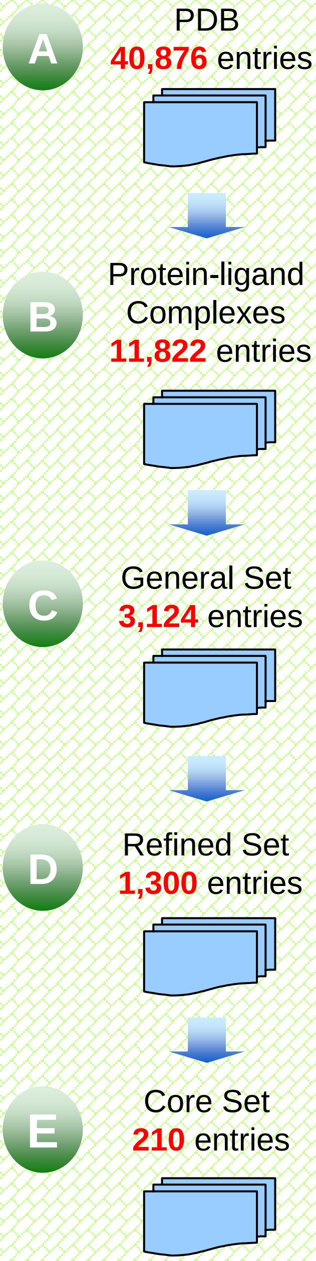
A Brief Introduction to the PDBbind Database v.2007

The PDBbind database, as prompted by its name, is a collection of the experimentally measured binding affinities exclusively for the protein-ligand complexes available in the Protein Data Bank (PDB). It thus provides a link between energetic and structural information of those complexes and may be of great value to various molecular recognition studies. Since its first public release in May 2004, several hundred users around the world have already registered to use the PDBbind database. The current release is **version 2007**.



A
PDB
40,876 entries

B
Protein-ligand Complexes
11,822 entries

C
General Set
3,124 entries

D
Refined Set
1,300 entries

E
Core Set
210 entries

In order to utilize the PDBbind database smartly, it is important to get familiar with its basic construction:

(A) The PDBbind v.2007 is based on the contents of PDB, which contains a total of **40,876** experimentally determined structures officially released before Jan 1st, 2007. Theoretical models are not considered by PDBbind.

(B) A sophisticated algorithm is designed to identify the complexes formed between proteins and small organic molecules. Please refer to the references cited in the end of this brochure for details. A total of **11,822** entries are identified as valid protein-ligand complexes in this release.

(C) The primary reference of each complex is reviewed manually to collect the experimentally determined binding affinity (IC_{50}, K_i, or K_d) of the complex. Binding affinity data of a total of **3,124** complexes are compiled in this way from nearly 7,000 references. They form the **"general set"** of PDBbind.

(D) A **"refined set"** is further selected out of the general set with a number of criteria. These criteria address the quality of binding data as well as complex structures. Each complex in the refined set has been double-checked to ensure that binding affinity matches the structure from PDB. The refined set is designed to serve as a high-quality standard data set for theoretical studies on protein-ligand binding. The refined set in this release consists of a total of **1,300** entries.

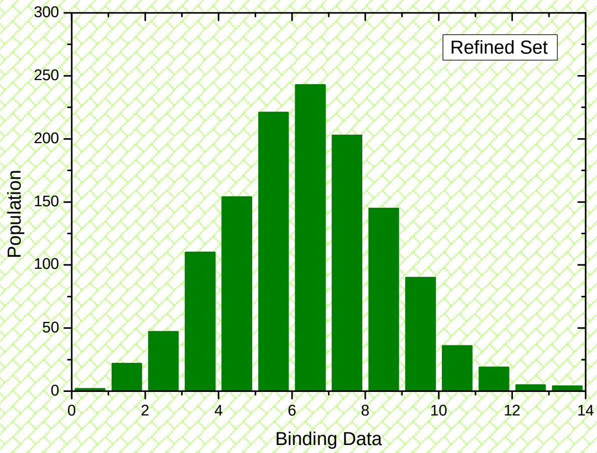
(To be continued on next page)

A given protein-ligand complex is accepted into the refined set if it meets all of the following criteria:

History of the PDBbind Database

Version	Protein-ligand Complexes	Entries in the General Set	Entries in the Refined Set	Entries in the Core Set
2002	5671	1446	800	---
2003	5897	1763	900	---
2004	6847	2276	1091	231
2005	9775	2756	1296	288
2006*	9775	2632	1122	234
2007	11822	3124	1300	210

*: V.2006 is in fact a correction on v.2005.



Distribution of the Binding Data in the Refined Set

1. It is a crystal structure with an overall resolution $\le$ 2.5 angstrom.
2. It is a "clean", binary complex formed between one protein and one ligand through non-covalent binding.
3. It has an experimentally determined K_d or K_i value.
4. The protein and the ligand used in binding assay exactly match the ones observed in the complex structure.
5. The ligand does not consist of any element other than C, N, O, S, P, H, and halogens, and its molecular weight is lower than 1,000.
6. There are no unnatural amino acid residues in the binding site on the protein.

For users' convenience, PDBbind provides processed structural files for each complex in the refined set, which can be readily utilized by molecular modeling software. The **biological unit** of each complex is split into a protein saved in the PDB format, and a ligand saved in the Mol2 format and the SD format. Atom types and bond types on the ligand are assigned carefully as appropriate.

(E) The refined set is further grouped into clusters by protein sequence similarity. BLAST is used to compute the similarity between any two sequences, and a cutoff of 90% is used in clustering. Each cluster normally consists of complexes formed by a particular type of protein. A total of **70** clusters from the refined set are found to contain at least four members. The one with the highest binding affinity, the one with the lowest binding affinity, and the one with a medium binding affinity in each cluster are selected as the representatives of this cluster. All of these representatives, **210** complexes in total, form the "**core set**" of PDBbind. Note that the core set is strictly a subset of the refined set. The core set is designed to provide a non-redundant sampling of the refined set, which may be more appropriate and more convenient to use for certain studies.

The PDBbind database is developed by a collaboration between Prof. Shaomeng Wang's group at the University of Michigan and Prof. Renxiao Wang's group at the Shanghai Institute of Organic Chemistry. To cite the PDBbind database, please refer to:

1. Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. *J. Med. Chem*. **2005**, *48*, 4111-4119.
2. Wang, R.; Fang, X.; Lu, Y.; Wang, S. *J. Med. Chem.* **2004**, *47*, 2977-2980.