Forging the Basis for Developing Protein-Ligand Interaction Scoring Functions

Zhihai Liu, Minyi Su, Li Han, Jie Liu, Qifan Yang, Yan Li*, and Renxiao Wang*

State Key Laboratory of Bioorganic and Natural Products Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai, P. R. China

Email: liuhai@mail.sioc.ac.cn

INTRODUCTION

In structure-based drug design, scoring functions are widely used for fast evaluation of protein-ligand interactions. Regardless of their technical difference, scoring functions all need data sets combining protein-ligand complex structures and binding affinity data for parameterization and validation. However, data sets of this kind used to be rather limited in terms of size and quality. On the other hand, standard metrics for evaluating scoring function used to be ambiguous, which do not directly reflect the genuine quality of scoring functions.

In a recently published paper (*Acc. Chem. Res.*, **2017**, 50: 302-309), we describe our long-lasting efforts to overcome these obstacles, which involve two related projects. On the first project, we have created the PDBbind database. It is the first database that systematically annotates the protein-ligand complexes in the Protein Data Bank (PDB) with experimental binding data. This database has been updated annually since its first public release in 2004. The latest release (v2016) provides binding data for 16179 biomolecular complexes in PDB. Data sets provided by PDBbind have been applied to many computational and statistical studies. In particular, it has become a major data resource for scoring function development. On the second project, we have established the Comparative Assessment of Scoring Functions (CASF) benchmark for scoring function evaluation. Our key idea is to decouple the "scoring" process from the "sampling" process, so scoring functions can be tested in a relatively pure context to reflect their quality. Importantly, CASF is designed to be an open-access benchmark.

METHODS

A **Protein Data Bank (114,344 entries)**

(A)The PDBbind v.2016 is based on the contents of PDB officially released on Jan 1st, 2016, which contains a total of **114,344** experimentally determined structures. Theoretical models are not considered by PDBbind.

B **Valid biomolecular complexes (53,838 entries)**

(B)The entire PDB is screened by a set of computer programs to identify four major categories of biomolecular complexes, including protein-small ligand, nucleic acid-small ligand, protein-nucleic acid, and protein-protein complexes. A total of **53,838** entries are identified as valid biomolecular complexes in this release.

C **The general set (16,179 entries)**

(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 **16,179** complexes are collected in this way out of 35,000 references. They are the main body of PDBbind, which is called the **"general set"**.

D **The refined set* (4,057 entries)**

(D)A **"refined set"** is compiled to provide a high-quality data set of protein-small ligand complexes especially for docking/scoring studies. The complexes in the refined set are selected out of the general set with a number of criteria addressing the quality of binding data as well as structures. Each qualified complex has been double-checked to ensure its binding data matches its structure from PDB. The refined set in this release is estimated to consist of a total of **4,057** entries.

E **The core set* (285 entries)**

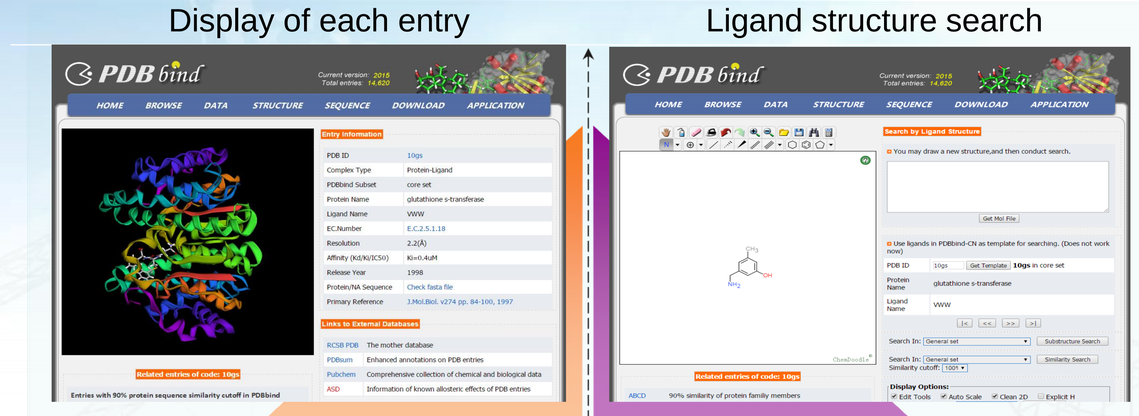
(E)A **"core set"** is further complied to provide a non-redundant sampling of the refined set. Briefly, the refined set is clustered by protein sequence similarity using a cutoff of 90%. In the latest CASF-2016 work, the core set will consist of a total of 57 clusters, 285 protein-ligand complexes in total (65 clusters and 195 complexes in CASF-2013).

* Only protein-ligand complexes are considered in this data set. New core set will be released with update of CASF benchmark dataset since v2014.

WEB INTERFACE

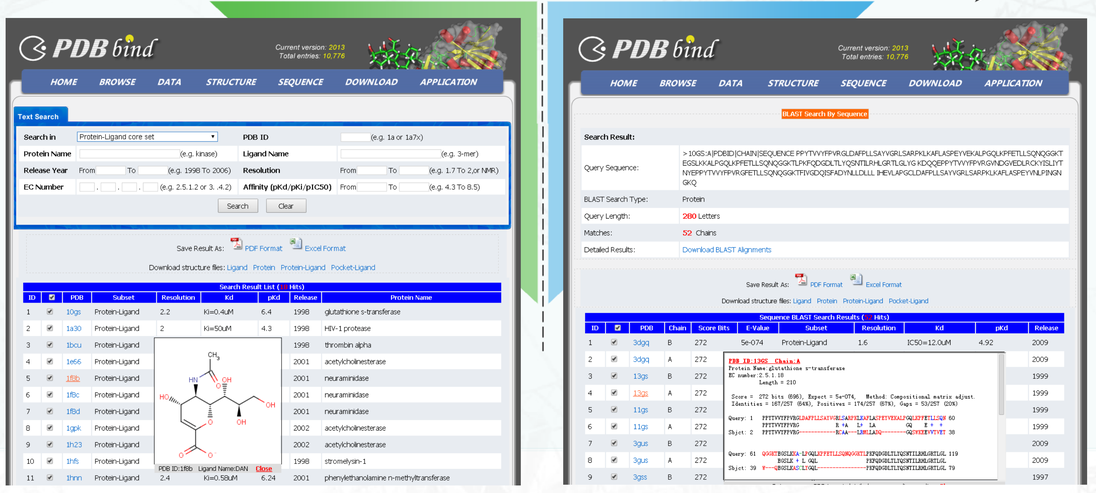
Display of each entry

Ligand structure search



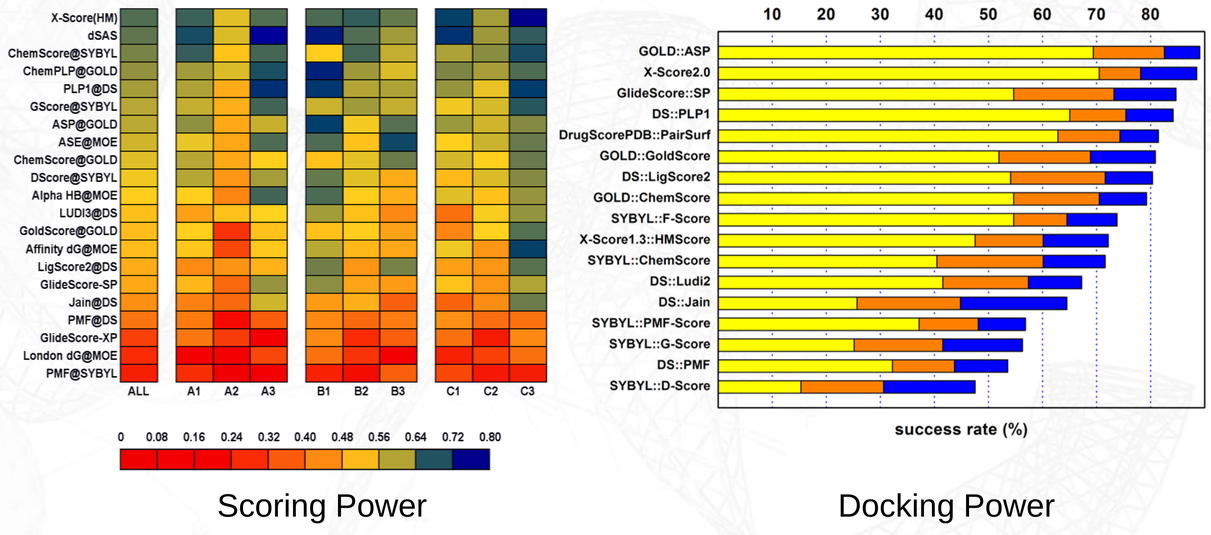
Text-based Search

Sequence-based search



CASF Benchmark

In our latest work on this track, i.e. CASF-2016, the performance of a scoring function was quantified in four aspects, including "scoring power", "ranking power", "docking power", and "screening power". All four performance tests were conducted on a test set containing **285** high-quality protein−ligand complexes selected from PDBbind. A panel of over 20 standard scoring functions were tested as demonstration. Importantly, CASF is designed to be an open-access benchmark, with which scoring functions developed by different researchers can be compared on the same grounds. Indeed, it has become a popular choice for scoring function validation in recent years.



Scoring Power

Docking Power

REFERENCES

[1] Yan Li, Minyi Su, Zhihai Liu, Jie Li, Jie Liu, Li Han, Renxiao Wang *, *Nature Protocols*, **2017**, In press.

[2] Zhihai Liu, Minyi Su, Li Han, Jie Liu, Qifan Yang, Yan Li*, and Renxiao Wang*, *Accounts of Chemical Research*, **2017**, 50 (2): 302-309.

[3] Zhihai Liu, Yan Li, Li Han, Jie Li, Jie Liu, Zhixiong Zhao, Wei Nie, Yuchen Liu and Renxiao Wang, *Bioinformatics*, **2015**, 31 (3): 405-412.

[4] Li, Yan; Liu, Zhihai; Li, Jie; Han, Li; Liu, Jie; Zhao, Zhi-Xiong; Wang, Renxiao *, *J. Chem. Inf. Model.* **2014**, 54, 1700-1716.

[5] Li, Yan; Han, Li; Liu, Zhihai; Wang, Renxiao *, *J. Chem. Inf. Model.* **2014**, 54, 1717-1736.

[6] Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. , *J. Med. Chem.,* **2005**; 48(12); 4111-4119.

[7] Wang, R.; Fang, X.; Lu, Y.; Wang, S. , *J. Med. Chem.,* **2004**; 47(12); 2977-2980.