The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures

Renxiao Wang, Xueliang Fang, Yipin Lu, and Shaomeng Wang*

Department of Internal Medicine and Comprehensive Cancer Center, University of Michigan Medical School, and Department of Medicinal Chemistry, University of Michigan College of Pharmacy, 3316 CCGC Building, 1500 E. Medical Center Drive, Ann Arbor, Michigan 48109-0934

Received November 19, 2003

Abstract: We have screened the entire Protein Data Bank (Release No. 103, January 2003) and identified 5671 protein−ligand complexes out of 19 621 experimental structures. A systematic examination of the primary references of these entries has led to a collection of binding affinity data (K_d, K_i, and IC_{50}) for a total of 1359 complexes. The outcomes of this project have been organized into a Web-accessible database named the PDBbind database.

One of the key issues in structure-based drug discovery is the prediction of binding affinities, which is often referred to as the “scoring” problem. Among a whole spectrum of methods developed for solving this problem, a group of approaches, called “scoring functions”, have gained popularity.[1−18] Because of their high speed and reasonable accuracy, scoring functions find major applications in molecular docking studies such as high-throughput virtual library screening[1] and they are gradually replacing the role of conventional force field computation in such studies.

Because of their empirical nature, scoring functions are usually developed and validated using a set of protein−ligand complexes with experimentally determined binding affinities and three-dimensional structures. Here, we use a term PLEXBAS to abbreviate “*p*rotein−*l*igand compl*ex*es with both known *b*inding *a*ffinity and three-dimensional *s*tructure”. Table 1 gives a summary of the major scoring functions published since 1990. One can see from this table that most of today’s scoring functions were developed and validated with a rather limited number of PLEXBAS. The largest set of PLEXBAS employed in those scoring functions was below 250. It is reasonable to expect that a larger, high-quality set of PLEXBAS will benefit scoring function development and may lead to more accurate scoring functions.

Table 1. Major Scoring Functions Published to Date

approaches	year published	no. of PLEXBAS used in training and test sets	ref
Böhm (Score1)	1994	54	2
Jain	1996	34	3
Head et al. (VALIDATE)	1996	65	4
Eldridge et al. (ChemScore)	1997	112	5, 6
Böhm (Score2)	1998	94	7
Wang et al. (SCORE)	1998	181	8
Muegge et al. (PMF)	1999	225	11−13
Mitchell et al. (BLEEP)	1999	90	14, 15
Gohlke et al. (DrugScore)	2000	>100	16, 17
Cozzini et al. (HINT)	2002	53	9
Ishchenko et al. (SMoG2001)	2002	119	18
Wang et al. (X-Score)	2002	230	10

Apparently, the lack of PLEXBAS is not due to the lack of available three-dimensional structures. At the time when this manuscript was being prepared, over 24 000 structures had been deposited into the Protein Data Bank (PDB),[19] among which there were more than 6000 protein−ligand complexes. However, the binding affinity data that match these complex structures are difficult to find because they are scattered in the scientific literature. The PLEXBAS sets listed in Table 1 were typically compiled by assembling other researchers’ previous compilations rather than collecting data directly from the original references. This approach has two major drawbacks. First, it will probably never give a real boost to the collection of known PLEXBAS because most entries are copies of data collected earlier; new entries are added only occasionally. This explains the slow growth in the total number of known PLEXBAS (see Table 1). Second, data quality problems arise because there is little rigor associated with such data collection. When people copy the data of others, they often do not verify the data with original publications themselves, and so errors will be propagated. For example, when we attempted to confirm the binding affinity data of the PLEXBAS sets listed in Table 1, we found that sometimes IC_{50} values were misrepresented as K_d or K_i values, and there were also typographical errors. In addition, not every PLEXBAS is suitable for developing or validating scoring functions. For example, some are actually covalently bound complexes, others have a major cofactor molecule bound with the ligand inside the same binding pocket, and still others have low structural resolution. If such unacceptable entries were excluded, there would be even fewer PLEXBAS available for scoring function development.

Since lack of a large, high-quality set of PLEXBAS has become a bottleneck for developing more accurate scoring functions, we have decided to screen the entire PDB to identify all of the complexes formed between proteins and small organic ligand molecules and then collect the experimentally measured binding affinity data for these complexes from the scientific literature. Our current work was based on PDB Release No. 103 (January 2003), which contained all of the entries released by PDB before 2003. It had a total of 19 621 experimental structures and 551 theoretical models. Only experimentally determined structures were considered in this work.

The first step of our work was to identify the protein−ligand complexes in PDB because PDB itself does not provide such a classification. Surprisingly, this task turned out to be nontrivial. At the early stage of our work, we tried to search PDB with key-word-based queries such as “complex”, “complexation”, “ligand”, and “inhibitor” with the hope that this approach would extract all of the complexes in PDB, albeit with some

* To whom correspondence should be addressed. Phone: (734) 615-0362. Fax: (734) 647-9647. E-mail: shaomeng@med.umich.edu.

10.1021/jm030580l CCC: $27.50 © 2004 American Chemical Society
Published on Web 05/04/2004

false hits. However, we soon found that this approach missed too many *bona fide* complexes. The reason is very simple: not every deposited complex structure is required to use a key word like "complex" or "complexation" in the PDB file. Application of more carefully designed text-based queries may reduce the number of missing hits, but it was expected that this approach would not solve the problem completely, and consequently, we adopted a more elaborate approach for this task. First, we filtered out all of the apparent noncomplex structures, including proteins, nucleic acids, and carbohydrates. It is noted that a structure is classified as "noncomplex" if it contains only inorganic or solvent molecules in addition to the main molecule. Second, we filtered out protein−protein complexes and protein−nucleic acid complexes because these two types of complexes were not our primary interest. However, we did include the complexes formed between proteins and oligopeptides because oligopeptides (and their mimics) are a very important class of molecules in drug discovery. Unlike proteins, most oligopeptides do not adopt stable secondary structures by themselves and may be considered as common organic molecules for the purpose of developing scoring functions. In our study, peptides containing 10 or fewer amino acid residues were defined as oligopeptides. All of the above classifications were performed by computer programs written for this project.

After the above classifications, each remaining structure contained a protein molecule and at least one organic molecule. However, since not every organic molecule attached to a protein is necessarily a valid ligand bound to the protein, we have applied several criteria to filter out "invalid" ligand molecules as follows: (i) A valid ligand molecule should be specific. Therefore, if a ligand molecule is observed in a large number of PDB structures, it is probably not a valid ligand. Such nonspecific ligands include buffer and solvent molecules, such as DMSO and ethanol, and some other types of molecules, such as *N*-acetyl-D-glucosamine and α-D-mannose. (ii) Some organic cofactors such as heme, NAD, CoA, and FAD can also be found in a large number of PDB structures. Unlike those nonspecific molecules in the first category, the molecules in the second category have important biological functions and are usually an indispensable part of the protein−ligand complex. Their binding affinities to their host proteins can be measured and have indeed been measured in many cases. However, these molecules are not "drug-like" and thus were not considered as valid ligand molecules in our study. (iii) We specified that a valid ligand molecule should contain at least six non-hydrogen atoms and its molecular weight should not exceed 1000. Placing a limit on size is another measure to ensure the "druglikeness" of the ligand molecules under examination. The above three criteria eliminated a large number of undesired entries. To ensure quality, all of the remaining entries were visually examined to confirm that they met all the criteria we specified. This task was greatly facilitated by utilizing the information available on the PDB web site and its linked web sites. Our final list of protein−ligand complexes contained 5671 entries, and Figure 1 provides the distribution of these entries sorted by their release year.

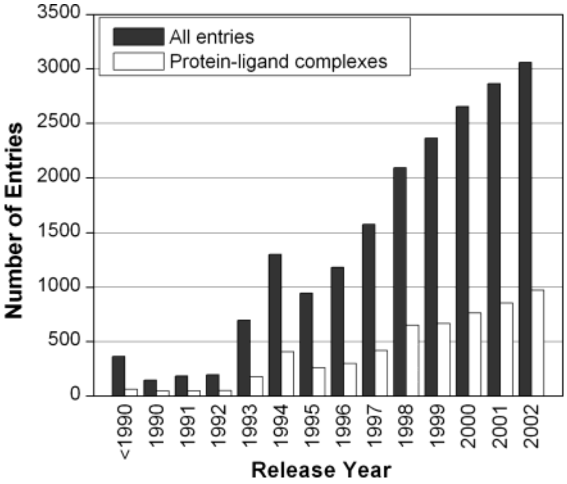

Figure 1. Protein−ligand complexes defined in the PDBbind database (sorted by their year of release from PDB).

The second step of our work was to search for binding affinity data of these 5671 protein−ligand complexes in the literature. The key problem was how to identify and retrieve the right references out of a large body of scientific literature. Fortunately, almost every structure deposited in PDB is associated with a primary reference, which can be read from the corresponding PDB file. If a PDB structure is indeed a protein−ligand complex and if its authors have also measured the binding affinity for the complex, it is expected that the result would be reported in the original publication. Alternatively, if the binding affinity of the complex was measured previously, it may be cited in the authors' publication of the PDB structure. In either case, if the binding affinity of a complex has indeed been measured, it is likely that such information can be retrieved from the primary reference listed in the PDB file. The primary reference information of each complex was retrieved from its associated PDB file, and the corresponding publication was requested from the University of Michigan library system; this resulted in a total of 3348 published papers. References for 336 complexes could not be obtained because they were either not indicated in their corresponding PDB files or were not available from our library system. All of the obtained publications were manually reviewed, which turned out to be the most time-consuming step in this project. During this process, we recorded three major forms of binding affinity data: dissociation constant (K_d), inhibition constant (K_i), and concentration at 50% inhibition (IC_{50}). We did not record enzymatic kinetic parameters such as K_m and k_{cat}, although they also reflect the binding affinity between an enzyme and its substrate. If binding affinity data of a given complex were available in different forms, we applied a priority order of $K_d > K_i > IC_{50}$ and recorded only the data with the highest priority. If binding affinities of a given complex were measured under different temperatures and pH levels, we recorded only the results measured at room temperature and at neutral pH or in assay conditions closest to room temperature and neutral pH. Finally, we collected K_d values for 431 entries, K_i values for 714 entries, and IC_{50} values for 214 entries from those 3348 papers. The overall "yield ratio" at this step was 25% ((431 + 714 + 214)/(5671−336) = 1359/5335).

The last step of our work was to build a high-quality set out of the 1359 PLEXBAS identified in the previous step. Since this data set will be applied to scoring and docking studies, a number of additional criteria were applied to filter out entries that may not be fully suitable for the purpose of scoring and docking studies. To enter the final refined list, a protein−ligand complex must meet the following five criteria: (i) It must have a K_d or K_i value. It is well-known that K_d and K_i are equilibrium constants and thus thermodynamic properties. In contrast, IC_{50} values largely depend on the binding assay conditions and only have relative meanings for comparison of the binding affinities of molecules measured in the same assay.[20] Thus, for a data set that includes diverse families of proteins and ligand molecules, IC_{50} values should be left out. (ii) It must be a noncovalently bound complex. We wrote a computer program to examine if a given complex may be covalently bound by considering interatomic distance as well as chemical feasibility. (iii) It may not have more than one ligand molecule bound in the binding pocket of the protein. This often occurs when one ligand molecule is a substrate analogue while the other is a cofactor. It becomes much more complicated in such cases to define the binding affinity of each individual molecule because the existence of the other molecule must be taken into account. For the sake of simplicity, we excluded such entries. (iv) The ligand molecule in the complex must contain only common organic elements, i.e., C, N, O, P, S, F, Cl, Br, I, and H. Although some ligand molecules contain elements such as Be, B, Si or metals, a practical concern is that the parameters for these elements are not always available in molecular modeling software. For the same reason, if a protein molecule contains any nonstandard amino acid residue as part of its binding pocket, the complex was also rejected. (v) The resolution of the given complex structure must be equal to or better than 2.5 Å. Since only 15 NMR structures were found among the 1359 PLEXBAS, they were not included in the refined list.

Our final refined list of PLEXBAS contains 800 entries. Each complex structure in this list has been processed properly and saved in a uniform format so that it can be readily utilized by molecular modeling software. This task was done with the Sybyl software.[21] Briefly, each complex structure was split into a protein molecule saved in the PDB format and a ligand molecule saved in the Tripos Mol2 format. Water molecules and other inorganic components were saved with the protein. Atom types and bond types of each ligand molecule were automatically assigned by the Sybyl software, followed by visual inspection and correction where necessary. No structural optimization or any type of transformation of the coordinates was made; the stored coordinates of the complex are exactly the same as those in the original PDB file. Hydrogen atoms were added to the protein and the ligand molecule by the Sybyl software with simple geometrical criteria. Positions of “rotatable” hydrogen atoms, such as the one on a hydroxyl group, were not optimized.

In summary, out of 19 621 experimental structures in the PDB Release No. 103 (January 2003), 5671 protein−ligand complexes were identified that met our selection criteria. Experimentally determined binding affinity data of 1359 complexes were collected from the associated references. After elimination of the complexes that are not suitable for the purpose of scoring/docking studies, our final refined list contains 800 PLEXBAS, all crystal structures with resolution equal to or better than 2.5 Å. Over 200 different types of proteins are found in this refined list. Their binding constants (K_d or K_i) range from 0.25 M to 11 fM, spanning more than 13 orders of magnitude (Figure 2). This set of PLEXBAS is several times larger than any previous compilation of this kind (Table 1), and because we have systematically collected and compiled this data set with stringent criteria and obtained the binding affinity data from original references, we expect this set of PLEXBAS to be of a high quality. For example, we previously used a set of 230 PLEXBAS for the development and validation of our X-Score scoring function.[10] Of these 230 entries, we found only 156 (about $^2/_3$) that qualify for entry into the final refined list of 800 PLEXBAS.

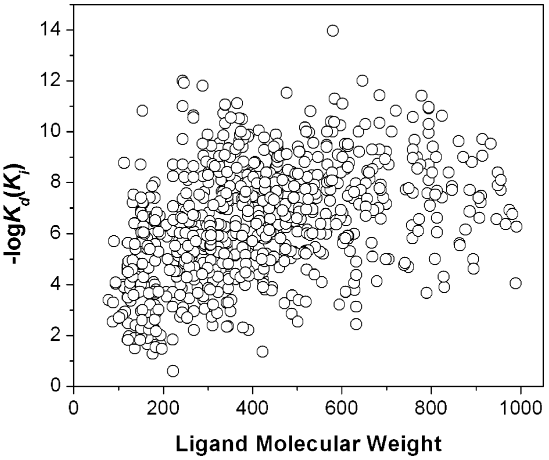

Figure 2. Distribution of the binding constants of the 800 protein−ligand complexes in the refined set of the PDBbind database.

We have organized the major outcomes of this project, including classification tables, binding affinity data, reference citations, and structural files, into a Web-accessible database named the PDBbind database. In its current form, it allows the users to browse and search the contents using a number of SQL queries based on textual and numerical criteria, such as binding affinity range, molecular weight of a ligand, and classification of a protein. In the next step we will enable non-SQL-based queries such as chemical structure and pharmacophore searching. Furthermore, external researchers are encouraged to deposit known binding affinity data of protein−ligand complexes through on-line preformatted forms, which will serve as a supplementary method to enrich the contents of our database. Our PDBbind database can be readily updated and expanded to better serve the scientific community and our plans are to (i) update our database annually to keep up with the rapid growth of the PDB, (ii) screen the other references listed in the PDB file if the binding affinity data of a complex of interest cannot be found in the primary reference, (iii) add key information of the binding assays if available in the original references, (iv) expand the scope of the database, which is currently limited to protein−ligand complexes, by including protein−protein and protein−nucleic acid complexes. The goal is to make the PDBbind database a valuable information resource for a larger research community.

Several existing databases are also dedicated to the study of protein−ligand binding. Our PDBbind database complements them and has a number of appealing aspects. Relibase+,[22] for example, collects protein−ligand complexes deposited in the PDB and provides various tools for analyzing these structures but does not provide any binding affinity information for these complexes. The Binding Database (BindingDB)[23] aims at collecting binding affinity data for a wide range of biological and chemically synthesized complex systems. But presently it has information only for a limited number of entries and not every entry in the BindingDB has available three-dimensional structural information. Another two existing databases are the Ligand−Protein Database (LPDB)[24] and the Protein−Ligand Database (PLD).[25] These two databases have a very similar theme to our PDBbind database: they both emphasize the link between the binding affinities and the structures of the protein−ligand complexes in PDB and have provided convenient Web-based tools for data retrieval and analysis. However, since the binding affinity data in these two databases were largely obtained from previously known compilations, neither LPDB (∼220 entries) nor PLD (∼270 entries) has yet to substantially increase the collection of known PLEXBAS.

Our current work represents the first accomplished attempt to collect experimental binding affinity data of protein−ligand complexes on the entire PDB level. Because of the theoretical significance of protein−ligand interactions, there are other ongoing efforts in this area by several other research groups. For example, Carlson's group at the University of Michigan is currently compiling a comprehensive database of protein−ligand complexes based on PDB.[26] We expect that these collective efforts will lead to the creation of valuable databases and data mining tools for the studies of protein−ligand interactions in this structural genomics era.

The PDBbind database described in this paper is public-accessible at http://www.pdbbind.org/.

Acknowledgment. This work was supported in part by a seed grant from the Office of the Vice President of Research of the University of Michigan. The authors are grateful to Dr. Marc C. Nicklaus at the National Cancer Institute at NIH and to Dr. Heather A. Carlson and Mr. Michael Lerner at the University of Michigan for their valuable discussions and helpful input on this project. The authors also thank Dr. George W. A. Milne for his critical reading of the manuscript and his editorial assistance.

References

(1) Böhm, H. J.; Stahl, M. The use of scoring functions in drug discovery applications. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH Inc.: New York, 2002; pp 41−88.

(2) Böhm, H. J. The development of a simple empirical scoring function to estimate the binding constant for a protein−ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243−256.

(3) Jain, A. N. Scoring non-covalent protein−ligand interactions: A continuous differentiable function tuned to compute binding affinities. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 427−440.

(4) Head, R. D.; Smythe, M. L.; Oprea, T. I.; Waller, C. L.; Green, S. M., Marshall, G. R. VALIDATE: A new method for the receptor-based prediction of binding affinities of novel ligands. *J. Am. Chem. Soc.* **1996**, *118*, 3959−3969.

(5) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425−445.

(6) Murray, C. W.; Auton, T. R.; Eldridge, M. D. Empirical scoring functions. II. The testing of an empirical scoring function for the prediction of ligand−receptor binding affinities and the use of Bayesian regression to improve the quality of the model. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 503−519.

(7) Böhm, H. J. Prediction of binding constants of protein ligands: A fast method for the prioritization of hits obtained from de novo design or 3D database search programs, *J. Comput.-Aided Mol. Des.* **1998**, *12*, 309−323.

(8) Wang, R.; Gao, Y.; Lai, L. SCORE: A new empirical method for estimating the binding affinity of a protein−ligand complex. *J. Mol. Model.* **1998**, *4*, 379−394.

(9) Cozzini, P.; Fornabaio, M.; Marabotti, A.; Abraham, D. J.; Kellogg, G. E.; Mozzarelli, A. Simple, Intuitive Calculations of Free Energy of Binding for Protein−Ligand Complexes. 1. Models without Explicit Constrained Water. *J. Med. Chem.* **2002**, *45*, 2469−2483.

(10) Wang, R.; Lai, L.; Wang, S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 11−26.

(11) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein−ligand interactions: A simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791−804.

(12) Muegge, I. A knowledge-based scoring function for protein−ligand interactions: Probing the reference state. *Perspect. Drug Discovery Des.* **2000**, *20*, 99−114.

(13) Muegge, I. Effect of ligand volume correction on PMF scoring. *J. Comput. Chem.* **2001**, *22*, 418−425.

(14) Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Thornton, J. M. BLEEP: potential of mean force describing protein−ligand interactions. I. Generating potential. *J. Comput. Chem.* **1999**, *20*, 1165−1176.

(15) Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Forster, M. J.; Thornton, J. M. BLEEP: potential of mean force describing protein−ligand interactions. II. Calculation of binding energies and comparison with experimental data. *J. Comput. Chem.* **1999**, *20*, 1177−1185.

(16) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein−ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.

(17) Gohlke, H.; Hendlich, M.; Klebe, G. Predicting binding modes, binding affinities and "hot spots" for protein−ligand complexes using a knowledge-based scoring function. *Perspect. Drug Discovery Des.* **2000**, *20*, 115−144.

(18) Ishchenko, A. V.; Shakhnovich, E. I. Small molecule growth 2001 (SMoG2001): An improved knowledge-based scoring function for protein−ligand interactions. *J. Med. Chem.* **2002**, *45***,** 2770−2780.

(19) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, I. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242 (http://www.rcsb.org/pdb/).

(20) Kenakin, T. *Pharmacologic Analysis of Drug−Receptor Interaction*, 3rd ed.; Lippincott-Raven Press: New York, 1997.

(21) *SYBYL* (software), version 6.8; Tripos Inc.: St. Louis, MO; http://www.tripos.com/.

(22) Hendlich, M.; Bergner, A.; Gunther, J.; Klebe, G. Relibase: Design and development of a database for comprehensive analysis of protein−ligand interactions. *J. Mol. Biol.* **2003**, *326*, 607−620 (http://relibase.ccdc.cam.ac.uk/).

(23) Chen, X.; Lin, Y.; Liu, M.; Gilson, M. K. The Binding Database: Data Management and Interface Design. *Bioinformatics* **2002**, *18*, 130−139 (http://www.bindingdb.org/).

(24) Roche, O.; Kiyama, R.; Brooks, C. L., III. Ligand−Protein DataBase: Linking Protein−Ligand Complex Structures to Binding Data. *J. Med. Chem.* **2001**, *44*, 3592−3598 (http://lpdb.scripps.edu/).

(25) Puvanendrampillai, D.; Mitchell, J. B. O. Protein Ligand Database (PLD): additional understanding of the nature and specificity of protein−ligand complexes. *Bioinformatics* **2003**, *19*, 1856−1857 (http://www-mitchell.ch.cam.ac.uk/pld/index.html).

(26) Personal communications with Dr. Heather A. Carlson at the College of Pharmacy, the University of Michigan.

JM030580L