The PDBbind Database: Methodologies and Updates

Renxiao Wang, Xueliang Fang, Yipin Lu, Chao-Yie Yang, 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 December 23, 2004

We have developed the PDBbind database to provide a comprehensive collection of binding affinities for the protein−ligand complexes in the Protein Data Bank (PDB). This paper gives a full description of the latest version, i.e., version 2003, which is an update to our recently reported work. Out of 23 790 entries in the PDB release No.107 (January 2004), 5897 entries were identified as protein−ligand complexes that meet our definition. Experimentally determined binding affinities (K_d, K_i, and IC_{50}) for 1622 of these were retrieved from the references associated with these complexes. A total of 900 complexes were selected to form a "refined set", which is of particular value as a standard data set for docking and scoring studies. All of the final data, including binding affinity data, reference citations, and processed structural files, have been incorporated into the PDBbind database accessible on-line at http://www.pdbbind.org/.

Introduction

Mutual recognition and association of receptor and ligand molecules play a crucial role in virtually all fundamental biological processes. For example, binding of a substrate to an enzyme initiates a catalytic reaction, binding of a hormone molecule to its receptor triggers a metabolic event, and binding of a repressor or an activator to DNA regulates gene transcription. Development of molecules that can bind specifically to a targeted molecule is a basic strategy for modulating a biological process on the molecular level, and it is central to the discipline known as "chemical genomics".

An example of this kind is structure-based drug design,[1−3] whose success depends on a firm grasp of the principles of molecular recognition. Utilizing the available structural information of the targeted molecule, structure-based methods may lead to efficient discovery of lead compounds that are sterically and chemically complementary to the targeted molecule and thus accelerate the birth of new drugs. Structure-based drug design has been widely accepted as an established technology for drug discovery research in academia and the pharmaceutical industry. Several new drugs, whose invention was largely facilitated by structure-based drug design methods, have already reached the market.[4]

A central problem in structure-based drug design concerns the quantitative prediction of the binding affinities of ligand molecules to their biological target, which is often referred to as the "scoring problem". In recent years, a group of approaches called "scoring functions" have emerged as practical solutions to the scoring problem.[5] The best approach to the development and validation of these methods [6−22] requires the knowledge of protein−ligand complexes with both high-resolution three-dimensional structures and reliably determined binding affinities. Modern crystallographic and multidimensional NMR techniques have provided a wealth of structural information of various biological molecules. As this paper was prepared, nearly 28 000 structures have already been deposited into the Protein Data Bank (PDB);[23] a substantial proportion of these are complexes formed between various types of protein molecules and organic ligand molecules. Structural information, however, does not automatically reveal the energetic characteristics in protein−ligand binding, which is also essential to a full understanding of the protein−ligand binding process. Several publicly accessible databases currently exist, such as KiBank [24] and the PDSP K_i Database,[25] which provide comprehensive collections of experimentally determined binding affinity data. A weakness of these databases is that they do not supply the three-dimensional structural information for the entries in the databases.

We see the gap between structural and energetic information as one major obstacle to the development of finer theoretical and computational models for protein−ligand binding affinity prediction. Table 1 summarizes the collections of PLEXBAS (*p*rotein−*l*igand comp*lex* with both known *b*inding *a*ffinity and *s*tructure) that have been reported since the 1990s; most of these have been used to develop and validate scoring functions. Combining the collections listed in Table 1 and eliminating duplicates gives a total of about 300 PLEXBAS. Neither the size of the dataset nor the speed of its growth is impressive. Furthermore, quality problems arise since the collections listed in Table 1 were typically compiled from secondary sources with little rigor. For example, we used a set of 230 PLEXBAS before to develop and validate the X-Score scoring function.[22] This set was largely a combination of other researchers' previously published compilations.[9−19] After careful examinations, we now realize that a good portion of this set has various problems that thus may make it unsuitable for scoring function development or validation. For example, some entries are covalently bound complexes,

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

10.1021/jm048957q CCC: $30.25
Published on Web 05/24/2005

Table 1. Compilations of Protein−Ligand Complexes with Known Binding Affinities Prior to the Creation of the PDBbind Database

author (approach[a])	year published	no. of PLEXBAS cited	refs
Böhm (Score1)	1994	54	6
Jain	1996	34	7
Head et al. (VALIDATE)	1996	65	8
Eldridge et al. (ChemScore)	1997	112	9, 10
Böhm (Score2)	1998	94	11
Wang et al. (SCORE)	1998	181	12
Muegge et al. (PMF)	1999	225	13−15
Mitchell et al. (BLEEP)	1999	90	16, 17
Gohlke et al. (DrugScore)	2000	83	18, 19
Roche et al. (LPDB)	2001	∼200	26
Chen et al. (BindingDB)	2001	[b]	27−29
Ishchenko et al. (SMoG2001)	2002	119	20
Cozzini et al. (HINT)	2002	53	21
Wang et al. (X-Score)	2002	230	22
Puvanendrampillai et al. (PLD)	2003	∼300	30

[a] LPDB, Binding DB, and PLD are binding affinity databases; all of the other approaches listed in this table are scoring functions, which all utilized a number of PLEXBAS as training set or test set. [b] Binding DB consists of PLEXBAS as well as other entities without structural information. The number of PLEXBAS in BindingDB is unknown to us.

some are ternary complexes, some have low-resolution structures, while some entries have IC_{50} values rather than K_i or K_d values as binding affinities. In fact, only 95 entries in this set remain in the "refined set" of our PDBbind database (which will be described later in this paper). It is reasonable to expect that more accurate scoring functions can be developed if a more comprehensive and high-quality set of PLEXBAS is available.

Recognizing the importance of a comprehensive collection of protein−ligand binding affinities, we have launched a project to retrieve from the open literature experimentally determined binding affinities for all of the protein−ligand complexes in the PDB. The preliminary results of this project were reported recently.[31] In our previous work, PDB release No. 103 was screened and 5671 protein−ligand complexes out of 19 621 experimental structures were identified. A systematic examination of over 3000 publications associated with these entries led to a collection of binding affinity data (K_d, K_i, and IC_{50}) for a total of 1359 complexes. These results have been organized into a Web-accessible database, named PDBbind, which was released to the public in May 2004 and is now used by more than 500 groups around the world.

Compared to all the previous collections listed in Table 1, our database offers a significantly larger collection of PLEXBAS. The binding affinity data deposited in the database are also of higher quality, since they were all extracted from original references through a systematic effort. Importantly, we have demonstrated that binding affinity information for the protein−ligand complexes in PDB can be collected efficiently from literature using our strategy, i.e., examining the references associated with these complex structures. With this strategy, a regular update based upon the growing contents of PDB will ensure a steady growth of our database in the future.

Since the publication of our preliminary results, we have refined the methods used for the identification and classification of protein−ligand complexes in the PDB and have updated our database by using a new release of PDB. The results have been incorporated into the PDBbind database (version 2003), which is described in this paper.

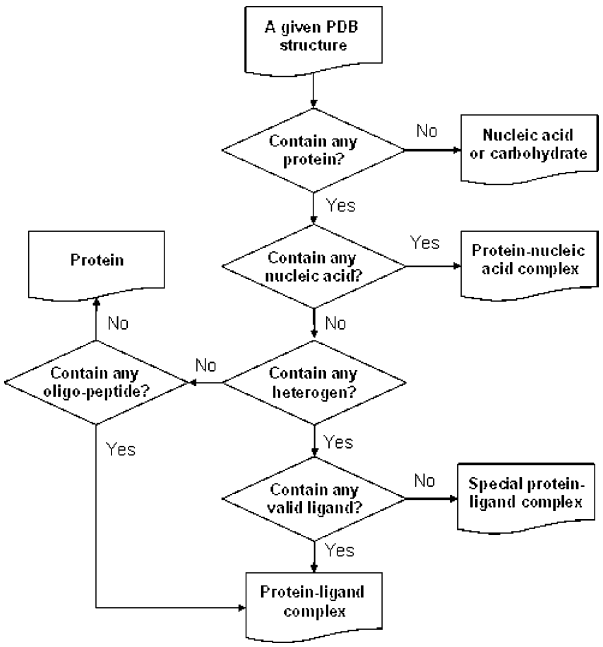

Figure 1. The scheme used for the identification and classification of protein−ligand complexes.

Methods and Results

The work described in this paper was based on PDB release No. 107 (January 2004). This release contains all of the entries made public by PDB before January 1, 2004, a total of 23 790 experimentally determined structures. A few hundred theoretical structural models were also included in this release, but they were not considered in our work.

Identification and Classification of Protein−Ligand Complexes. The first step was to identify the protein−ligand complexes in PDB, since PDB itself does not provide such a classification. The focus of our current work centers on the complexes formed between protein molecules and nonpolymer organic molecules. Complexes formed between biological macromolecules, such as protein−protein and protein−nucleic acids complexes, were not considered in this work.

As an improvement to the methods used in our previous work,[31] we have developed a new scheme to analyze and classify any given structure in the classical PDB format. The new classification scheme was programmed in C++ according to the flowchart shown in Figure 1 and proceeds by the following steps:

(1) Check if the given structure contains any polypeptide chain. Such a judgment can be made by analyzing the "SEQRES" and "ATOM" records in the given PDB file. If not, this structure must be either a nucleic acid molecule or a carbohydrate molecule.

(2) Check if the given structure contains any polynucleotide chain (DNA or RNA). If so, the given structure can be classified as a protein−nucleic acid complex.

(3) Check if this structure contains any eligible heterogen molecule. According to the definition of the PDB format, any species other than the standard components of protein or nucleic acid molecules are considered as "heterogens". These heterogens, including all types of organic and inorganic molecules, are typically specified in the "HET", "HETNAM", and "HETATM" records in a PDB-format file. In fact, very few structures in PDB do not contain any heterogen molecule. Not all of the heterogen molecules, however, may qualify as valid ligand molecules. First, inorganic heterogens, including water

Table 2. "Special" Heterogen Molecules Defined in Our Work[a,b]

Coenzyme A Family[c]							
COA (69)	2CP (2)	3CP (1)	3HC (1)	4CA (2)	4CO (2)	ACO (25)	AMX (1)
BCA (3)	CA3 (1)	CA5 (1)	CAA (17)	CAO (2)	CIC (1)	CMC (3)	CMX (1)
CO8 (13)	COD (1)	COF (2)	COS (4)	COT (6)	CS8 (4)	DAK (6)	DCA (5)
DCC (1)	FAM (1)	FCX (1)	HAX (1)	HMG (5)	HXC (7)	LYX (2)	MCA (14)
MCD (2)	MDE (1)	MLC (2)	MYA (4)	NHM (2)	NMX (1)	SCA (3)	SCD (2)
SCO (1)							
Heme (Metal-Containing Protoporphyrin) Family[d]							
HEM (1138)	1CP (3)	B12 (31)	BCB (28)	BCL (224)	BLA (4)	BLV (4)	BPB (16)
BPH (74)	CCH (1)	CL1 (96)	CL2 (1)	CLA (375)	CLN (1)	CNC (18)	COB (4)
COH (11)	CON (1)	COJ (1)	COY (4)	CP3 (3)	DDH (3)	DEU (1)	DHE (42)
F43 (14)	FEC (24)	HAS (1)	HDD (12)	HDM (1)	HE6 (1)	HEA (40)	HEB (1)
HEC (389)	HEG (4)	HEO (2)	HES (1)	HEV (3)	HIF (1)	HNI (22)	MHM (1)
MMP (6)	MP1 (1)	PC3 (1)	PCU (1)	PNI (2)	POR (1)	PP9 (1)	SRM (12)
ZEM (1)	ZNH (4)						
NAD (Nicotinamide Adenine Dinucleotide) Family[e]							
NAD (300)	ADJ (2)	CAN (1)	CND (2)	DND (18)	NAC (3)	NAE (2)	NAH (15)
NAI (60)	NAJ (10)	NAP (271)	NAQ (2)	NAX (2)	NBD (4)	NBP (1)	NDA (1)
NDC (1)	NDO (1)	NDP (221)	NHD (1)	NHO (1)	ODP (2)	PAD (3)	SND (5)
TAP (4)	ZID (1)						
FAD (Flavin Adenine Dinucleotide) Family[f]							
FAD (319)	6FA (4)	FAA (2)	FAB (8)	FAE (1)	FAS (1)	FDA (1)	FMA (4)
FMN (295)	FNS (2)	MGD (44)	RFL (1)				
Nucleotides[g]							
AMP (93)	ADP (289)	ANP (92)	ATP (160)	CMP (20)	CDP (6)	CTP (19)	GMP (6)
GDP (125)	GNP(74)	GTP (135)	2GP (71)	TMP (56)	TDP (25)	TTP (49)	UMP (125)
UDP (70)	UTP(7)	PSU (95)					
"Junk" Molecules[h]							
BMA (40)	BOG (66)	C8E (88)	CIT (105)	CRY (44)	DTT (32)	EPE (71)	F6P (33)
FUC (176)	GAL (184)	GLC (354)	GOL (690)	HED (99)	LDA (35)	LI1 (141)	MAL (44)
MAN (462)	MES (113)	MPD (167)	MYR (85)	NAG (1098)	NGA (32)	PEG (24)	PG4 (31)
POP (45)	PYR (26)	SPM (48)	TRS (114)	XYS (82)			

[a] Each molecule in this table is represented by its three-letter abbreviation as found in PDB files. To view the full name and the chemical structure of each molecule, we recommend the reader check the summary of hetero groups in the PDBsum database at http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/index.html. [b] The number in brackets is the total number of PDB entries that contain this particular molecule. [c] The keyword "coenzyme" was used to identify the members in this family. [d] The keyword "porphyrin" was used to identify the members in this family. [e] The keyword "nicotinamide" was used to identify the members in this family. [f] The keyword "flavin" was used to identify the members in this family. [g] The keywords "adenosine", "thymidine", "cytidine", "guanosine", and "uridine" were used to identify the members in this family. [h] These molecules were selected among the heterogen molecules with an occurrence over 20 in PDB.

molecules, metal ions, and other inorganic components included in the given structure, were not considered. Second, organic solvent molecules were not considered eligible. A list of organic solvent molecules commonly seen in PDB structures, such as methanol, ethanol, acetone, and DMSO, was implemented in our program to identify such molecules. Third, we also applied a simple rule that any molecule with fewer than six non-hydrogen atoms was not considered eligible. If the given structure does contain at least one eligible heterogen molecule, the procedure then goes to step 5; otherwise, it goes to step 4 below.

(4) Check if the given structure contains any oligopeptide besides the main protein molecule by analyzing the "SEQRES", "ATOM", and "HETATM" records in the given PDB file. Oligopeptides and their mimics are treated as valid ligand molecules in our work, since they represent a very important class of molecules in drug discovery. Unlike full-length peptides, oligopeptides normally do not form stable secondary structures by themselves and therefore may be considered as common organic molecules. In our work, peptides containing 10 or fewer amino acid residues were defined as oligopeptides. If the given structure contains at least one oligopeptide, it is classified as a protein−ligand complex.

(5) Check if any of the heterogen molecules in the given structure can be considered as a valid ligand molecule. Some particular heterogen molecules were considered as "special" and were distinguished from the others in our work. We have compiled a dictionary of such "special" molecules, and our program uses it as a reference when examining the given structure. If the given structure is found to contain a valid ligand molecule, it will be classified as a protein−ligand complex; otherwise, it will be classified as a complex formed with a "special" ligand molecule.

Here we need to explain in more detail what molecules are defined as "special" in our work. All of such molecules are listed in Table 2, which can be grouped into two categories. The first consists of several well-characterized biological cofactors/coenzymes, including coenzyme A (CoA), heme, nicotinamide adenine dinucleotide (NAD), flavin adenine dinucleotide (FAD), and nucleotides (A, T, C, G, and U). These molecules all have important biological functions and are often an indispensable part of the complex structures. However, they represent some very unique chemical structures and would better be distinguished from "normal" organic ligand molecules. Note that our dictionary of "special" ligand molecules contains not only the several particular biological cofactors/coenzymes mentioned above but also a number of heterogen molecules observed in the PDB which, being direct derivatives or analogues of them, should therefore be grouped with them. Such direct derivatives and analogues of these biological cofactors/coenzymes were identified through a keyword-based search among all of the heterogen molecules summarized in our survey. For example, we stipulated that any molecule with a porphyrin ring system in its chemical structure was an analogue of heme. Accordingly, every heterogen molecule with the word "porphyrin" in its name was extracted and then examined to confirm that it indeed contained such a chemical moiety. If so, it was then added to the heme family. The coenzyme A, NAD, FAD, and nucleotide families in our dictionary were all constructed through a similar approach (see Table 2).

The second category of special molecules defined in our work consists of some miscellaneous organic molecules, which are

Table 3. Classification of PDB Release No. 107 (23 790 entries in total)

category	total no.	% of the entire PDB
nucleic acids	1276	5.4
carbohydrates	18	<0.1
protein−nucleic acid complexes	982	4.1
apo-proteins and protein−protein complexes	10221	43.0
protein−ligand complexes formed with junk molecules	1955	8.2
protein−ligand complexes formed with biological cofactors/coenzymes	3441	14.5
protein−ligand complexes formed with organic molecules	5897	24.8

referred to as "junk" molecules in this paper. Such molecules are observed in a large number of PDB entries as heterogens, but they cannot be considered as meaningful, deliberately introduced ligand molecules. For example, a number of saccharide molecules, such as *N*-acetyl-D-glucosamine (NAG) and α-D-mannose (MAN), are standard components seen on many glycoprotein molecules. Some other molecules, such as 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (Hepes, EPE in Table 2), are components of buffers or other additive reagents that are frequently used in crystal growth. To identify such junk molecules, we have conducted a survey on all of the heterogen molecules observed in the entire PDB. Information of each heterogen, including name and formula, was extracted from the "HET", "HETNAM", and "HETATM" records in the PDB files using computer programs. All such information was summarized, and the occurrence of each unique heterogen molecule was determined. The guideline we used is that a meaningful ligand molecule should be specifically identified, and if a certain heterogen molecule is observed in a considerable number of structures, its role as a meaningful ligand molecule is dubious. We thus examined all of the heterogen molecules that occurred in ≥20 PDB entries in our survey, and the junk molecules were identified out of these candidates.

One more thing needs to be clarified here. Usually, each heterogen specified in a PDB file stands for an individual molecule. But in some cases, several heterogens are connected with each other via covalent bonds, forming one single molecule. A heterogen in such cases is actually a fragment in a molecule. Our program is able to detect the connections between heterogens by interpreting the "SEQRES", "LINK", and "CONECT" records in a PDB file and constructs the complete ligand molecule accordingly. If a ligand molecule was detected to be composed of multiple heterogens, our program judged it by two rules: (1) if any of the heterogen components of this molecule were a biological cofactor/coenzyme as defined in Table 2, it was classified as a biological cofactor/coenzyme molecule, or (2) if all of the heterogen components of this molecule were junk component as defined in Table 2, it was classified as a junk molecule. Otherwise it was classified as a valid ligand molecule.

Using the classification scheme described above, the 23 790 experimentally determined structures in PDB release No. 107 were grouped into several categories (see Table 3). The protein−ligand complexes formed with normal organic ligand molecules, 5897 in total, were considered in the subsequent steps in our work. Note that they included both noncovalently and covalently bound complexes.

Collection of Binding Affinity Data from Literature. The second step of our work was to collect binding affinity data of protein−ligand complexes from the literature. We employed the strategy used in our previous work:[31] the "primary reference" listed in the PDB file of each complex was reviewed to retrieve the experimentally measured binding affinity of this complex. The primary reference of a PDB entry is normally provided to give a full description of the solution of the structure. However, if the PDB structure is a protein−ligand complex, and if its authors have also measured the binding affinity of this complex, it is possible that they will also include this information in the same paper; alternatively, if the binding affinity of this complex has been measured previously, such information may be cited in this paper. Thus, in either case, if the binding affinity of a complex has indeed been measured, it is likely that such information can be retrieved, directly or otherwise, from the "primary reference" listed in the PDB file of that complex.

The "primary reference" of each complex was retrieved from the "JRNL" record in its PDB record. Some PDB files released in the early years do not have a JRNL record, and in such cases, the first reference cited in the "REMARK" record was retrieved as the "primary reference". All of the references were either downloaded on-line or requested from the University of Michigan library system, which yielded a total of over 4000 published papers. For convenience of reviewing and archiving, all of these references were saved electronically in PDF format. The references for 644 complexes could not be obtained because they were stated in the corresponding PDB files as "to be published" or were published in some journals that were unavailable in our library system.

All of the obtained references have been manually reviewed. The experimentally measured binding affinity that matched the complex structure presented in the corresponding PDB file was recorded if it was reported or cited in the given reference. We accepted three major forms of binding affinity data: dissociation constant (K_d), inhibition constant (K_i), and concentration at 50% inhibition (IC_{50}). Kinetic parameters for enzymatic reactions, such as K_m and k_{cat}, were not accepted. In a very small number of cases, binding affinity data of a given complex were available in multiple forms. In such cases we applied an order of preference ($K_d > K_i > IC_{50}$) and recorded only the preferred form. Similarly, if the binding affinity data of a given complex were measured under different environmental settings, we recorded only the result measured at room temperature and neutral pH or the one measured under a condition closest to this setting.

Combining the results obtained in this work and our previous work,[31] we have collected binding affinity data for 1622 protein−ligand complexes (468 K_d values, 840 K_i values, and 314 IC_{50} values) among the 5897 legitimate protein−ligand complexes identified in the previous step.

Selection of the Refined Set. We anticipated that our PDBbind database will be of major interest to researchers working on structure-based drug design, particularly in the docking/scoring field. Not every 1622 PLEXBAS identified through the previous steps, however, is the right material for docking/scoring studies. Consequently, we have applied a number of criteria to filter out the unqualified entries from them, and the remaining entries formed a "refined set". The following criteria were applied to the selection of the refined set:

(1) Only crystal structures with an overall resolution better than or equal to 2.5 Å were accepted into the refined set. This cutoff of 2.5 Å was chosen to select the structures with relatively high resolution. Structures resolved by NMR techniques were neglected in our work, since they represent a minor population: there were only 39 NMR structures among the 1622 candidates. However, as the number of NMR structures grows, they will be included in our future updated version.

(2) Only noncovalently bound complexes were accepted into the refined set. Our program detected any covalent connection between the protein and the ligand by analyzing the "LINK" and "CONECT" records in the given PDB file. It further checked the distance between the protein and the ligand using the available coordinate information. If the distance between any two non-hydrogen atoms on the protein and the ligand was found to be shorter than 2.0 Å, this complex was rejected, even if it was not identified as covalently bound.

(3) Only binary complexes were accepted into the refined set. In other words, the complex must be a structure formed unambiguously between one protein molecule and one ligand molecule. In some cases, two ligand molecules bind together inside the same binding site on the protein molecule to form

a ternary complex. In such cases, experimental measurement of the binding affinity of each individual ligand molecule becomes much more complicated, since the existence of the other ligand molecule must be taken into account, and none of the known docking/scoring tools is really capable of handling ternary protein−ligand complexes. Thus, we decided not to include them in the refined set. If there were more than one ligand molecule in the given complex structure, our program would check the distance between them. If any non-hydrogen atom on one ligand molecule was within 8.0Å of any non-hydrogen atom on another ligand molecule, the given complex was considered to be a ternary complex and, accordingly, was rejected.

(4) Only complexes with known K_d or K_i values were accepted into the refined set. K_d and K_i are equilibrium constants, which are in principle associated with the thermodynamic properties of a protein−ligand binding process. They are also independent of the concentration levels of the protein and the ligand used in binding assay and thus can be considered as a measurement of absolute binding affinities. In contrast, IC_{50} values depend on the design of the binding assay and reflect only the relative binding affinities of the inhibitor molecules evaluated in the same binding assay. Therefore, we decided not to mix IC_{50} values with K_d and K_i values.

(5) Some specific restrictions on the ligand and the protein were also applied. To be included in the refined set, the ligand molecule in the given complex must consist of only common organic elements, i.e., C, N, O, P, S, F, Cl, Br, I, and H. Otherwise, the complex was rejected. This is a practical concern, since not all of today's molecular modeling software has the necessary parameters to handle organic molecules with uncommon elements, such as Be, B, Si, or metals. Similarly, if the protein molecule in the given complex had nonstandard amino acid residues close (i.e. within 8.0 Å) to where the ligand molecule was bound, the complex was rejected. In addition, to control the size of the ligand molecules, the molecular weight of the ligand molecule was restricted to 1000 or less.

With the above criteria, 900 protein−ligand complexes were selected for inclusion into the refined set. Their binding affinities (K_d or K_i) range from 0.25 M to 11 fM, spanning more than 13 orders of magnitude. Over 200 different types of proteins, judged by their names and enzyme classification (EC) numbers, are included in this set.

The structural file of each complex in the refined set was processed into appropriate formats. Each complex structure was split into a protein molecule and a ligand molecule. The protein molecule was saved in PDB format and the ligand molecule in Tripos Mol2 format. Other auxiliary components in the original PDB structural file, such as water molecules and metal ions, were saved together with the protein molecule. Atom types and bond types on each ligand molecule were first assigned by using the Sybyl software on the basis of the information available in the original PDB file and then were visually inspected. Any incorrect atom types and bond types that were found were corrected. Hydrogen atoms were added to both the protein and the ligand molecule with standard bond lengths and bond angles using Sybyl software. A special program was developed to place the hydrogen atoms on hydroxyl and thiol groups, which could be placed at different positions along a circular path, at a position where they could form maximal contacts with the nearby hydrogen-bond-acceptor atoms on both the protein and the ligand molecule. For convenience, the "standard" protonation states at neutral pH were applied to the amino acid residues on the protein side. Thus, Asp, Glu, and His residues were negatively charged, and Arg and Lys residues were positively charged; carboxylic, sulfonic, and phosphoric acid groups in the ligands were built in deprotonated forms, while aliphatic amine groups, guanidine, and amidine groups were built in protonated forms. Neither the protein nor the ligand was subjected to any structural optimization. Therefore, all of the non-hydrogen atoms on both the protein and the ligand molecule retained their original coordinates from PDB.

In the process described above, the "biological unit" of each complex structure was used instead of the structure given in the standard PDB file, which typically only consists of one asymmetric unit of the crystal cell. Technically, a biological unit is not necessarily equivalent to an asymmetric unit: it may be composed of multiple copies or part of the asymmetric unit. Using the biological unit of each complex ensures that the structure is functionally complete and nonredundant. The structure of the biological unit of each complex was downloaded from the RCSB FTP server at ftp://ftp.rcsb.org/pub/pdb/. If the biological unit was bound with more than one copy of the same ligand molecule at different positions, only the first copy appearing in the given file was kept, while all the others were removed.

Design of the Web Interface. We have organized the major results of our project into a Web-based database called PDBbind, which is accessible at http://www.pdbbind.org/. The database was constructed using the Oracle relational database software (v 9.20). The utilities on the front end of the database were programmed primarily in Java. All of the programs were integrated and powered by Apache on a platform running Red Hat Linux (v 9.0 professional). Apache Jakarta Tomcat (v5.0) was used to implement Java Servlet and Java Server Pages to provide the environment required by Java programs.

All of the 1622 protein−ligand complexes with known binding affinity are included in the database. For each of them, our web site provides a "fact sheet", summarizing its 4-letter PDB code, resolution of the structure, name and EC number of the protein molecule, and its binding affinity (K_d, K_i, or IC_{50}). This sheet provides a link to the corresponding page of this complex on the RCSB PDB web site, from which one can conveniently get information regarding the complex. The primary reference of this complex, in which the experimentally measured binding affinity was cited, is linked on this sheet to the corresponding web page in the NCBI PubMed on-line literature database. For each complex in the refined set, this sheet also provides a two-dimensional sketch of the chemical structure of the ligand molecule.

Search engines have been implemented on the front end of our database to allow the users to analyze its contents. One can use a range of text-based SQL queries to search for and retrieve entries matching the given queries. Currently accepted queries include the PDB code, the release year of the structure, the protein name, EC number, and binding affinity range. Another web page has been designed to perform non-SQL structure searches among the ligand molecules in the refined set. The popular JChem Java applet[32] was used to construct the interface. With this tool, one can draw a chemical moiety as query and then retrieve all of the ligand molecules containing the input query or a structure similar to the input query. The hits are summarized and returned on a separate page. This tool helps one to investigate the role of a particular chemical moiety in protein−ligand binding in a very convenient way.

Discussion

Our PDBbind database is a secondary database based on the contents of PDB. In constructing this database, we attempted to establish a PDB-wide link between the structural information of protein−ligand complexes and their energetic information. Several databases of protein−ligand complexes already exist. For example, Relibase+ database[33] and the PDBLIG database[34] also offer PDB-wide classifications of protein−ligand complexes. These two databases, however, do not provide any binding affinity information for the protein−ligand complexes in their current contents. The Ligand−Protein DataBase (LPDB)[26] and the Protein−Ligand Database (PLD)[30] provide collections of binding affinity data for protein−ligand complexes in PDB. However, neither of these databases have collected binding affinity data from

original references and have not frequently updated their contents to keep up with the rapid growth of PDB.

To work on a database like PDB, which is already comprehensive and is still growing rapidly, the entire project needs to be carefully planned, and it is important to select the right strategy at every step. In this connection, several particular aspects in our work deserve further discussion.

The Definition of Protein−Ligand Complex. The first step in our work was to identify the protein−ligand complexes in PDB, this being a necessary prerequisite for all of the subsequent steps. This task, as was discovered, is by no means trivial. As mentioned in our previous work,[31] our initial attempt was to search PDB with keyword-based queries, in a hope that this approach would extract all of the protein−ligand complexes in PDB, albeit with some false hits. It was soon found however, that this approach was naïve, since it missed far too many bona fide protein−ligand complexes. To illustrate this, we have designed a computer program that reports a hit if it finds the spelling of "complex", "ligand", or "inhibit" in the descriptive records of a given PDB file. Any word containing the given query is considered as a hit. Thus, the word "complex" for example also retrieves "complexing", "complexation", "complexed", and so on. This program was applied to all of the PDB entries that do not contain any nucleic acid molecule in its contents (otherwise they are either nucleic acid molecules or protein−nucleic acid complexes and therefore rejected) and reported 10 952 hits out of a total number of 21 513 qualified entries. By comparing these hits with the results given by the more elaborate classification scheme adopted in our work, we found that 960 complexes formed with organic ligands and 1341 complexes formed with biological cofactors/coenzymes were not included among the hits given by this simple keyword-driven search. Apparently, these entries simply do not have a word such as complex, ligand, or inhibit in their PDB files. In addition, among the 10 952 hits given by this approach, a total of 3915 structures were not complexes formed with either organic ligands or biological cofactors/coenzymes, giving a rate of false hits of 3915/10 952 = 36%. If we were to have relied on the results of this keyword-driven search, the subsequent literature-reviewing step in our work would have become even more tedious because of these false hits. Using more complicated queries in the search may reduce false positives (hits), but it will inevitably lead to a greater number of false negatives. Our results suggest that a reliable classification of a given PDB entry must be based on a thorough analysis of its structure.

We use the term "protein−ligand complex" to describe a complex formed between a protein molecule and a nonpolymer organic molecule. Such a complex can be defined in a very general sense. As a matter of fact, nearly half of the PDB structures contain a protein molecule and at least one organic heterogen molecule (see Table 3). A notable feature in our work is that we have attempted to identify the "legitimate" protein−ligand complexes, which are relevant to drug design studies. As described in the methods section, our classification scheme differentiated the complexes formed with normal organic molecules, the complexes formed with specific biological cofactors/coenzymes, and the complexes formed with some irrelevant junk molecules. Only the complexes in the first category were considered in the subsequent steps in our work.

Exclusion of the second and the third category of complexes of course reduced the labor required at the subsequent steps, especially the literature-reviewing step, but the exclusion was also based upon technical grounds. First, these complexes typically involve some particular ligand molecules that are observed repeatedly in many PDB entries. For example, heme alone is observed in more than 1100 entries in PDB (see Table 2). Inclusion of these complexes in our database will, as can be readily perceived, bias its contents toward some particular types of molecules; this works counter to our goal of a database of protein−ligand complexes with a maximal diversity. Second, these complexes often have some undesired features. For example, some biological cofactors/coenzymes, such as NAD and CoA, have oxidized and reduced forms. Different forms of the same cofactor/coenzyme may exhibit binding affinities differing by several orders of magnitude. Technically, it is not always straightforward to identify the correct form of a biological cofactor/coenzyme only with the information available in the given PDB record. Besides, if a biological cofactor/coenzyme is not in a neutral form, it is also difficult to portray its chemical structure accurately with common atom types, such as the ones used by the Mol2 format. Third, biological cofactors or coenzymes often bind together with another molecule to form a ternary complex, and in some cases, they are connected to the protein molecule via covalent bonds. All of these considerations suggest that it is advisable to differentiate the complexes in the second and the third category from the other protein−ligand complexes.

The classification scheme we developed in this work was one major improvement to the one used in our previous work.[31] It was fully automated after some preliminaries, such as the derivation of the dictionary of "special" ligand molecules, and it processed the entire PDB in less than 1 h. The dictionary of "special" ligand molecules implemented in our program was more complete, since it was derived from a thorough survey of the heterogen molecules observed in PDB. More types of molecules, such as nucleotides, were classified as special ligand molecules in this work. With this change, the total number of legitimate protein−ligand complexes defined in this work was only marginally larger (5897 entries vs 5671 entries) than the one reported in our previous work, although this work was based on a later release of PDB (23 790 entries vs 19 621 entries).

We would like to mention that this classification scheme of protein−ligand complexes reflects our own concerns. The definition of protein−ligand complex could be very flexible. Some other studies, such as Relibase+[33] and PDBLIG,[34] provide different lists of protein−ligand complexes on the full PDB scale. As suggested by the data in Table 3, it should not be too surprising if between 25% and 50% of the entire PDB are classified by somebody as protein−ligand complexes.

Literature Reviewing. As described in the methods section, the binding affinity data in our database were all retrieved from the open literature. The literature

review step was definitely the most time-consuming step in our project, since each reference had to be reviewed manually, sometimes more than once. It would be helpful if a computer program could be deployed to read the papers and to retrieve the wanted information automatically, but this possibility does not seem to be practical. Computer programs have been used to interpret spoken and written languages, but we do not know of any program that can read and, more importantly, understand a scientific paper. At a very early stage of this project, we tested keyword-based searches on HTML-format and PDF-format references with queries such as "dissociation constant", "inhibition constant", and "IC50", with the hope of narrowing down the number of references that had to be reviewed manually. Unfortunately, similar to the keyword-driven search for protein−ligand complexes in PDB we have discussed above, this approach returned far too many false hits and also missed many real hits. Moreover, we found in our work that the most difficult aspect in reviewing a given reference is to find in a pool of reported data the binding affinity that matches the protein−ligand complex structure in the corresponding PDB file. We believe that human interpretation is still essential for this task.

One major advantage of our database over the previous collections of protein−ligand binding affinities is that all of the binding affinity data recorded in our database are well-documented, all of them having been retrieved from original sources, i.e., the primary references indicated in the corresponding PDB files. As mentioned above, we have collected binding affinity data for 1622 complexes from a total of 5897 − 644 = 5253 complexes whose primary references were available to us. Therefore, the overall "yield ratio" at this literature reviewing step was 1622/5253 = 31%. Since not every protein−ligand complex's binding affinity has been measured, this yield ratio is actually very encouraging and suggests that examining the references associated with those protein−ligand complex structures is an effective strategy for collecting their binding affinities.

It is conceivable that manual review of a large number of references will not be totally free of error. It came to our attention that Carlson and co-workers were developing a database of protein−ligand complexes from the PDB, i.e., Binding MOAD, with their own specific aims.[35] Part of their work involved retrieval of binding affinity data from open literature using a strategy similar to ours. Their list of protein−ligand complexes was not identical to ours but there was considerable overlap between the two. We compared our literature reviewing results with those provided by Carlson et al. regarding the protein−ligand complexes on both lists and found a small number (<3%) of disagreements. These disagreements basically fell into two categories: (1) one group recorded binding affinity for a certain protein−ligand complex while the other did not or (2) the binding affinities recorded by the two groups did not match. In all such cases, we have carefully reexamined the corresponding reference to investigate how the disagreement occurred. This effort found that both PDBbind and Binding MOAD had an error rate around 1% in their binding affinity data. On the basis of the results of this reexamination, we have made appropriate corrections to the binding affinity data collected from our first round of literature reviewing. This comparison of the results derived independently by two groups has certainly reduced the man-made errors in our database to a very low level.

The Refined Set. We anticipated that our database would be of interest to researchers working on structure-based drug design, particularly in the docking and scoring field. In addition to providing a comprehensive collection of binding affinity data, we attempted to compile a standard test set for use in docking and scoring studies. As an example, we recently reported a comparative evaluation of 14 scoring functions by using the 800 protein−ligand complexes in the refined set of PDBbind (version 2002).[36]

The rules that we have applied to the selection of the refined set reflected our concerns with several aspects: rules 1−3 addressed the quality of the complex structure; rule 4 addressed the quality of the binding affinity data, and rule 5 addressed chemical components of the ligand molecule and the protein molecule. There are no further rules to tailor this refined set, but users of our database can always apply additional filters at their own discretion. For example, one may want to use a resolution cutoff of 2.0 Å instead of 2.5 Å, if it is desired to select better defined structures for one's own purposes.

One more convenience we have provided in connection with the refined set is the properly processed structural files. This is important because the original PDB files of protein−ligand complexes often cannot be readily utilized by today's molecular modeling software, and this could prove to be a serious impediment for researchers who want to study these structures. First, the original PDB files save both the protein molecule and the ligand molecule in a single file. This treatment is neither flexible nor efficient for studying the interactions between two entities. Therefore, we have split each complex into a protein molecule and a ligand molecule, which are saved separately. Second, the classic PDB format is most suitable for characterizing biopolymer molecules, which are comprised exclusively of standard residues. However, it is normally not straightforward in a PDB-format file to specify atom types and bond types for normal organic molecules, the crucial information with which molecular modeling software interprets a chemical structure correctly. In our work, the atom types and bond types of each ligand molecule were carefully examined, and each ligand molecule was saved in the Tripos Mol2 format, one of the most popular formats for characterization of organic molecules. Other formats for characterizing small organic molecules, such as the SDF format, can be generated from the Mol2 format easily.

A question frequently asked by the users of PDBbind is why the protein−ligand complex structures in the refined set are not optimized. It is our basic philosophy that we do not distribute geometrically altered PDB structures. It is true that a small number of complex structures, even under an acceptable overall resolution, still have some severe clashes between the protein and the ligand. If not treated with proper optimization, such structures will lead to meaningless results when a force field is used to compute the interactions between the protein and the ligand. To address this problem, as mentioned in the methods section, rule 2 is designed to

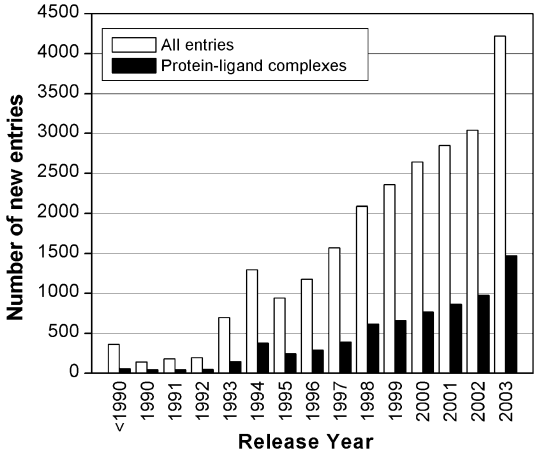

Figure 2. Growth of the Protein Data Bank (based on PDB release No. 107).

avoid the inclusion of complexes with serious clashes between the protein and the ligand in the refined set. If this does not solve the problem completely, users can always perform the structural optimization needed in their own studies.

Update Strategy and Future Developments. Our PDBbind database is a secondary database based on the contents of PDB. PDB itself is growing rapidly. For example, more than 4000 new structures were released by PDB during 2003 (see Figure 2). To allow our database to keep up with the expansion of PDB, we conceived the strategy for its update from the very beginning of our project. We have planned to update our database annually, and by each update, we will process the new entries released by PDB in the previous year, add the results to our database, and name it accordingly. For example, here we report PDBbind version 2003, which covers the contents available in PDB before January first, 2004. The work described in this paper was carried out mainly from March to July 2004. Another annual update, i.e., PDBbind version 2004, is currently under way. We expect that it will become available shortly after the publication of this paper.

In addition to the regular update, we also plan to pursue several new directions in the future to achieve a broader coverage of binding affinity information as follows. (1) If the binding affinity of a given complex cannot be found in the primary reference, we will examine the other references listed in the PDB file. (2) For each protein−ligand complex included in our database, we will also supply the key information of its binding assay, if such information is also available in the cited references. The key information of each binding assay that will be gathered includes binding assay method, major material used in assay, major instruments used for detecting signals, and environmental settings such as temperature, pH level, buffer composition, and salt concentration. Summary of the key information cited above should help users to obtain a quick assessment on the binding affinity data deposited in our database without checking the complete description in the corresponding reference. (3) We will extend our efforts to the other types of protein−ligand complexes. For example, the complexes formed with biological ligand molecules have been ignored in our current work, but they represent certain types of protein−ligand interactions and therefore are also of theoretical significance. In the future, protein−protein and protein−nucleic acid complexes will also be incorporated into our database.

Summary

We have screened the entire PDB (release No. 107, January 2004) to identify the complexes formed between protein molecules and small organic ligand molecules. A systematic review of the references associated with these complexes led to a collection of experimentally determined binding affinities (K_d, K_i, and IC_{50}) for more than 1600 protein−ligand complexes. A total of 900 complexes were selected to form a "refined set", which is of particular value as a standard data set in docking and scoring studies. All of the results have been organized into the Web-accessible PDBbind database, running at http://www.pdbbind.org/. We have also made a concrete plan for developing and updating of our database to keep up with the growth of the PDB. Consequently, a steady growth of our database is expected in the future.

Acknowledgment. This work was supported in part by a pilot grant from the Office of the Vice President of Research, the University of Michigan. The authors thank Dr. Heather Carlson's group at the University of Michigan for generously sharing their collection of protein−ligand binding affinity data before publication. The authors also thank Dr. George W. A. Milne for his critical reading of the manuscript and his editorial assistance.

References

(1) Kuntz, I. D. Structure-based strategies for drug design and discovery. *Science*, **1992**, *257*, 1078−1082.
(2) Verlinde, C. L. M. J.; Hol, W. G. J. Structure-based drug design: Progress, results, and challenges. *Structure* **1994**, *2*, 577−587.
(3) Gane, P. J.; Dean, P. M. Recent advances in structure-based rational drug design. *Curr. Opin. Struct. Biol.* **2000**, *10*, 401−404.
(4) Hardy, L. W.; Malikayil, A. The impact of structure-guided drug design on clinical agents. *Curr. Drug Discov.* **2003**, *Dec*, 15−20.
(5) 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; Vol. 18, pp 41−88.
(6) 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.
(7) Jain, A. N. Scoring noncovalent protein−ligand interactions: A continuous differentiable function tuned to compute binding affinities, *J. Comput.-Aided Mol. Des.* **1996**, *10*, 427−440.
(8) 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.
(9) 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.
(10) 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.
(11) 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.
(12) 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.

(13) 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.
(14) Muegge, I. A knowledge-based scoring function for protein−ligand interactions: Probing the reference state. *Perspect. Drug Discov. Des.* **2000**, *20*, 99−114.
(15) Muegge, I. Effect of ligand volume correction on PMF scoring. *J. Comput. Chem*. **2001**, *22*, 418−425.
(16) 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.
(17) 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.
(18) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein−ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.
(19) 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 Discov. Des.* **2000**, *20*, 115−144.
(20) 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.
(21) 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.
(22) 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.
(23) 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/.
(24) Zhang, J. W.; Aizawa, M.; Amari, S.; Iwasawa, Y.; Nakano T.; Nakata, K. Development of KiBank, a database supporting structure-based drug design. *Comput. Biol. Chem.* **2004**, *28*, 401−407. http://kibank.iis.u-tokyo.ac.jp/.
(25) Roth, B. L.; Kroeze, W. K.; Patel, S.; Lopez, E. The multiplicity of serotonin receptors: Uselessly diverse molecules or an embarrassment of riches? *Neuroscientist* **2000**, *6*, 252−262. http://kidb.cwru.edu/.
(26) 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/.
(27) Chen, X.; Liu, M.; Gilson, M. K. Binding DB: A web-accessible molecular recognition database. *J. Combi. Chem. High-T. Scr.* **2001**, *4*, 719−725. http://www.bindingdb.org/.
(28) Chen, X.; Lin, Y.; Liu, M.; & Gilson, M. K. The Binding Database: Data Management and Interface Design. *Bioinformatics* **2002**, *18*, 130−139.
(29) Chen, X.; Lin, Y.; Gilson, M. K. The Binding Database: Overview and User's Guide. *Biopolymers Nucleic Acid Sci.* **2002**, *61*, 127−141.
(30) Puvanendrampillai, D. and 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/.
(31) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **2004**, *47*, 2977−2980.
(32) Csizmadia, F. JChem: Java Applets and Modules Supporting Chemical Database Handling from Web Browsers. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 323−324.
(33) 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/.
(34) Chalk, A. J.; Worth, C. L.; Overington, J. P.; Chan, A. W. E. PDBLIG: Classification of Small Molecular Protein Binding in the Protein Data Bank. *J. Med. Chem.* **2004**, *47*, 3807−3816.
(35) Hu, L.; Benson, M. L.; Smith, R. D.; Lerner, M. G.; Carlson, H. A. Binding MOAD (Mother of All Databases). *Proteins: Struct. Funct. Bioinf.* In press.
(36) Wang, R.; Lu, Y.; Fang, X.; Wang, S. An Extensive Test of 14 Scoring Functions Using the PDBbind Refined Set of 800 Protein−Ligand Complexes. *J. Chem. Inf. Comput. Sci.* **2004**, *44,* 2114−2125.

JM048957Q