Binding MOAD (Mother of All Databases)

Liegi Hu,[1] Mark L. Benson,[2] Richard D. Smith,[3] Michael G. Lerner,[3] and Heather A. Carlson[1–3]*
[1]*Department of Medicinal Chemistry, College of Pharmacy, University of Michigan, Ann Arbor, Michigan*
[2]*Bioinformatics Program, University of Michigan, Ann Arbor, Michigan*
[3]*Biophysics Research Division, University of Michigan, Ann Arbor, Michigan*

ABSTRACT **Binding MOAD (Mother of All Databases) is the largest collection of high-quality, protein–ligand complexes available from the Protein Data Bank. At this time, Binding MOAD contains 5331 protein–ligand complexes comprised of 1780 unique protein families and 2630 unique ligands. We have searched the crystallography papers for all 5000+ structures and compiled binding data for 1375 (26%) of the protein–ligand complexes. The binding-affinity data ranges 13 orders of magnitude. This is the largest collection of binding data reported to date in the literature. We have also addressed the issue of redundancy in the data. To create a nonredundant dataset, one protein from each of the 1780 protein families was chosen as a representative. Representatives were chosen by tightest binding, best resolution, etc. For the 1780 "best" complexes that comprise the nonredundant version of Binding MOAD, 475 (27%) have binding data. This significant collection of protein–ligand complexes will be very useful in elucidating the biophysical patterns of molecular recognition and enzymatic regulation. The complexes with binding-affinity data will help in the development of improved scoring functions and structure-based drug discovery techniques. The dataset can be accessed at http://www.BindingMOAD.org. Proteins 2005;60:333–340.**

Key words: protein–ligand complex; binding affinity; redundancy; PDB; scoring function; inverse docking; structure-based drug discovery; molecular recognition

INTRODUCTION

Binding datasets for protein–ligand complexes were first used in computational chemistry to develop scoring functions for ligand docking and de novo design of enzyme inhibitors. The earliest relevant dataset was only 45 complexes[1] and more recent sets are 200–800.[2–4] Some sets have been made available online, changing their nature from a flat list of data in a paper to a dynamic and searchable tool for the scientific community. The largest and most useful datasets are outlined below. The strengths of each are noted and the comparative strengths of Binding MOAD are highlighted. Our aim is to make Binding MOAD the largest possible collection of high-quality, protein–ligand complexes available from the Protein Data Bank (PDB)[5] and augment that set with the inclusion of binding data. At this time, Binding MOAD contains 5331 protein–ligand complexes. We have compiled binding data for 1375 (26%) of the protein–ligand complexes.

LPDB

The Ligand–Protein Database (LPDB) has 195 complexes with binding data.[2] LPDB also provides computer generated "docking decoys" to help researchers in developing more accurate scoring functions. We do not plan to add decoys to Binding MOAD, but our dataset is an order of magnitude larger. LPDB has been analyzed to address redundancy of the protein structures. The 195 complexes consist of 51 unique proteins in 21 protein classes.[2]

Binding DB

In one of the first papers announcing the Binding Database (Binding DB), it was reported to contain very high-quality thermodynamic data for 400 binding reactions (90 for biopolymers).[3] Binding DB has recently started to accept the deposition of K_i data, and the number of entries has grown significantly to 3300 binding reactions (http://www.bindingdb.org/bind/stat.jsp). Most of the data is now inhibition constants for biopolymer binding. Binding DB's strength lies in the volumes of information given on experimental conditions used in determining binding information, including raw data in some cases. Though we do not provide isothermal titration calorimetry details like Binding DB, our dataset is larger and we supply structural data from the PDB. The complexes in Binding DB are not cross-linked to their structural data.

PDBbind

PDBbind was created by Shaomeng Wang and coworkers.[4] It contains binding data on 800 complexes with resolution ≤ 2.5 Å (559 structures > 2.5 Å are also provided as a secondary set). PDBbind does not address redundancy, but does note that approximately 200 different types of proteins are present. This set was curated in a similar fashion as Binding MOAD but focuses on complexes with only one ligand in a pocket. PDBbind also excludes any complex binding a simple cofactor such as ATP. Binding MOAD is larger because we do not ignore

*Correspondence to: Heather A. Carlson, University of Michigan, College of Pharmacy, 428 Church St., Ann Arbor, MI 48109-1065. E-mail:carlsonh@umich.edu

Received 16 September 2004; Revised 26 December 2004; Accepted 30 January 2005

Published online 21 June 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20512

cofactors or protein–cofactor–ligand complexes. We also provide information on the structures when we do not have binding data because they are still a valuable resource in database mining. PDBbind only provides structures of complexes for which it has binding data.

PDBbind and Binding MOAD were developed independently at the University of Michigan, Ann Arbor. When we learned of our similar research efforts, we found that our goals were synergistic. The research projects around PDBbind focus on developing scoring functions and searching ligand substructures. Our focus with Binding MOAD is more on protein binding sites and protein flexibility. In sharing binding data between our groups, we found a disagreement of only 1%, which highlights the high accuracy and quality of binding data collected in both groups. Disagreements were simple typos that were easily corrected by consulting the reference again. This arrangement allows both groups to double check all of the data, basically eliminating the errors inherent in hand-processed data. This high level of quality control is unheard of for datasets of this size.

Other Online, Protein–Ligand Databases Without Binding Data

Of course, various improvements are constantly being added to the PDB to provide additional information and viewers to aid understanding protein–ligand complexes.[6,7] However, several other online resources deserve discussion. These databases do not present binding data for the protein–ligand complexes in the PDB, but they do provide useful search tools, various analyses, and viewers of PDB complexes.

Relibase+ and MSDsite are similar datasets that specifically focus on protein–ligand complexes. In 2002, Relibase+ contained 15,454 PDB entries, 50,514 individual ligand sites, and 4530 unique ligands.[8,9] MSDsite is the newest resource in the MSD suite of web-based tools from the European Bioinformatics Institute.[10] However, the description of ligands in both datasets is unusual for our application. We have taken great care to make extensive lists of molecules to exclude as ligands in Binding MOAD. Metal cations like magnesium, inorganic salts such as sulfate, and common crystal additives like polyethylene glycol are not counted as ligands in Binding MOAD, but they are ligands in Relibase+ and MSDsite. They even count modified amino acids in the protein chain as ligands. The strengths of Relibase+ and MSDsite are that they provide powerful search tools for mining their datasets for interaction patterns. A benefit to the description of ligands in Relibase+ and MSDsite is that it allows a user to investigate a protein's interactions with a feature like a modified residue, a structural zinc ion, or an inorganic reactive center in the active site. These groups are simply considered to be part of the protein in Binding MOAD because of its focus on substrates, organic cofactors, and inhibitors. Such an investigation is not possible with Binding MOAD at this time.

PDBsum and MMDB do not focus on protein–ligand interactions, but they provide resources that are very useful for those interests. PDBsum is an online resource from Laskowski and Thornton[11–13] that provides analyses for all structures in the PDB (not just protein–ligand structures). PDBsum provides chemical, enzymatic, and genomic information about the entry, and it provides viewers to analyze protein–ligand interactions. The viewers display secondary structure, ligand interactions, and cavities. MMDB is Entrez's 3D-structure database.[14] Its focus is protein data, but several resources for comparing related sequence and structure have direct relevance for ligand binding.

Redundancy in Protein–Ligand Databases

Binding databases available to-date usually do not address the issue of redundancy. Many protein complexes have more than one bound structure. Many small datasets contain several examples of HIV protease, dihydrofolate reductase, thrombin, trypsin, lysozyme, etc. To address this issue in Binding MOAD, we have analyzed for redundancy and grouped proteins by 90% sequence identity. Of 5331 complexes in Binding MOAD, there are 1780 unique protein families when clustered at 90% identity. In our nonredundant version of Binding MOAD, each protein family is represented by the structure of the tightest binder. Of the 1780 complexes in the nonredundant set, we have obtained binding data for 475. (In cases were binding data was not available, best resolution and other factors were used to choose representatives of the protein families). As we mine this database for general biophysical properties, our results for redundant and nonredundant Binding MOAD can be compared to measure the influence of bias in the structures available in the PDB. Also, inverse docking techniques, where a single ligand molecule is screened against a set of many proteins, will require a nonredundant set of protein complexes.[15,16]

METHODS

Top-Down Approach

Older protein–ligand databases were originally created by reading through the literature and compiling lists of appropriate complexes and their binding affinities. This sort of "bottom up" approach relies on finding good information in a relatively random fashion. We chose a "top down" approach to create Binding MOAD so that it contained every protein–ligand complex with a 3D structure. We started with the entire PDB,[5] removed inappropriate structures, and used the remaining structures to guide our literature searches in a systematic fashion. Since almost all protein structures are annotated with the authors' names and the appropriate reference, a starting point for the literature search is straightforward.

Paring Down the PDB

Perl scripts were written to determine whether each protein structure was an appropriate entry for Binding MOAD (Fig. 1). Our original scripts were written to search through PDB files, and more recently, we have rewritten the rules engine to analyze flat mmCIF files. Our new scripts take advantage of the STAR parsers[17] from the

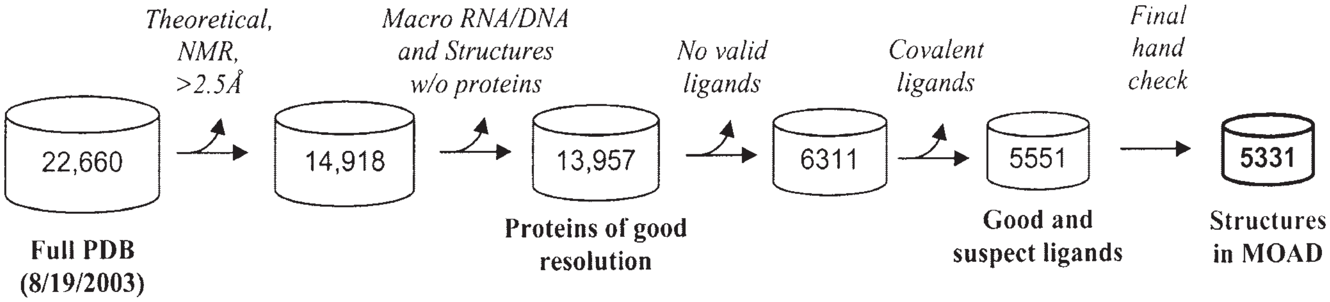

Fig. 1. Criteria to judge all PDB structures for entry into Binding MOAD. The scripts evaluate each structure—one at a time—against all criteria, but this step-by-step diagram is given to show the impact of each criterion.

Research Collaboratory for Structural Bioinformatics (RCSB) and the new mmCIF format from the uniformity project. The mmCIF files have gone through additional checks to correct sequence and EC errors that may exist in the legacy PDB files.[18] By using the mmCIF files, we can keep abreast of the newest improvements in data from the RCSB, making our resource more timely, accurate, and valuable. Our technique is similar to that used by Rognan and coworkers to create sc-PDB, a set of protein binding sites for inverse docking.[16] The major difference is that we did not use a keyword search to identify complexes. Our group and others[4] have found that keyword searches miss complexes that can be identified through analyzing the individual structures. Starting with the entire PDB (22,660 structures on 8/19/2003), we eliminated theoretical models, NMR structures, and structures with poor resolution ($>$ 2.5Å). Large macromolecular complexes between proteins and nucleic acids were removed. However, we wanted to keep any metabolic enzymes that process nucleic acids, so structures with chains of four nucleic acids or less were kept in Binding MOAD. Short chains of 10 amino acids or less were counted as peptide ligands. Short-chain ligands were identified in the SEQRES section (_pdbx_poly_seq_scheme data items in mmCIF). Small molecule ligands were identified in the HET and FORMUL sections (_chem_comp in mmCIF) or in ATOM and HETATM (_atom_site in mmCIF).

Covalently linked ligands were identified by calculating the minimum distance between the protein and each ligand. Minimum distances greater than 2.4 Å were defined as noncovalent. Values between 2.1–2.4 Å were examined visually to determine covalency. Distances less than 2.1 Å were considered covalent unless the short contact was to a metal ion (we considered many common catalytic metals to be part of the protein during this analysis). All short contacts to metals were examined visually. This was crucial in the case of zinc-containing enzymes where a zinc–ligand distance $<$ 2.1 Å is not necessarily a covalent bond.[19] HET groups within 2 Å of another HET were identified as multipart ligands (unless they had partial occupancy and were actually two ligands occupying the same space). If any group of a multipart ligand was covalently linked to the protein, all components are identified as a covalent modification. This was important in the case of sugar chains on glycosylated proteins. Proteins with covalent modifications can still be part of the database if they have another acceptable ligand. If all ligands are covalent or inappropriate (see Table I), the crystal structure is rejected.

Extensive Hand Curation of the Data

The literature citations for all final structures were read to confirm the validity of the ligands and find binding data. Our preference for affinity data is K_d over K_i over IC_{50}. Table I shows the great care that was taken to ensure that entries in Binding MOAD contain only appropriate protein–ligand structures. Short protein–ligand distances and "suspect" ligands were flagged for visual inspection in a more careful hand-check stage. Suspect ligands are crystal additives that are valid only in some cases. "Partial" ligands are molecules that cannot be a ligand on their own but are often a component of multipart ligands. Any HET with $\leq$ 3 heavy atoms is automatically part of this list. The covalency check identifies if these HET are modifications to the protein or a ligand.

The reason for our choice to reject or suspect various HETs in Table I is obvious in many cases. The reader may notice that β-D-*N*-acetylglucosamine (GlcNac, NAG in the PDB) is not on the suspect lists. We found that GlcNac was never used as a crystal additive. It was either part of a ligand or a covalent modification that was readily identified by our scripts.

Modifications to amino acids are on the partial ligand list because they can be part of the protein or part of a peptide ligand. Complexes containing heme groups were rejected because the covalent association of ligands to the central metals made it difficult for us to properly identify the true ligands. In many cases, it was a small molecule (oxygen, carbon dioxide). Of course, this neglects P450s which are very important in medicinal chemistry, toxicology, and pharmacology.[20] We plan to add P450s to Binding MOAD in the future to make it more useful.

Grouping the Proteins to Address Redundancy in the Data

It is desirable to group proteins by related structure and function so that users can compare related systems. Enzyme classification (EC) numbers are used to broadly group entries into "classes" with similar chemical function-

TABLE I. Definition of Unusual HET Groups†

Classification	Type of HET (Examples)
75 Suspect ligands	Sugars (glucose, galactose, fructose, xylose, sucrose, β-D-xylopyranose, trehalose. . .) Small organic molecules (phenol, benzene, toluene, t-butyl alcohol. . .) Membrane components (phosphatidylethanolamine, palmitic acid, decanoic acid. . .) Small metabolites that may be buffer components (citric acid, succinate, tartaric acid. . .)
74 Partial ligands	Chemical groups (amino group, ethyl group, butyl group, methoxy, methyl amine. . .) Inorganic centers of transition state or product mimics (aluminum fluorides, beryllium fluorides, boronic acids. . .) Modifications to amino acids (oxygens of oxidized cys, phosphate group on tyr. . .)
398 Rejected ligands	Unknown or dummy groups (UNK, DUM, “unknown nucleic acid,” “fragment of. . .”) Salts and buffers (Na^+, K^+, CI^-, $PO_4{}^{-3}$, CHAPS, TRIS, tetramethyl ammonium ion. . .) Solvents (DMSO, hexane, acetone, hydrogen peroxide. . .) Crystal additives and detergents (polyethylene glycol, oxtoxynol-10, dodecyl sulfate, methyl paraben, 2,3 propanediol, pentaethylene glycol, cibacron blue. . .) Metal complexes that associate to the protein surface and are used for phase resolution (terpyridine platinum, bis bipyridine imidazole osmium. . .) Metal ions that are part of the protein (Mg^{+2}, Zn^{+2}, Mn^{+2}, Fe^{+2}, Fe^{+3}. . .) Catalytic centers that are part of the protein (4Fe-4S cluster, Ni-Fe active center. . .) Heme groups (heme D, bateriochlorophyll, cobatamin, protoporphyrin IX. . .)

†For brevity, not all compounds are listed.

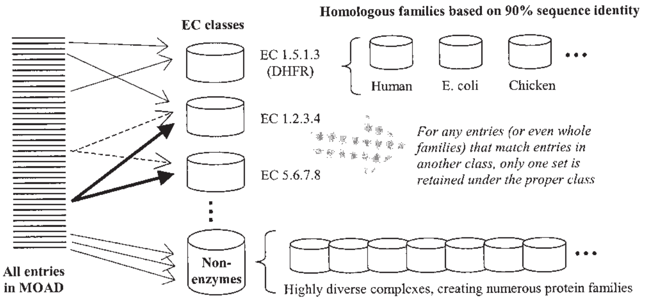
Homologous families based on 90% sequence identity
EC classes
EC 1.5.1.3
(DHFR)
Human
E. coli
Chicken
EC 1.2.3.4
EC 5.6.7.8
For any entries (or even whole families) that match entries in another class, only one set is retained under the proper class
All entries in MOAD
Non-enzymes
Highly diverse complexes, creating numerous protein families

Fig. 2. Currently, 1780 protein families exist over all EC classes. Our routine for grouping proteins by EC number and 90% sequence identity is shown schematically above. The dashed arrows represent a protein with two EC numbers being added to two EC classes. The bold arrows show how a protein with no EC number is added to an EC class by sequence identity. The bold arrows represent a protein that is nearly identical to the dashed protein, so it is added to the same two classes. The gray arrow notes that the homologous protein families are compared in the end, and entries found multiple families are corrected.

ality. Within these classes, proteins are grouped into homologous protein “families” based on sequence.

The EC numbers and protein sequences are pulled from the mmCIF files of all appropriate structures. To compare the sequences in Binding MOAD, we use BLASTp v2.2.7.[21] Defaults are used (E = 10, BLOSOM62 matrix, gap cost = 11, gap extend cost = 1). To create protein families, we use a cutoff of 90% sequence identity like HOMSTRAD,[22] but our grouping of proteins is slightly different than the clustering used for grouping similar sequences at the PDB.[23] The routine is as follows (Fig. 2):

1. Use BLASTp to compare each protein chain of each entry to all other chains.
2. All protein sequences are initially grouped into classes by the EC numbers. If a protein has more than one EC number, it is a member of more than one EC class (dashed arrows in Fig. 2).
3. Structures that do not have an EC number are checked against the existing EC classes. If the sequence is 90% identical to any protein in an EC class, the sequence is added to that class. These entries can be added to more than one class (see bold arrows in Fig. 2).
4. Any structures that do not have matches in the EC classes are initially grouped into a “nonenzyme” class. The “nonenzyme” class can contain enzymes that lack EC numbers or proteins that bind ligands but do not catalyze a reaction.
5. Homologous protein families in each EC class are created using the comparison matrix generated from step 1. At this stage, two entries (A and B in a class) are grouped together into a homologous family if one of the sequences in A is $\geq 90\%$ identical to one of the sequences in B. With 90% sequence identity being so strict for clustering, we always found that any additional chains in entries A and B were also 90% sequence identical.
6. In some cases, every entry in an EC class may be at least 90% identical to all other entries. In those cases, the entire EC class is grouped into one homologous protein family. In the nonenzyme class, there are many, different homologous protein families because of the greater structural diversity.
7. At this point, the homologous families within all EC classes are compared to identify any potential errors.

   a. For proteins with more than one EC number, we find nearly identical protein families in more than one EC class. Only one of the families is retained and placed in the most appropriate EC class.
   b. If an error was made in the EC number of an entry, it will initially be placed into the wrong EC class, but it will have little similarity to the other entries in that class. The misplaced entry will have high similarity

to the entries in another protein family in the correct EC class (e.g., HIV protease was given many different EC numbers for historical reasons, but the entries must be grouped together). The incorrectly labeled entry is moved to the proper class/family. At this time, a missing or incorrect EC number in Binding MOAD can only be corrected if the entry can be identified by its similarity to a homologous protein family in the proper EC class.

8. The "best" entry in a protein family is the structure with the tightest binder. In cases where a family has no entry with binding data, complexes of ligand–protein or ligand–cofactor–protein are chosen over protein–cofactor complexes. The priority for choosing a representative of the protein family is:

   a. Tightest binder (when binding data available)
   b. Best resolution (complexes with ligands preferred over complexes with just cofactors)
   c. Wild-type over structures with site mutations
   d. Most recent deposition date
   e. When all criteria are the same, the representative is chosen based on comments in the crystallography paper.

Annual Updates

We will conduct updates annually to incorporate more structures into Binding MOAD as they become available in the PDB. Our 2004 update began in August. The update procedure is:

1. Use the PDB's list of obsolete entries to identify any existing structures in Binding MOAD that should be removed.
2. Download a new set of mmCIF files. The previous version will be compared to identify all new structures that have been added to the PDB since the last version of Binding MOAD was created.
3. Identify good protein–ligand complexes in the new structures using our current scripts.
4. Any new HETs must be classified as suitable ligands or added to the suspect, partial, or reject lists.
5. The literature portion of the updates should be faster because the number of complexes will be significantly smaller than the existing set and almost all references will be available as online PDF files.
6. Sequences will be added to existing classes and protein families, but regrouping all sequences from scratch may be necessary to periodically confirm our protein classes and families.
7. Each new structure will be compared with the leader of its homologous protein family to determine if the new structure is a better representative of the family.

RESULTS AND DISCUSSION

After examining the PDB contents from August 19th, 2003 (22,660 entries), a total of 5331 valid protein–ligand complexes was obtained. Table II provides detailed information about the functional roles of the proteins contained in Binding MOAD. Our distribution of structures is a little different than that of sc-PDB[16] due to slightly different selection criteria. Three-fourths of the proteins are enzymes, with hydrolases and transferases having the most representatives.

TABLE II. Functional Classification of Entries in Binding MOAD

	Entries in Binding MOAD[b]
Proteins identified with EC numbers[a]	
1.–.–.– (OXIDOREDUCTASE)	810 (15.2%)
2.–.–.– (TRANSFERASE)	1109 (20.8%)
3.–.–.– (HYDROLASE)	1559 (29.2%)
4.–.–.– (LYASE)	335 (6.3%)
5.–.–.– (ISOMERASE)	236 (4.4%)
6.–.–.– (LIGASE)	122 (2.3%)
Total enzymes	4171 (78.2%)
Proteins without EC numbers	
Binding (lectin, streptavidin, agglutinins, etc.)	271 (5.1%)
Signalling, cell cycle, apoptosis	170 (3.2%)
Folding (chaperones, etc.)	13 (0.2%)
Immune (antibodies, immunoglobulins, cytokines, etc.)	141 (2.6%)
Mobility/structural (actin, myosin, etc.)	36 (0.7%)
Toxin/Viral	55 (1.0%)
Transcription, translation, replication proteins	129 (2.4%)
Transport (amino acid transporters, electron transport, etc.)	227 (4.3%)
Enzymes without EC numbers (eg., isopenicillin N synthase)	36 (0.7%)
Other	82 (1.5%)
Total proteins without EC numbers	1160 (21.8%)

[a] Enzyme counts include entries without EC numbers that could be identified through keywords or enzyme names. Some were also identified by 90% sequence identity to entries with EC numbers.

[b] Number of entries and their percentage of all 5331 entries in Binding MOAD.

Binding MOAD contains 2630 unique, valid ligands within the 5331 complexes. Cofactors, inhibitors, and substrates are all considered "ligands" in Binding MOAD. Figure 3 provides the distribution of valid ligands by size. The ligands range from 4–176 heavy atoms. The average number of heavy atoms in Binding MOAD's ligands is 31; an example of the average ligand is ATP which has 31 heavy atoms and a molecular weight of ~500. Figure 3 shows that the number of significantly larger ligands drops off quickly. The largest ligands are peptide, nucleic acid, and sugar chains.

Clustering Binding MOAD into Homologous Protein Families

The protein sequences of the entries in Binding MOAD were grouped into homologous protein families. When the set is clustered at 100% sequence identity, 3639 unique protein sequences were identified. As one would expect

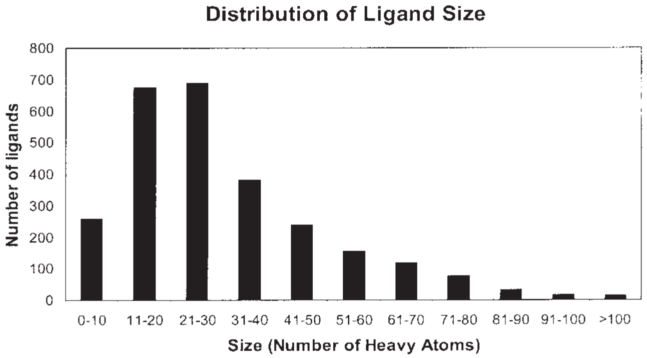

Fig. 3. Distribution of the 2630 unique ligands by size. The average ligand in Binding MOAD has 31 heavy atoms. The largest are small chains of sugars, amino acids, and nucleic acids.

TABLE III. Characteristics of Binding MOAD When Grouped Into Families by Sequence Identity

Clustering criterion	Number of homologous protein families	Size of the largest family (second largest family is also noted)
100% Sequence identity	3639	77 complexes[a] (52)[b]
90% Sequence identity	1780	135 complexes[c] (94)[a]
75% Sequence identity	1526	138 complexes[c] (94)[a]
50% Sequence identity	1330	138 complexes[c] (111)[a]

[a] Trypsin.

[b] Thrombin.

[c] HIV protease.

when the criterion for sequence identity is relaxed, fewer protein families are found and the size of the protein families increases (Table III). Clustering at 90% sequence identity (our preference) produces 1780 homologous protein families with the largest family containing 135 complexes. The largest families are for systems that have been well studied for molecular recognition between proteins and ligands (e.g., trypsin, thrombin, HIV protease, lysozyme, dihydrofolate reductase, etc.). In Figure 4, a histogram of the homologous protein families shows that most of the families have only a few entries. This reflects the emphasis in structural biology to identify new structures and folds, rather than solve many structures of the same protein. Generally, families contain multiple complexes when mutagenesis studies have been performed or various ligands have been co-crystallized.

Nonredundant Binding MOAD

To create a nonredundant version of the dataset, we had to choose unique representatives for each protein family. As outlined in the Methods, we made every effort to identify the tightest binder to represent each family. For the dataset clustered at 90% sequence identity, 1002 of the 1780 families contained only one complex, and so the choice for the representative was obvious. The remaining families contained multiple complexes. For 312 of the families, the representative was easily identified by binding data. Resolution was the deciding factor for 335 of the families (either because there was no binding data or the binding affinity was the same for more than one ligand). Of the remaining families, 46 were chosen based on complexes with ligands being preferred to complexes with only cofactors, 13 were chosen by wild-type over mutated protein, 24 by most recent deposition date, and 48 by other criteria (R factor, comments about ligands in the paper, etc.)

The nonredundant version of Binding MOAD contains 1780 unique proteins. After choosing the complexes for the nonredundant set as outlined above, this set contains binding data for 475 of the unique structures.

Binding-Affinity Data

The binding-affinity data contained within Binding MOAD ranges 13 orders of magnitude, from low fM to high mM values (Fig. 5). The dataset contains mostly K_d and K_i values. Only 159 entries have IC50 data, ranging 60 pM–14 mM. For the 516 entries with K_d data, values range 190 fM–250 mM. The 700 entries with K_i data have the largest range of binding affinity, 11 fM–400 mM.

One of our primary goals is to obtain binding data for all entries in the full set of Binding MOAD (all 5331 complexes). At this time, only 1375 complexes (26%) in Binding MOAD are augmented with binding data. Though this is much larger than other datasets with a few hundred binding affinities,[1–3] we were disappointed to find that so few of the structure papers notes binding-affinity data. A survey of the literature by Wang and coworkers found a similar rate of binding data included in the crystallography papers.[4]

Of course, some of our complexes inherently lack binding data; protein–cofactor structures do not have K_d, K_i, or IC_{50} data for us to report. K_M is the more appropriate binding data for most cofactor–protein complexes, and we have started to collect that information for our complexes. Protein–cofactor structures should be part of the dataset because they can be very important in studying molecular recognition and drug design. For example, patterns in ATP recognition can be extracted from ATP-binding domains to explain enzymatic regulation or develop inhibitors.[24,25]

CONCLUSION

As stated above, we will continue to expand Binding MOAD to contain more binding-affinity data (including the addition of K_M for cofactors). We have also committed to annual updates of the dataset to keep pace with the growth in the PDB. Our most recent annual update began in August 19, 2004. When compared to the previous download of the PDB from August of 2003, almost 5000 new structures were identified (4763 structures). Of those new structures, 877 appeared to be good protein–ligand complexes and an additional 517 required hand-checking to determine if the ligands are appropriate and noncovalently bound. We have recently completed searching the literature to verify these new structures and find binding data. The August 2004 update will soon add a total of 1307 new structures to Binding MOAD, bringing to total to 6638 complexes. Of the new complexes, 418 (32%) have binding

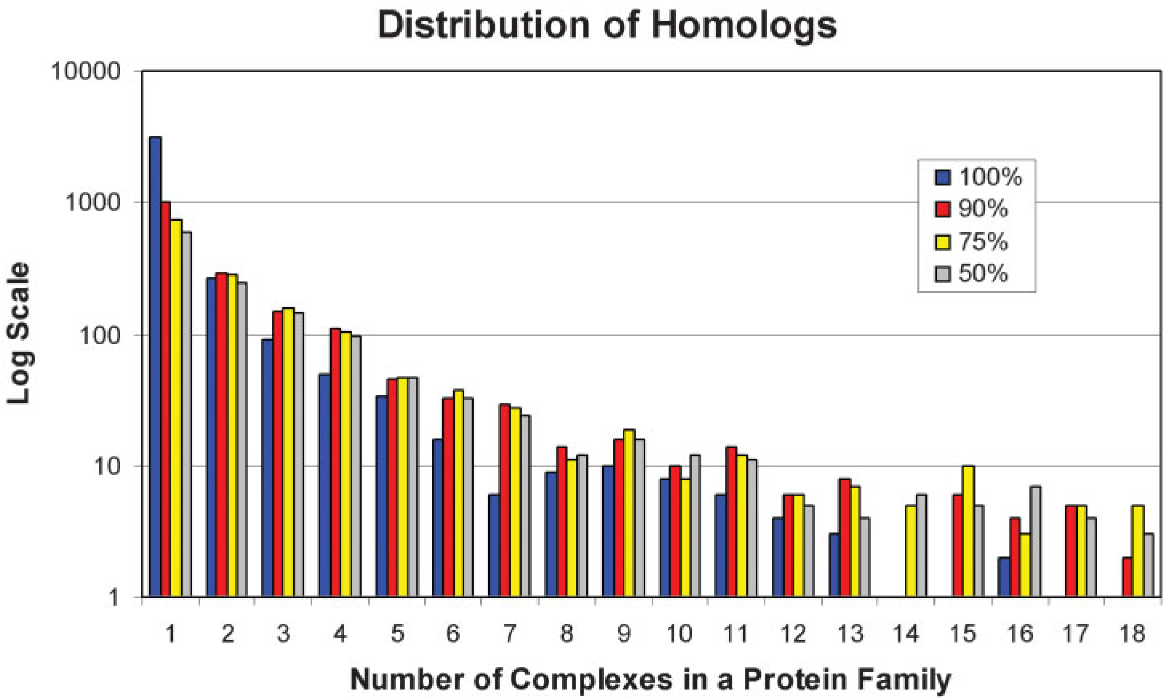

Fig. 4. Histogram of the homologous protein families shows that most families have only a few complexes. There is a near-exponential decrease in the number of larger and larger families. This trend is basically the same for clustering at 100% sequence identity (blue), 90% (red), 75% (yellow), and 50% (gray).

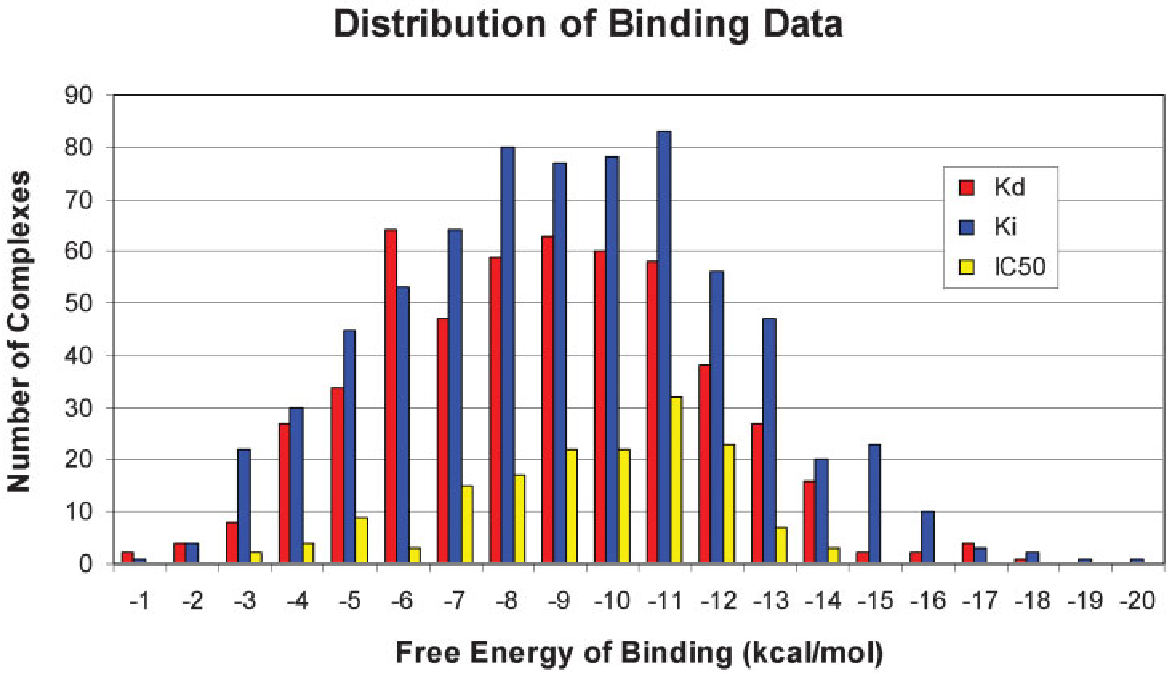

Fig. 5. The distribution of binding-affinity data within Binding MOAD. Data is available as K_d (red), K_i (blue), or IC_{50} (yellow). For this histogram, binding data were converted to free energies by $-RT \ln$ (data). Though not strictly appropriate for many K_i or IC_{50}, this simply provides a comparison for the reader.

data. This addition will bring the total number of complexes with binding data to 1793. We are in the process of annotating this new data for addition to the dataset.

We have made the dataset available online at www.BindingMOAD.org. This web-accessible resource makes our information freely available to other research groups at non-profit organizations (annual licenses are available to the private sector). Data from our perl scripts and our hand curation include PDB id, EC class, homologous protein family, binding-affinity data, and classification of each ligand in the entry (valid vs. invalid). The datapage for each complex in Binding MOAD provides this information to the user. Our scripts also note the reason any PDB structure was excluded (resolution > 2.5 Å, no appropriate ligand, etc.). If a user tries to access a PDB entry that is not part of Binding MOAD, a datapage provides the reason for its exclusion from the dataset.

We are choosing to make the structures available as biological units rather than PDB files. The biological units provide the proper multimer for biological activity. For instance, only the proper dimer is provided when multiple dimers occupy a unit cell, or the proper tetramer is provided from symmetry operations of a unit cell containing only the monomer. This will provide users with the structures that are most related to biological activity and therefore the most appropriate for study.

NOTE ADDED IN PROOF

PDBbind has been updated recently, and it now includes 900 complexes in the refined set and 722 structures in the secondary set. (See Wang R, Fang X, Lu Y, Yang C-Y, Wang S. The PDBbind Database: Methodologies, Refinements and an Updated Version. J Med Chem, in press.)

ACKNOWLEDGMENTS

We thank Prof. Shaomeng Wang and Dr. Renxiao Wang (Medical School, University of Michigan, Ann Arbor) for many helpful discussions and sharing PDBbind prior to publication. This work has been funded through the generosity of the Beckman Young Investigator Award Program (HAC) and the National Institute of General Medical Sciences (HAC GM065372). RDS and MGL would like to thank the University of Michigan's Molecular Biophysics Training Program for support (NIGMS grant GM008270).

REFERENCES

1. Bohm 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.
2. Roche O, Kiyama R, Brooks CL. Ligand-protein database: linking protein-ligand complex structures to binding data. J Med Chem 2001;44:3592–3598.
3. Chen X, Lin Y, Gilson MK. The binding database: overview and user's guide. Biopolymers 2002;61:127–141.
4. 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.
5. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
6. Information about PDBbeta's Ligand Explorer can be found at http://pdbbeta.rcsb.org/pdb/Welcome.do
7. Feng Z, Chen L, Maddula H, Akcan O, Oughtred R, Berman HM, Westbrook J. Ligand Depot: a data warehouse for ligands bound to macromolecules. Bioinformatics 2004,20:2153–2155.
8. Bergner A, Günther J, Hendlich M, Klebe G, Verdonk M. Use of Relibase for retrieving complex three-dimensional interactions patterns including crystallographic packing effects. Biopolymers 2002;61:99–110.
9. Hendlich M, Bergner A, Günther J, Klebe G. Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. J Mol Biol 2003;326:607–620.
10. Golovin A, Dimitropoulos D, Oldfield T, Rachedi A, Henrick K. MSDsite: a database search and retrieval system for analysis and viewing of bound ligands and active sites. Proteins 2005;58:190–199.
11. Laskowski RA, Hutchinson EG, Michie AD, Wallace AC, Jones ML, Thornton JM. PDBsum: a Web-based database of summaries and analyses of all PDB structures. Trends Biochem Sci 1997;22: 488–490.
12. Luscombe NM, Laskowski RA, Westhead DR, Milburn D, Jones S, Karmirantzou M, Thornton JM. New tools and resources for analysing protein structures and their interactions. Acta Cryst D Biol Crystallogr 1998;54:1132–1138.
13. Laskowski RA. PDBsum: summaries and analyses of PDB structures. Nucleic Acids Res 2001;29:221–222.
14. Chen J, Anderson JB, DeWeese-Scott C, Fedorova ND, Geer LY, He S, Hurwitz DI, Jackson JD, Jacobs AR, Lanczycki CJ, and others. MMDB: Entrez's 3D-structure database. Nucleic Acids Res 2003,31:474–477.
15. Rockney WM, Elcock AH. Progress toward virtual screening for drug side effects. Proteins 2002;48:664–671.
16. Paul N, Kellenberger E, Bret G, Müller P, Rognan D. Recovering the true targets of specific ligands by virtual screening of the protein data bank. Proteins 2004;54:671–680.
17. Westbrook JD, Bourne PE. STAR/mmCIF: an ontology for macromolecular structure. Bioinformatics 2000;16:159–168.
18. Westbrook J, Feng Z, Jain S, Bhat TN, Thanki N, Ravichandran V, Gilliland GL, Bluhm WF, Weissig H, Greer DS, and others. The protein data bank: unifying the archive. Nucleic Acids Res 2002;30: 245–248.
19. Christianson DW. Structural biology of zinc. Adv Protein Chem 1991;42:281–355.
20. Verras A, Kuntz ID, Ortiz de Montellano PR. Computer-assisted design of selective imidazole inhibitors for cytochrome P450 enzymes. J Med Chem 2004;47:3572–3579.
21. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
22. Stebbings LA, Mizuguchi K. HOMSTRAD: recent developments of the homologous protein structure alignment database. Nucleic Acids Res 2004;32:D203–D207.
23. Weissig H, Bourne PE. Protein structure resources. Acta Crystallogr D Biol Crystallogr 2002;58:908–915.
24. Mao L, Wang Y, Liu Y, Hu X. Molecular determinants for ATP-binding in proteins: a data mining and quantum chemical analysis. J Mol Biol 2004;336:787–807.
25. Lamb ML. Targeting the kinome with computational chemistry. In: Spellmeyer DC, editor. Annual Reports in Computational Chemistry, Volume 1. Amsterdam: Elsevier BV; 2005. p 185–202.