PDBLIG: Classification of Small Molecular Protein Binding in the Protein Data Bank

Andrew J. Chalk, Catherine L. Worth, John P. Overington, and A. W. Edith Chan*

Department of Molecular Design, Inpharmatica, 60 Charlotte Street, London, W1T 2NU, U.K.

Received March 3, 2004

It is known that proteins can adopt different folds while sharing similar features for recognition of similar substrates or ligands, for example, in the binding sites of enzyme cofactors such as ATP. On the other hand, proteins that have highly flexible binding sites or belong to large and diverse protein families can bind structurally dissimilar ligands, as, for example, in the case of the matrix metalloprotease family. We have developed a database, PDBLIG, that classifies protein domains and ligands. The information stored includes each protein's function, domain class(es), which ligand(s) it binds, and so on. The database can provide valuable knowledge for drug discovery, supporting the answering of questions such as whether the same drug molecule can bind different target protein families and whether these families are related functionally or structurally, which ligand classes (such as metabolites or organic molecules) bind to a particular protein family and whether the ligands are druglike, and which target families bind a wide variety of ligands and whether different ligands are associated with different subfamilies.

Introduction

Now that drug discovery has entered the postgenomic era, attention is being paid to chemogenomics,[1] wherein the research effort no longer concentrates on drug discovery for one target but rather looks at several members of the target family. It is well-known that protein function is intimately related to three-dimensional structure. High-throughput structural genomics projects[2] are now starting to increase the structural information available for genome sequences using various advanced experimental techniques.[3,4] Analysis of protein structure can provide insight into the biochemical functions and mechanisms of proteins.[5,6] The relationship between protein fold and protein function in general is complex.[7] Proteins with similar folds usually have similar function,[8,9] but a single protein fold can sometimes perform many different functions,[10] while a specific biological function can have many different structural supports.[11]

Ligand binding,[12] through the geometrical and chemical complementarities between small molecules and their macromolecular partners, is a key aspect of protein function, mediating the ability of proteins to recognize their natural ligands for transport, signal transduction or catalysis, and also our ability to modulate function through the discovery of drugs.[13] Ligands that bind specifically to certain proteins can lead to enzyme inhibition or modulation of signal transduction and thus can be used as drugs.[14]

However, there are examples, which have probably arisen as a result of divergent evolution, where structurally and/or functionally similar proteins having very low sequence identity bind related but nonidentical ligands. One such example is that of adenyl cyclase and DNA polymerase I. Despite having quite different sequences, the catalytic domains of these two proteins can be superimposed onto one other and they catalyze analogous reactions on similar substrates, which strongly hint at their having diverged from a common ancestor protein. In some enzyme superfamilies, the family members catalyze different overall reactions but share a common chemical strategy for stabilizing the transition state for the formation of a reactive intermediate.[15] Mitchell[16] has studied the relationship between the sequence similarity of α-helical proteins and the molecular similarity of the ligands they bind. He came to the conclusion that related proteins tend to bind similar ligands, although the study was limited to all α-helical proteins only.

Conversely, there are examples of structurally very different proteins performing very similar functions, often as a result of functional convergence. A classic case is that of subtilisin and chymotrypsin, which are both serine proteases and both contain the catalytic triad in remarkably similar configurations despite the proteins themselves having very different structural folds.[17]

Thus, the idea that molecular recognition patterns may be conserved throughout the binding pockets of proteins of similar function imply ligand similarity; i.e., members of the same protein family bind similar ligands. However, there are many examples in which the active site conformations are similar but the entire folds are completely different and in which the folds are similar but the biochemical functions are completely different.[18−20]

The 3D structure of a protein provides the basis for the structure-based design of active compounds. By use of the properties of the ligand binding site along with the assumption of the "lock-and-key" and "induced fit" principle,[21] many computational techniques can be employed to identify and/or design a potential drug molecule. The techniques include virtual screening[22] using a pharmacophore[23] generated from the binding site followed by docking[24] of compound libraries into the

* To whom correspondence should be addressed. Phone: +44 2070744642. Fax: +44 2070744700. E-mail: e.chan@Inpharmatica.co.uk.

10.1021/jm040804f CCC: $27.50
Published on Web 06/19/2004

active site. De novo methods where favorable fragments are inserted into the binding site and grown into molecules are also commonly used.[25,26] Additionally, many cheminformatics techniques, such as topomeric searching and diversity analysis, have been applied to "lead hopping"[27] in order to select chemical structures with similar shape but different chemistry.[28,29] The hope is that these similar-shape molecules may have certain desired biological properties. By hopping to a new lead series, chemistry projects can be steered away from patent protected chemistries, or the new series may have more attractive pharmacokentic profiles. Similar ligands that bind to different protein folds may provide compound hopping information for similar compounds that bind functionally different proteins.

About one-quarter of the protein structures in the Protein Data Bank[30] (PDB, which now has around 27K entries) consist of proteins with bound ligands. In addition, the PDB contains many examples of structures of the same protein with and without a bound ligand or with a variety of different ligands bound. The diversity or similarity of ligands binding to the same protein can reflect the potential for making different interactions within the binding site. The majority of these structures provide valuable information on how the true substrates, cofactors, inhibitors, or ligands bind to their cognate targets. Moreover, the structures provide some degree of comparative information, where, for example, different ligands bind to the same protein of a different species or the same ligand (often a cofactor) binds to structurally different proteins.

In this study, the development of a database, PDBLIG, that classifies the relationships between ligand chemical classes and the protein structural classes to which they bind is reported. Some of the features of PDBLIG are also available in the Relibase,[31] MSD,[32] and Ligand Depot[33] databases. Relibase calculates bond orders and atoms types for ligands in the PDB from their atom coordinates,[34] while MSD and Ligand Depot make use of the macromolecular crystallographic information file (mmCIF)[30,33] for ligand annotation. All these databases have transformed the ligands from PDB format into searchable chemical structures, allowing researchers to perform structural, substructural, and similar searches. In addition, the databases provide links and searchable fields to information in the original PDB files, such as resolution, taxonomy, protein sequences, etc. They have also stored the results of various computational experiments, such as statistics for all interactions of ligands throughout the PDB and automatic superposition of related binding sites.

Primarily, our database stores the chemical structures and their physical and chemical properties, such as chemical names, ligand class, molecular weight, etc. of all the ligands in the PDB, as well as the structural classification of their binding protein using CATH.[35] The contact details between the residues of the protein and the ligand are calculated and schematically represented by the LIGPLOT program.[36] Using the database, we have analyzed the structural diversity/similarity of the ligands, proteins that bind them, and the relationship between the two. We present three examples to illustrate cases where (1) the same ligand binds (implying same function) to a variety of protein folds (structurally and sequence different proteins), (2) a diverse set of ligands (measured by the Tanimoto coefficient) bind to the same protein family (structurally similar proteins), and (3) similar ligands bind to similar or identical protein structures.

We hope that the results of our study can guide the understanding of the relationship between protein fold and type of ligand it binds, providing information for library design or serving as a means to identify regions of nonconservation (and hence specificity) across the binding sites of different protein family members.

Materials and Methods

A. Extraction of Ligands from the PDB Files. The PDBLIG ligand database is generated from the structures in the PDB. Identification of ligands is not always straightforward, particularly in some of the older format PDB files. Molecules can consist of a single hetero group (such as a sulfate ion) or a connected set of groups, which can include amino acid residues or nucleotides. Thus, a molecule is defined as any distinct group of covalently bonded atoms. Connectivity between atoms was calculated using distance cutoffs. Each molecule is named according to the sequence of hetero groups and/or amino acid residues present in the molecule (e.g. ASP-ARG-LEU). Peptide sequences longer than 30 residues and nucleotide sequences are excluded from the molecule list. Protein modifications, such as carbohydrates or covalently bound protease inhibitors, are separated from the protein chain and treated as ligands, and the fact that the ligand is bonded to the protein is noted. The 3D coordinates of the ligand are stored without any energy minimization to reflect the bound conformation of the ligand. Since it is impossible to reconstruct the coordinates of unresolved atoms, ligands are stored exactly as they appear in the PDB; i.e., missing atoms are not recorded.

The bond orders and atomic formal charges for each ligand are derived using a modified version of the HBADD[37] program, which matches het groups to definitions in the PDB het group dictionary.[30] The modified version of HBADD uses a graph algorithm[38] to find correspondence between atoms in the ligand and those in the dictionary. This dictionary lists the bond orders and charges for each het group code found in the PDB, which are then mapped onto the ligands using calculated correspondences. The het groups not present in the dictionary have their dictionary entries created manually. There are also cases where a het group name corresponds to more than one structure, only one of which is defined in the dictionary. For example the het group BNN corresponds to 1,3-diaminobenzylphenylalanine in PDB entry 1a86 and acetyl-*p*-amidinophenylalanine in entry 7kme. Such missing entries are also created manually. The modified HBADD program is sufficiently robust that it can examine several possibilities and select the most appropriate match. Hydrogen addition and conversion of 3D to 2D structures are performed using the dbtranslate utility from Unity 4.3.[39] Acidic and basic compounds (e.g., molecules containing carboxylic acids and amines) are stored in their neutral form to ensure consistency. Physical properties are calculated using the facilities provided by Unity 4.3.

B. Ligand Database. PDBLIG is an Oracle relational database consisting of several components. First, it holds the 2D and 3D ligand structures and their associated property data, including ClogP, molecular weight, number of hydrogen bond donors and acceptors, and rotatable bond count. We have also grouped the ligands into classes based on the het groups present in the molecule. The following classes have been defined: peptides, modified peptides (peptides with nonstandard residues), cofactors (e.g., ATP), modified cofactors, metabolites (compounds found in metabolic pathways, for example, citrate), near metabolites (compounds similar to metabolites) and carbohydrates. Any ligand not falling into any of the above classes is assigned as inorganic where metallic elements are present or as an organic where they are not.

To structurally classify the protein and distinguish their functional units or domains of the proteins, we have used the CATH protein classification system. CATH is a hierarchical classification of protein domain structures, which clusters proteins at four major levels: class (C), architecture (A), topology (T), and homologous superfamily (H). Each protein has a CATH number identifying its classification at each of these four levels (e.g., CATH number 3.10.20.4). Proteins with identical CATH numbers are considered structurally homologous. Each CATH code is further divided into sequence families, where each family has >35% sequence identity. In simple terms, if the first 4 CATH codes are identical, the proteins have the same structural fold. In order for this domain assignment to be useful with respect to ligand binding sites, we need to know their locations within the protein. We extract this information from the CATH domain description file.[34] This along with the domain sequence and CATH code is stored in the database.

If we wish to examine the binding of ligands to CATH domains, it is essential to know which protein residues the ligand interacts with. Hence, we also store protein−ligand contact data. The contacts were determined using two programs, LIGPLOT and HBPLUS.[40] To determine the protein contact residues, information regarding bonding between the ligand and protein is required. HBPLUS calculates both hydrogen bonding and hydrophobic interactions between two non-hydrogen atoms in different molecules within 3.9 Å (default value). We store a list of each residue for which an interaction is found to occur.

Structures in the PDB contain many chemical species that are artifacts of the crystallization process and are potentially not bound at the active site or do not reflect the endogenous function of the protein. From a drug discovery perspective, these compounds are usually irrelevant, and hence, we wish to exclude them from the database. It is reasonable to expect that a ligand bound at the active site of a protein would have many interactions with nearby protein residues. Therefore, a simple method of excluding molecules that are not of interest is to restrict the database to compounds interacting with more than a set number of protein residues. Similarly, we neglect ligands that occur between neighboring molecules in the crystal lattice as well as water. The most common noninteracting small molecule is *N*-acetylglucosamine (NAG), which is found in its noninteracting state in 399 PDB structures, usually as a modification to the protein. In some cases, however, it is the substrate of an enzyme and we wish to retain it. For example, PDB entry 1j92 contains NAG bound to the active site of the protein, and in this case, the compound exceeds our interaction limit and is included in the database. Other common noninteracting organic ligands include glycerol (GOL), 2-hydroxyethyl disulfide (HED or SEO−SEO), 2-methyl-2,4-pentanediol (MPD), and 2-(*N*-morpholino)ethanesulfonic acid (MES). It has already been mentioned that ligand−protein contact information is stored in the database, making it a simple matter to filter the ligands based on the number of contacts.

C. Comparison of Molecules. To allow a meaningful comparison to be made between the ligand molecules in the database, we must first define the features that we wish to compare and have a method to compare these features. For the present study, we have chosen to describe the molecules by the 2D fingerprint produced by tools included with Unity 4.3.[33] This fingerprint is a 988 bit binary string where bits are set to 1 or 0 depending on the presence or absence of specific fragments, heteroatoms, or substructures. Such binary fingerprints are commonly compared using the Tanimoto coefficient[41] (TC), and we have employed this method also. The Tanimoto coefficient is defined as follows:

$$\mathrm{TC} = \frac{N_c}{N_a + N_b - N_c}$$

where N_a and N_b are the number of bits set for fingerprints A and B, respectively, and N_c is the set bits that A and B have in common. In simple terms, the Tanimoto coefficient represents the ratio of the number of features the fingerprints have in common to the number of features that they could potentially have in common. The Tanimoto coefficient ranges from 0 for fingerprints having no bits in common to 1 for identical fingerprints. For some applications, it is more useful to know the dissimilarity or distance between two molecules. This can be achieved by simply subtracting the Tanimoto coefficient from 1.

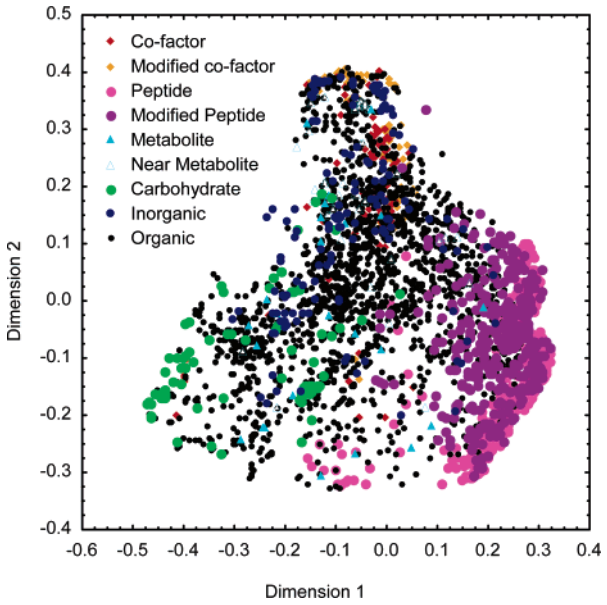

Figure 1. MDS results of the ligand set showing different chemical classes.

D. Analysis of Protein−Ligand Relationships. We have used PDBLIG to study the relationship between protein structures and the structures of the molecules they bind. The version of PDBLIG used for this study was generated from all PDB entries as of April 16, 2002, comprising 17 730 structures of which 9283 had some type of small molecule bound to the proteins. The protein domains were classified using version 2.4 of the CATH database.[34] Ligands were defined as small molecules having contacts with at least five residues from the parent protein. This gives a ligand data set of 3865 compounds with unique structure. Some of these ligands appear in multiple PDB entries. For example, the most frequently occurring ligands at the time of this study are HEME (682), FAD (flavin−adenine dinucleotide 258), and NAD (nicotinamide−adenine dinucleotide 188). Trypsin-like serine proteases (336 unique ligands) and acid proteases, for example, HIV (86 unique ligands), are the proteins that have the most distinct ligands.

The ligand molecules were compared by first calculating an all-by-all dissimilarity (or distance) matrix. Such a matrix is difficult to interpret and visualize. Hence, the multidimensional scaling (MDS)[42] technique was employed to extract the key details in the data. MDS is a method whereby distance data can be visualized in a small number of dimensions, in this case as a 2D map. The points representing molecular fingerprints are arranged in the 2D plane in such a way that the root-mean-square change in the distance when going from the original matrix to the new representation is minimized. To gauge the accuracy of the transformation, we quote the percentage of the original variance retained in each dimension when presenting MDS results. The total of these results gives an indication of the accuracy of the resulting map.

Results and Discussion

A. Analysis of the Ligand Database. The results of the MDS analysis on the entire ligand set are shown in Figure 1. The points are colored according to the chemical class of the ligand. Although only 22% of the

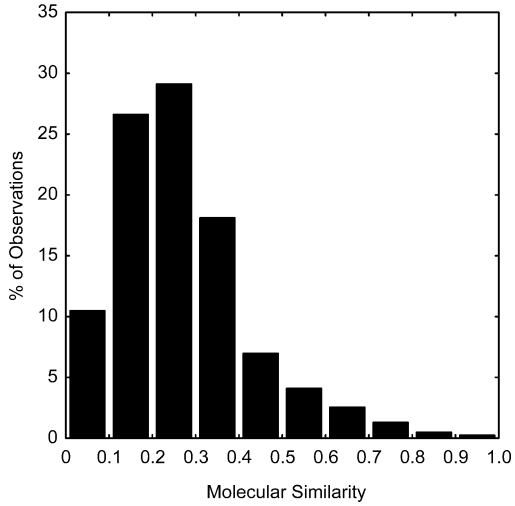

Figure 2. Histogram of similarities from all-by-all ligand comparison. Self-similarities are excluded.

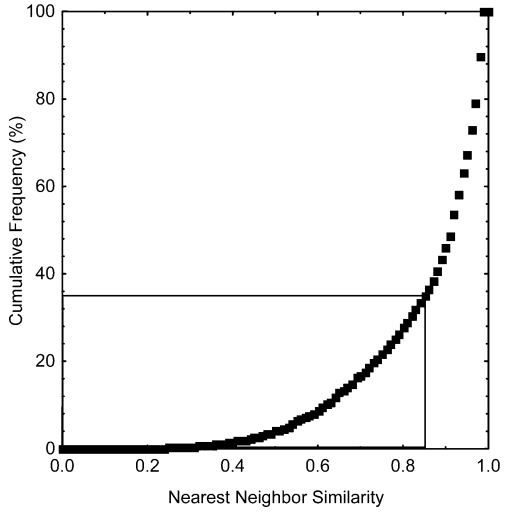

Figure 3. Nearest-neigbor similarity cumulative histogram for the entire ligand set.

original distance matrix is reproduced in this plot, it can be seen that the ligands cluster quite well according to their chemical classes. It is particularly interesting to note that the cofactors and metabolites, ligands found naturally in an organism, cluster together. It can be also seen from this plot that the ligands form quite a diverse set, a Tanimoto distance from left to right being about 0.8 and that from top to bottom being around 0.7.

To examine the diversity of the ligand set in more detail, we examined the distribution of values in the all-by-all similarity matrix (Figure 2). Most of the results show low similarity, clustered in the region from 0.2 to 0.6, indicating that globally the ligand set contains a diverse set of molecules. However, it does not indicate the degree of redundancy that is present in the data. To address this issue, we have examined the similarity of each molecule to its nearest (according to the Tanimoto coefficient) neighbor. In a nonredundant set of compounds we would expect that compounds are dissimilar to their nearest neighbors and we would therefore expect low values for nearest-neighbor similarities. Conversely, for redundant sets of molecules, we would expect high values. The results of this analysis are shown as a cumulative histogram in Figure 3. If we take a Tanimoto coefficient of 0.85 to be indicative of similar molecules,[43] it can be seen that only around 35% of the compounds have nearest-neighbor similarities less than this value, indicating a high degree of redundancy in the database.

The majority of the ligands in the PDB will have been chosen on their ability to bind to the protein of interest, and only a fraction will represent molecules designed as orally available drugs. In fact, about 102 compounds (listed in Table S1 in the Supporting Information), or 2.6% of PDBLIG, which covers about 8.6% of drugs, are present in the Orange Book compiled by the U.S. Food and Drug Administration.[44] Thus, it is interesting to examine the "druglikeness" of the ligand data set. Lipinski and co-workers[45] have examined the properties required for drug molecules to show good absorption properties and proposed a set of four rules, the so-called "rule of five", named after the number 5, which appears in the rules. According to these rules, the following criteria should be satisfied for good oral absorption: number of H-bond donors, ≤5; number of H-bond acceptors, ≤10; ClogP ≤ 5; molecular weight, ≤500. Histograms showing the distribution of these properties are given in Figure 4. We find that 1504, or around 40% of the compounds, are druglike according to these criteria.

B. ATP Binding Sites. Same Ligand, Diverse Protein Structures. Next, we examined the range of proteins that a ligand will bind to. We chose ATP as an example because many classes of current targets, e.g., kinases, phosphodiesterases, adenosine receptors, etc., are targetable through the adenine binding site. In its role as an energy carrier within an organism, ATP is involved in many diverse processes and interacts with a wide range of proteins. It therefore represents a good test case for examining the function and protein structure relationships for a ligand.

At the time of the study, 73 PDB entries that have fully classified CATH domains were found to contain ATP. Of these, some had multiple or similar CATH domains, so after removing these, we were left with 36 unique fully classified domains. This alone demonstrates that a significant amount of sequence variation is present in ATP binding domains. These 36 domains are characterized by 7 unique architectures, orthogonal bundle, barrel, two-layer sandwich, three-layer ($\alpha\beta\alpha$) sandwich, roll, up−down bundle, and complex, indicating that a wide range of protein architectural diversity is also present within the ATP binding domains. The full list of PDB codes, CATH codes, and architectures can be found in Table S2 in the Supporting Information.

However, because only a small portion of any domain binds to ATP, it is useful to examine the variability of residues that actually interact with the ATP (as calculated by LIGPLOT[35]). Figure 5 shows a histogram of the propensities of each of the amino acids to interact with the ATP. The propensities were calculated by normalizing the counts of contacts for each residue type by the number of occurrences of that residue type in the entire PDB.[46] The triphosphate group in ATP contains many negatively charged groups, and so it is not surprising to find the positively charged amino acids arginine and lysine occurring 2.2 and 1.6 times more often, respectively, than would be expected for an average protein. Threonine and glycine are also com-

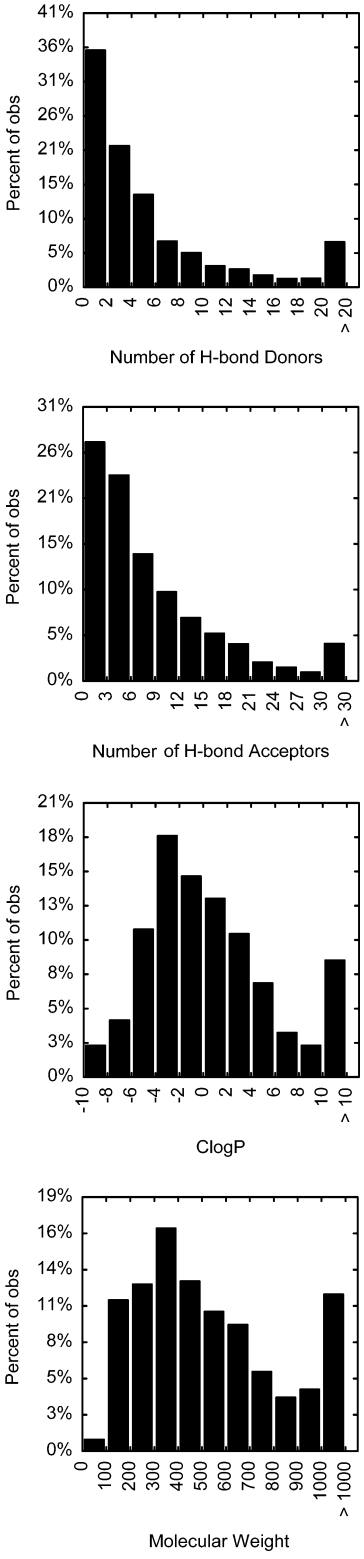

Figure 4. Histograms of ligand properties.

mon, consistent with the p-loop motif GXXXXGKS/T present in many ATPases.[47] Therefore, although the ATP binding domains differ in their primary and tertiary structures, there is some conservation of key residues in the binding site. An analysis of the 3D structure of the binding sites would reveal the three-dimensional features common to such binding sites, and such an analysis has been performed in a number of studies.[19,48]

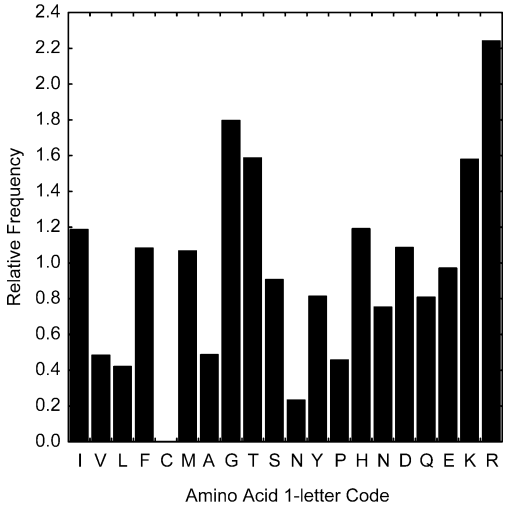

Figure 5. Normalized propensities of amino acids in ATP binding sites. The polar character of the amino acids increases from left to right.

C. Matrix Metalloprotease Ligand Set. Similar Protein Stuctures, Diverse Binding Ligands. Matrix metalloproteases (MMPs) are a family of zinc-dependent endopeptidases that have been shown to play a significant physiological role in tissue remodeling in normal growth and development.[49] CATH defines 13 separate sequence classes for the MMP families. They all have CATH code 3.40.390.10, with their sequence classes numbered: 1.*X*, where *X* ranges from 1 to 13. Of the 13 MMP subfamilies only 8 had structures with bound ligands in the PDB at the time of this study. These ligand names, along with the PDB codes and CATH sequence classes, are shown in Table 1, while the mapping between CATH sequence classes and the MMP members can be found in Table 2. There were 39 distinct ligands in the PDB in 50 MMP PDB entries, two of which were not bound in the active site (see Table 1), bringing the number of ligands considered to a total of 37. It should be also noted that there are several occurrences of the same ligand having different names in different PDB entries. For example, 345 in PDB entry 456c is identical to CBP in 1cxv. Full details are available in Table 1. Figure 6 depicts a diverse selection of 6 of these 37 ligands.

As for the whole ligand set, we have plotted both an all-by-all and a cumulative nearest-neighbor similarity histogram, which can be seen in Figures 7 and 8, respectively. We observe that about 60% of the values in the all-by-all comparison (Figure 7) occur in the range 0.2−0.4, suggesting that the MMP ligands are quite diverse, at least when judged using Unity 2D fingerprints. The results of the nearest-neighbor histogram, shown in Figure 8, suggest that the set is also nonredundant. Around 80% of the compounds have nearest neighbors with a similarity value less than 0.85,[43] compared with the value of 35% for the complete ligand database.

Figure 9 shows the MDS analysis of the MMP ligands. Around 40% of the original information was retained upon conversion to the 2D map.A left−right division, corresponding to the two main classes of MMP ligands, is immediately obvious in Figure 9. The ligands on the right correspond mainly to sulfones and sulfonamides (e.g., DPS in Figure 6) containing multiple aromatic

Table 1. PDB Codes, Ligand Names, and CATH for the MMP Family

PDB code	ligand name	CATH sequence code
830c	RS1	1.9
456c	345	1.9
1cxv	CBP[a]	1.10
966c	RS2	1.1
1bzs	EPE[b]	1.2
1a86	HMI-ASP−BNN	1.2
1a85	HMI-ASN−BNN	1.2
3ayk	CGS	1.1
4ayk	CGS	1.1
1bm6	HAV-3MP-MSB[c]	1.3
1eub	HAV-3MP-MSB[c]	1.9
1b3d	S27	1.3
1b8y	IN7	1.3
1biw	S80	1.3
1bqo	N25	1.3
1bzs	BSI	1.2
1i76	BSI	1.2
1c3i	TR1	1.3
1c8t	TR1	1.3
1caq	DPS	1.3
1cgl	CBZ-ABU-LEU-PHE-EMR	1.1
1ciz	DPS	1.3
1d5j	MM3	1.3
1d7x	SPC	1.3
1jk3	BAT	1.13
1mmb	BAT	1.2
1jao	BTP-ASP-GM1	1.2
1hv5	CPS[b]	1.12
1fbl	HTA	1.4
1fls	WAY	1.9
1fm1	WAY	1.9
1g4k	HQQ	1.3
1hfc	HAP	1.1
1mnc	PLH[d]	1.2
1hfs	L04	1.3
1hv5	RXP	1.12
1i73	PRO-LEU−PAT	1.2
1jan	PRO-LEU-GLY-HOA	1.2
1jap	PRO-LEU-GLY-HOA	1.2
1jaq	HMP-ASP-GM1	1.2
1jj9	BBT	1.2
1kbc	HLE-RIN	1.2
1sln	INH	1.3
2srt	INH	1.3
2tcl	RO4	1.1
1ums	HAE-MOP-LEU-PHE-NH2	1.3
1umt	HAE-MOP-LEU-PHE-NH2	1.3
2usn	IN8	1.3
1usn	IN9	1.3
3usn	ATT	1.3

[a] Identical to 345 in PDB entry 456c. [b] Not in active site. [c] Identical to CGS in PDB entries 3ayk and 4ayk. [d] Identical to HAP in pdb code 1hfc.

Table 2. Mapping of CATH Sequence Classes to MMP Subfamilies

CATH sequence class (3.40.390.10)	MMP subfamily[a]
1.1	mmp1
1.2	mmp8
1.3	mmp3
1.4	mmp1 (pig)
1.9	mmp13
1.10	mmp13 (mouse)
1.12	mmp11
1.13	mmp12

[a] Refers to the human protein unless otherwise stated.

groups, while those on the left contain peptide or peptide-like molecules.

There also appears to be little relationship between the CATH sequence class and the clustering of the ligands. For the peptide ligands (i.e., those in the left

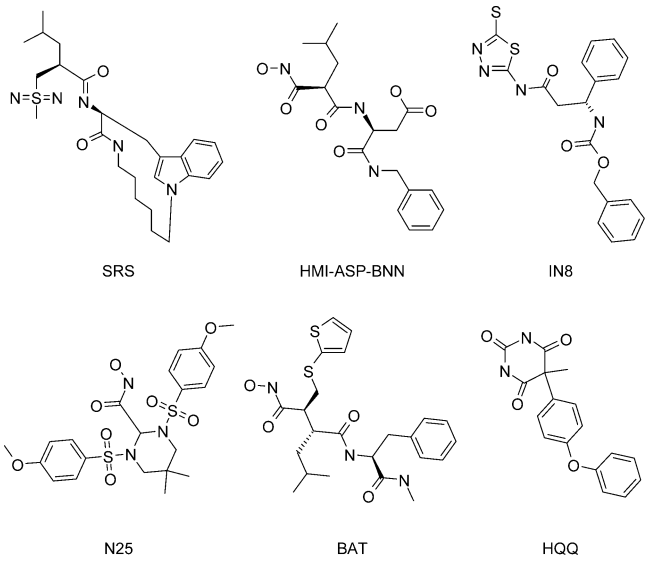

Figure 6. Selection of MMP ligands.

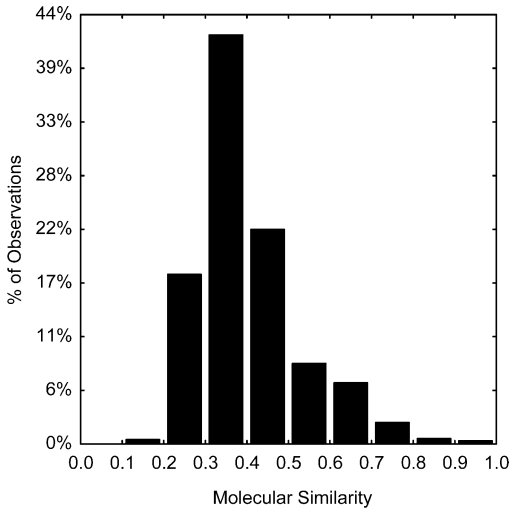

Figure 7. Histogram of the all-by-all similarity matrix for the MMP ligands.

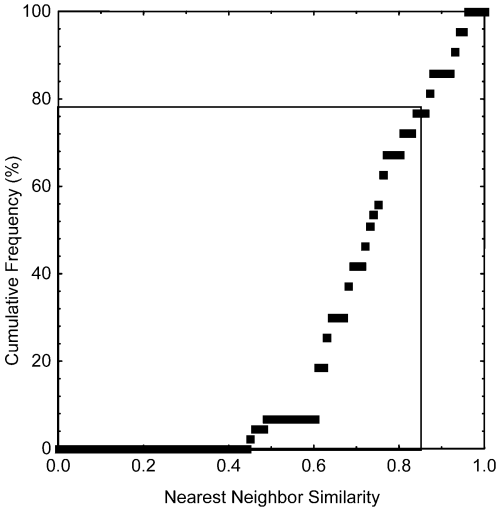

Figure 8. Nearest-neighbor similarity for the MMP data set.

half of the plot), there is no obvious clustering by subfamily. In some cases, very similar or even identical ligands are bound to proteins in different sequence classes. On the other hand, the organic ligands on the

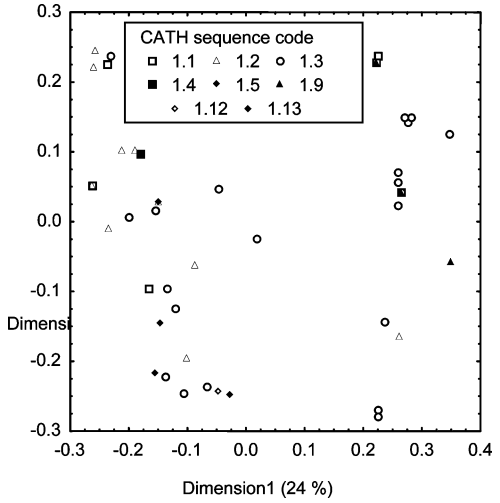

Figure 9. MDS map for the MMP ligand set.

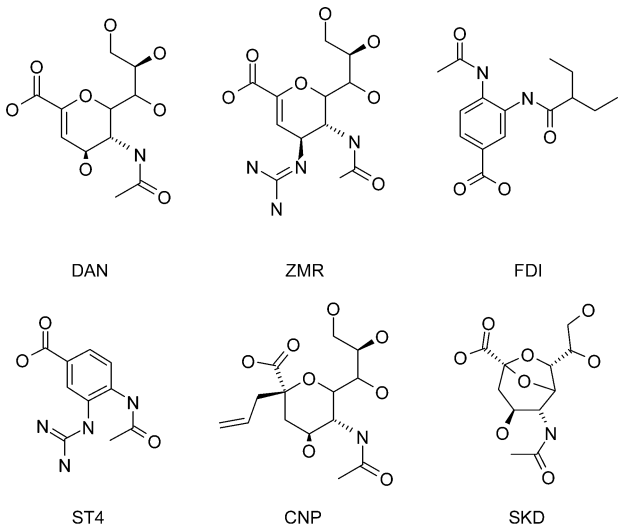

Figure 10. Selection of neuraminidase ligands.

right of Figure 9 tend to mostly bind to proteins in sequence class 1.3 (mmp3). However, ligands bound to proteins in four other sequence subfamilies are also present here. The fact that similar ligands bind to different proteins in the same protein family could indicate that achieving high selectivity among the members of this family might be difficult.

D. Neuraminidase. Similar Protein Structures, Similar Ligands. Neuraminidase is a member of CATH class 2.120.10.10, which contains seven sequence classes whose PDB entries contain ligands. The majority of these proteins are from the influenza virus. PDB codes, ligand names, and CATH sequence codes for the neuraminidase family are shown in Table 3, while the mapping of CATH sequence codes to the corresponding species or influenza strain is shown in Table 4. Of the 30 unique small molecules found in members of this class, 5 were found to be protein modifications (see Table 3) and have therefore been excluded from the following discussion. Some examples of the remaining 25 structures, corresponding to 44 PDB entries are shown in Figure 10.

Figure 11 shows a histogram of ligand similarity values from the all-by-all similarity matrix for the neuraminidase ligands together with comparable results for the MMP ligand set. The first feature of note is the

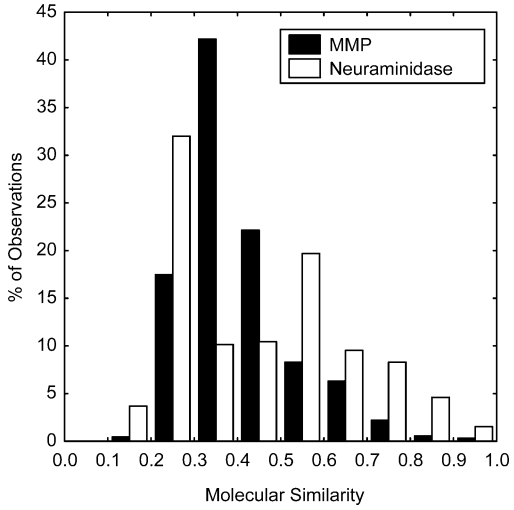

Figure 11. Histogram of all-by-all similarities for the neuraminidase and MMP ligand sets.

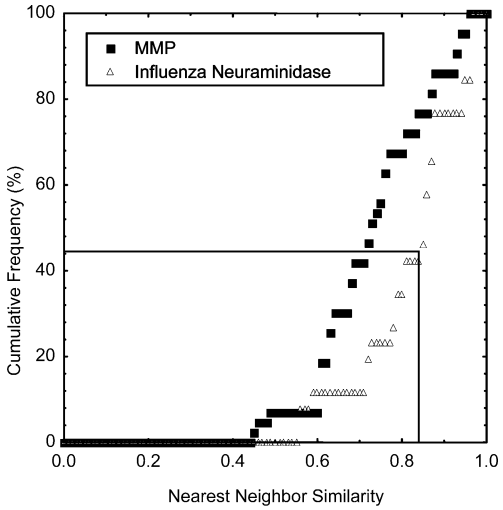

Figure 12. Cumulative histogram of nearest-neighbor similarity for the neuraminidase and MMP ligand sets.

peak between 0.2 and 0.3 caused by two distinct clusters being present. This is discussed further below. It can also be seen that apart from this peak, the Tanimoto coefficients of the neuraminidase ligands have significantly greater occurrence at high values relative to those of MMP, indicative of the neuraminidase ligands currently sampled in PDB entries having higher similarity. The neuraminidase set is also much more redundant than the MMP set, as can be seen in Figure 12. Only around 40% of the ligands have nearest-neighbor similarities less than 0.85 compared to 80% of the MMP ligands. In fact, when we take a closer look at the ligands, they fall into four series. Examples of each ring system can be seen in Figure 10. The first corresponds to DAN and ZMR, the second to FDI and ST4, the third to CNP, and the fourth to SKD. Apart from a small number of exceptions, the four series are built upon variation at two points in the parent structure. This information is very valuable in the design of inhibitors or studying the SAR of the ligand−protein interaction when activity data are available. Very often, the diversity of molecules that bind to a protein reflects the flexibility of the binding site. However, we point out that in many cases, low diversity of the ligands found to bind

Table 3. PDB Codes, Ligand Names, and CATH Sequence Classes for Members of the Neuraminidase Family

PDB code	ligand name	scaffold type[a]	CATH sequence code
1a4g	ZMR	alip	2.1
1a4q	DPC	alip	2.1
1b9s	FDI	arom	2.2
1b9t	RAI	arom	2.2
1b9v	RA2	arom	2.2
1bji	G21	alip	1.1
1dim	EQP	alip	3.1
1eus	DAN	alip	4.1
1f8b	DAN	alip	1.1
1f8c	4AM	alip	1.1
1f8d	9AM	alip	1.1
1f8e	49A	alip	1.1
1inf	NAG[b]		2.2
1inf	ST4	arom	2.2
1ing	NAG-NAG-MAN-MAN-MAN-MAN[b]		1.2
1ing	ST5	arom	1.2
1inh	NAG-NAG-MAN-MAN-MAN-MAN[b]		1.2
1inh	ST6	arom	1.2
1inv	EQP	alip	2.2
1inw	AXP	alip	1.2
1inx	EQP	alip	1.2
1iny	EQP	alip	1.1
1ivb	NAG[b]		2.2
1ivb	ST1	arom	2.2
1ivc	NAG-NAG[b]		1.2
1ivc	ST2	arom	1.2
1ive	ST3	arom	1.2
1ivf	DAN	alip	1.2
1mwe	SIA	alip	1.1
1nma	NAG-NAG-MAN-MAN-MAN-MAN[b]		1.1
1nmb	NAG-NAG-MAN-MAN-MAN-MAN-MAN[b]		1.1
1nnb	DAN	alip	1.1
1nnc	GNA	alip	1.1
1nsc	SIA	alip	2.1
1nsd	DAN	alip	2.1
1sli	DAN	alip	6.1
2bat	SIA	alip	1.2
2qwb	SIA	alip	1.1
2qwc	DAN	alip	1.1
2qwd	4AM	alip	1.1
2qwe	GNA	alip	1.1
2qwf	G20	alip	1.1
2qwg	G28	alip	1.1
2qwh	G39	alip	1.1
2qwi	G20	alip	1.1
2qwj	G28	alip	1.1
2qwk	G39	alip	1.1
2sim	DAN	alip	3.1
2sli	SKD	alip	6.1
3sil	GOL[b]		3.1
3sli	SKD	alip	6.1
4sli	CNP	alip	6.1

[a] Arom refers to aromatic scaffold, alip referes to aliphatic. [b] These groups are protein modifications and were not considered.

Table 4. Mapping of CATH Sequence Classes to Species or Influenza Strain for the Neuraminidase Family

CATH sequence class (2.120.10.10)	species (strain)
1.1	Influenza (A/tern/Australia/g70c)
1.2	Influenza (A/subtype n2)
2.1	Influenza (B/Beijing/1/87)
2.2	Influenza (B/lec/40)
3.1	*Salmonella typhimurium*
4.1	*Macromomospora viridifaciens*
6.1	*Macrobdella decora*

to a specific protein may be due to synthetic reasons or the specific interests of the scientists performing the crystallization.

Figure 13 shows the MDS analysis for the neuraminidase ligands; 69% of the original information was retained in this transformation. Once again, we see two distinct classes of molecules, the main difference between them being an aliphatic or aromatic scaffold (see Table 3). Ligands on the left have an aromatic ring

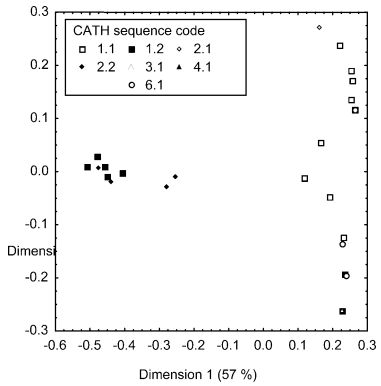

Figure 13. MDS map for the neuraminidase ligand set. Points are labeled by CATH sequence code.

scaffold (e.g., FDI, Figure 10), and those on the right are aliphatic (e.g., ZMR). However, the groups attached to the scaffolds are found to be quite similar; for example, functional groups such as guanidine, glycerol, and acetamide are common to all classes. This suggests a possible pitfall in the use of 2D fingerprints. A lower similarity score can result from a difference in the scaffold, even if the groups interacting with the protein are similar or even identical.

It can also be seen from Figure 13 that the ligands cluster much better with the sequence subfamily of the protein than observed in the case of MMP. The ligands with aromatic scaffolds on the left correspond to sequence classes 1.2 and 2.2, and those on the right correspond largely to class 1.1. This result is somewhat counterintuitive; it would be expected that because of the higher sequence identity of classes 1.1 and 1.2, the ligands of class 1.2 would be more similar to those of class 1.2 than class 2.2. It is quite possible that this unusual clustering is due to the fact that the relevant structure determinations, for example, proteins from class 1.2 with aliphatic ligands, have simply not being performed. Indeed, it is known that several members of the aliphatic series are active against both type A and type B influenza neuraminidases.[50] It is also likely that were the arrangement of groups in 3D space to be considered, we would find the clusters much less well defined than the 2D fingerprints suggest.

Conclusions

We have described the database PDBLIG that incorporates a range of information relevant to ligands and their interaction with proteins. Some of the usefulness of this database has been illustrated by three examples. The first is ATP, a ligand that binds to a wide range of proteins. We are able to quickly identify and classify the proteins that bind to a specific ligand or to a set of similar ligands.

The second example is a protein family that binds a diverse range of ligands. It is generally assumed that molecular similarity implies similar biological activity for both protein and ligand, but the converse is not always true. The MMP family of proteins, whose ligands can be divided into two highly distinct classes, illustrates this point. Although they must have similar arrangements in 3D space of groups interacting with the protein, this similarity does not necessarily imply that the 2D structures exhibit high similarity.

Finally, the case of neuraminidase further reinforces this issue. We again see the presence of two clusters, in this case caused by two differing scaffolds. However, in this case the scaffold does not interact with the protein; it only ensures that the attached groups are in the optimal position. Hence, the existence of the two classes is not relevant to the binding of these ligands. This distinction would most likely disappear if the 3D structure were taken into account.

This study has related cheminformatics data, such as ligand structures, and their physical and chemical properties to bioinformatics data, such as protein structural folds and protein−ligand interactions. The data in PDBLIG might help answer some of the questions relating to drug discovery programs. For example, how many functionally different proteins bind the same cofactor, metabolite, or ligand? How similar (or dissimilar) are the ligands that bind to different members of the same protein target family? Obviously, further work is needed to answer more specific questions. For example, how similar (or dissimilar) are the binding sites of all the proteins that bind to ATP? If the binding sites of different members of the same target family are similar, how is selectivity targeted? In our study, we have only measured small-molecule structure similarity using 2D fingerprints. With the 3D information provided by the X-ray data from the PDB, comparison can in fact be made in 3D, although it is likely to be much slower. The methodology would involve the molecular alignment of compounds binding to the same site followed by generation of 3D pharmacophores. Alternatively, it would also be possible to generate 3D pharmacophores from the protein binding sites.

In summary, PDBLIG provides valuable information that is crucial for drug design. In the future, a 3D approach to both the ligands and binding site analysis will be necessary to answer even more specific questions.

Acknowledgment. We thank Roman Laskowski and Alex Michie for providing the programs LIGPLOT and HBADD.

Supporting Information Available: Table S1 listing the compounds in Orange Book (which are also found in the PDB), their PDB codes, and ligand names; Table S2 listing the PDB codes, CATH codes, and architectures of the ATP binding domains. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Chan, A. W. E.; Overington, J. P. Recent development in chemoinformatics and chemogenomics. *Annu. Rep. Med. Chem.* **2003**, *38*, 285−294.
(2) The Human Genome. *Nature* **2001**, *409*, 813−849.
(3) Drenth, J. *Principles of Protein X-ray Crystallography*; Springer: New York, 1999.
(4) Siegal, G.; van Duynhoven, J.; Baldus, M. Biomolecular NMR: recent advances in liquids, solids and screening. *Curr. Opin. Chem. Biol.* **1999**, *3*, 530−536.
(5) Orengo, C.; Todd, A.; Thronton, J. From protein structure to function. *Curr. Opin. Struct. Biol.* **1999**, *9*, 374−382.
(6) Martin, A.; Orengo, C.; Hutchinson, E.; Michie, A.; Wallace, A.; Jones, M.; Thronton, J. Protein folds and functions. *Structure* **1998**, *6*, 875−884.
(7) Todd, A.; Orengo, C.; Thronton, J. Evolution of function in protein superfamilies. *J. Mol. Biol.* **2002**, *307*, 1113−1143.
(8) Zarembinski, T. I.; Hung, L.-W.; Mueller-Dickmann, H.-J.; Kim, K.-K.; Yokota, H.; Kim, R.; Kim, S.-H. Structure-based assignment of the biochemical function of a hypothetical protein: A test case of structural genomics. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 15189−15193.
(9) Hwang, K. Y.; Chung, J. H.; Kim, S.-H.; Han, Y. S.; Cho, Y. Structure-based identification of a novel NTPase from *Methanococcus jannaschii*. *Nat. Struct. Biol.* **1999**, *6*, 691−696.
(10) Orengo, C.; Pearl, F.; Bray, J.; Todd, A.; Martin, A.; Lo, C.; Thronton, J. The CATH database provides insight into protein structure/function relationships. *Nucleic Acids Res.* **1999**, *27*, 275−279.
(11) Russell, R.; Sasieni, P.; Sternberg, J. Supersites within superfolds. Binding site similarity in the absence of homology. *J. Mol. Biol.* **1998**, *282*, 903−918.
(12) Gohlke, H.; Klebe, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem., Int. Ed.* **2002**, *41*, 2644−2676.
(13) Campbell, S. J.; Gold, N. D.; Jackson, R. M.; Westhead, D. R. Ligand binding: functional site location, similarity and docking. *Curr. Opin. Struct. Biol.* **2003**, *13*, 389−395.
(14) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.
(15) Gerlt, J. A.; Babbit, P. C. Mechanistically diverse enzyme superfamilies: the importance of chemistry in the evolution of catalysis. *Curr. Opin. Chem. Biol.* **1998**, *2*, 607−612.

(16) Mitchell, J. B. O. The relationship between the sequence identities of alpha helical proteins in the PDB and the molecular similarities of their ligands. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1617−1622.
(17) Todd, A. E.; Orengo, C. A.; Thornton, J. M. Evolution of protein function, from a structural perspective. *Curr. Opin. Chem. Biol.* **1999**, *3*, 548−556.
(18) Brenner, S. E.; Chothia, C.; Hubbard, T. J. P. Population statistics of protein structures: Lessons from structural classifications. *Curr. Opin. Struct. Biol.* **1997**, *7*, 369−376.
(19) Murzin, A. G. How far divergent evolution goes in proteins. *Curr. Opin. Struct. Biol.* **1998**, *8*, 380−387.
(20) Russell, R. B.; Sasieni, P. D.; Sternberg, M. J. E. Supersites within superfolds. Binding site similarity in the absence of homology. *J. Mol. Biol.* **1998**, *282*, 903−918.
(21) Koshland, D. E., Jr. The key−lock theory and the induced fit theory. *Angew. Chem., Int. Ed. Engl.* **1994**, *33*, 2375−2378.
(22) Böhm, H. J.; Stahl, M. Structure-based library design: molecular modelling merges with combinatorial chemistry. *Curr. Opin. Struct. Biol.* **2000**, *4*, 283−286.
(23) Martin, Y. C. In *Designing Bioactive Molecules*; Martin, Y. C., Willett, P., Heller, S. R., Eds.; Americal Chemical Society: Washington, DC, 1995.
(24) Lengauer, T.; Rarey, M. Computational methods for biomolecular docking. *Curr. Opin. Struct. Biol.* **1996**, *6*, 402−406.
(25) Murcko, M. A**.** Recent Advances in Ligand Design Methods. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: New York, 1997; Vol. 11, pp 1−66.
(26) Clark, D. E.; Murray, C. W.; Li, J. Current Issues in de Novo Molecular Design. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: New York, 1997; Vol. 11, pp 67−125.
(27) Hecht, P. High-throughput screening: beating the odds with informatics-driven chemistry. *Curr. Drug Discovery* **2002**, January, 21−24.
(28) Andrews, K. M.; Cramer, R. D. Toward general methods of targeted library design: Topomer shape similarity searching with diverse structures as queries. *J. Med. Chem.* **2000**, *43*, 1723−1740.
(29) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behaviour: a useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.
(30) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242. Website: http://www.rcsb.org/pdb/.
(31) Hendlich, M. Databases for Protein−Ligand Complexes. *Acta Crystallogr.* **1998**, *D54*, 1178−1182. 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. Website: http://relibase.ebi.ac.uk.
(32) Golovin, A.; Oldfield, T. J.; Tate, J. G.; Velankar, S.; Barton, G. J.; Boutselakis, H.; Dimitropoulos, D.; Fillon, J.; Hussain, A.; Ionides, J. M. C.; John, M.; Keller, P. A.; Krissinel, E.; McNeil, P.; Naim, A.; Newman, R.; Pajon, A.; Pineda, J.; Rachedi, A.; Copeland, J.; Sitnov, A.; Sobhany, S.; Suarez-Uruena, A.; Swaminathan, J.; Tagari, M.; Tromm, S.; Vranken, W.; Henrick, K. E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res.* **2004**, *D32*, 211−216. Website: http://www.ebi.ac.uk/msd/.
(33) Bourne, P.; Berman, H. M.; Watenpaugh, K.; Westbrook, J.; Fitzgerald, P. M. D. The macromolecular Crystallographic Information File (mmCIF). *Methods Enzymol.* **1997**, *277*, 571−590. Website: http://ligand-depot-i.rutgers.edu/.
(34) Hendlich, M.; Rippmann, F.; Barnickel, G. BALI: Automatic assignment of bond and atom types for protein ligands in the Brookhaven Protein Databank. *Chem. Inf. Comput. Sci.* **1997**, *37*, 774−778.
(35) Orengo, C. A.; Michie, A. D.; Jones, S.; Jones, D. T.; Swindells, M. B.; Thornton, J. M. CATH—A hierarchic classification of protein domain structures. *Structure* **1997**, *5*, 1093−1108.
(36) Wallace, A. C.; Laskowski, R. A.; Thornton, J. M. LIGPLOT: A program to generate schematic diagrams of protein−ligand interactions. *Protein Eng.* **1995**, *8*, 127−134.
(37) HBADD: one of the computer programs from LIGPLOT.
(38) Sedgewick, R. *Algorithms in C: Part 5: Graph Algorithms*, 3rd ed.; Addison-Wesley: Boston, MA, 2002.
(39) *Unity 4.3*; Tripos Inc., 1699 S. Hanley Road, St. Louis, MO 63144.
(40) McDonald, I. K.; Thornton, J. M. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **1994**, *238*, 777−793.
(41) For a review of many similarity measures, see the following. Willet, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(42) Cox, T. F.; Cox, M. A. A. In *Multidimensional Scaling*; Chapman & Hall: London, 1994.
(43) Matter, H. Selecting optimally diverse compounds from structure databases: A validation study of two-dimensional and three-dimensional molecular descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.
(44) *Electronic Orange Book. Approved Drug Products with Therapeutic Equivalence Evaluations*, 23rd ed.; U.S. Department of Health and Human Services: Washington, DC, 2003. Website: http://www.fda.gov/cder/ob/default.htm.
(45) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.
(46) Singh, J.; Thornton, J. M. *Atlas of Protein Side-Chain Interactions*; IRL Press: Oxford, U.K., 1992; Vol. I, p 10. Website: http://www.biochem.ucl.ac.uk/bsm/sidechains/index.html.
(47) Walker, J. E.; Saraste, M.; Runswick, M. J.; Gay, N. J. Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.* **1982**, *1*, 945−951.
(48) De Moliner, E.; Brown, N. R.; Johnson, L. N. Alternative binding modes of an inhibitor to two different kinases. *Eur. J. Biochem.* **2003**, *270*, 3174−3181.
(49) Matrisian, L. M. The Matrix Degrading Proteinases. *BioEssays* **1992**, *14*, 455−463.
(50) Smith, P. W.; Sollis, S. L.; Howes, P. D.; Cherry, P. C.; Starkey, I. D.; Cobley, K. N.; Weston, H.; Scicinski, J.; Merritt, A.; Whittington, A.; Wyatt, P.; Taylor, N.; Green, D.; Bethell, R.; Madar, S.; Fenton, R. J.; Morley, P. J.; Pateman, T.; Beresford, A. Dihydropyrancarboxamides related to zanamivir: A new series of inhibitors of influenza virus sialidases. 1. Discovery, synthesis, biological activity, and structure−activity relationships of 4-guanidino- and 4-amino-4*H*-pyran-6-carboxamides. *J. Med. Chem.* **1998**, *41*, 787−797.

JM040804F