sc-PDB: an Annotated Database of Druggable Binding Sites from the Protein Data Bank

Esther Kellenberger, Pascal Muller, Claire Schalon, Guillaume Bret,† Nicolas Foata, and Didier Rognan*

CNRS UMR7175-LC1, Institut Gilbert Laustriat, 74 Route du Rhin, F-67401 Illkirch Cédex, France

Received September 2, 2005

The sc-PDB is a collection of 6 415 three-dimensional structures of binding sites found in the Protein Data Bank (PDB). Binding sites were extracted from all high-resolution crystal structures in which a complex between a protein cavity and a small-molecular-weight ligand could be identified. Importantly, ligands are considered from a pharmacological and not a structural point of view. Therefore, solvents, detergents, and most metal ions are not stored in the sc-PDB. Ligands are classified into four main categories: nucleotides (< 4-mer), peptides (< 9-mer), cofactors, and organic compounds. The corresponding binding site is formed by all protein residues (including amino acids, cofactors, and important metal ions) with at least one atom within 6.5 Å of any ligand atom. The database was carefully annotated by browsing several protein databases (PDB, UniProt, and GO) and storing, for every sc-PDB entry, the following features: protein name, function, source, domain and mutations, ligand name, and structure. The repository of ligands has also been archived by diversity analysis of molecular scaffolds, and several chemoinformatics descriptors were computed to better understand the chemical space covered by stored ligands. The sc-PDB may be used for several purposes: (i) screening a collection of binding sites for predicting the most likely target(s) of any ligand, (ii) analyzing the molecular similarity between different cavities, and (iii) deriving rules that describe the relationship between ligand pharmacophoric points and active-site properties. The database is periodically updated and accessible on the web at http://bioinfo-pharma.u-strasbg.fr/scPDB/.

INTRODUCTION

The biological function of a macromolecule is the consequence of molecular interactions with other biologically relevant molecules. The characterization of the three-dimensional (3-D) structure of ligand-binding sites is, therefore, very informative for better understanding functional and physical aspects of proteins and is frequently used for the design and the optimization of protein ligands.[1] Numerous specialized databases exploit the structural data of the Protein Data Bank (PDB)[2] to provide information about complexes between macromolecules and bound ligands. As an example, Ligbase is a database of ligand binding sites aligned to structural templates.[3] Information about the 3-D environment of ligand−protein binding sites (e.g., intermolecular) could be retrieved using PDBsite,[4] Relibase,[5] or the MSDsite part of the Macromolecular Structure Database.[6] All three databases are supplied with query tools and a web interface and provide information about ligands, especially atom and bond types. Chemical features of ligands bound to macromolecules deposited in the PDB can also be retrieved from other electronic libraries of chemicals (HIC-up,[7] PDB-sum,[8] Ligand Depot[9]). In the process of drug discovery, these databases are very helpful for structure-based rational design by exploiting protein-bound conformations of known ligands, depicting their environment and the local flexibility of well-identified protein binding sites. However, described ligands are considered from a structural point of view. This means that no difference is implied between a compound known to activate/inhibit the corresponding target and a molecule (e.g., solvent, detergent, and metal ion) devoid of pharmacological effect on that target.

Recently, Carlson's group linked experimental binding data to 3-D structures from the PDB in the Binding MOAD database.[10] The selection procedure combined with the bibliographic search ensured the choice of the appropriate ligand within biologically relevant complexes. This database resource covers 5331 ligand−protein complexes and binding data for 1375 of them. This outstanding collection of data greatly benefits the characterization of molecular recognition as well as the development of scoring functions and structure-based drug discovery techniques.

To address the issue of ligand selectivity, we recently described an inverse in silico screening method in which a single ligand is serially docked to a collection of protein X-ray structures.[11] This approach requires a well-defined collection of suitable binding sites including exact 3-D coordinates that are not available in any of the above-mentioned databases. We have, therefore, created such a specialized database by parsing PDB files. This database was called sc-PDB for screening-PDB. Initial developments of the database[11] revealed several defects and limitations. For instance, the keyword-based searches for the database creation missed complexes or failed in the valid ligand selection. The rigorous update of the sc-PDB involves numerous improvements which are presented in the current manuscript. First, we have written a new algorithm to ensure an efficient, robust, and accurate detection of valid ligands. Like in the Binding MOAD database, ligands are considered from a pharmacological and not from a structural point of view. Ligands consist only of small nucleotides (< 4-mer),

* To whom correspondence should be addressed. Phone: +33-3-90 24 42 35. Fax: +33-3-90 24 43 10. E-mail: didier.rognan@pharma.u-strasbg.fr.

† Present address: Idéalp'Pharma, Bâtiment CEI, 66 Bd Niels Bohr, F-69603 Villeurbanne Cédex, France.

10.1021/ci050372x CCC: $33.50
Published on Web 01/13/2006

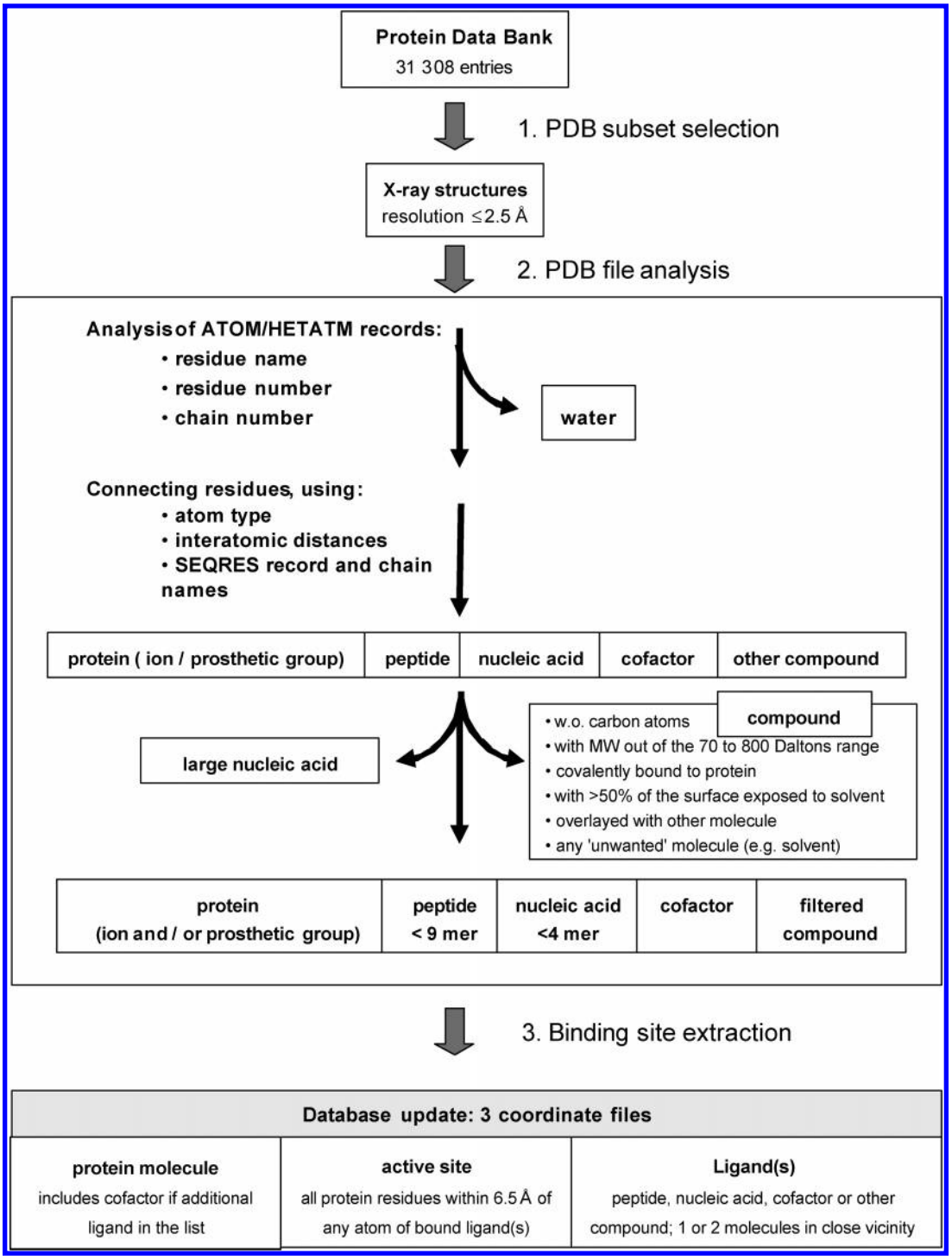

Figure 1. Flowchart of the sc-PDB database setup.

short peptides (< 9-mer), cofactors, and organic compounds. The content of the sc-PDB was also extended to ligand features, therefore, enabling the use of ligand-based virtual screening strategies to predict which protein(s) could be theoretically targeted by a given ligand. Hence, chemical similarity[12] of the cognate PDB ligands to the query ligand would identify the corresponding proteins as putative targets of the query ligand. We also carried out an extensive annotation as well as a clustering of both proteins and ligands. Last, files and annotations are freely accessible on a web resource. The present paper describes the creation, content, and query of the latest update of the sc-PDB.

METHODS

Scripts used in setting up and upgrading the sc-PDB were written in Perl 5.6.1. All data are stored in a relational database (MySQL 4.1) and can be queried through JSP (JavaServer pages).

Database Setup. A total of 21 207 high-resolution (<2.5 Å) crystal structures of the PDB (June 2005) were browsed for binding sites into which small-molecular-weight ligands may fit. Protein cavity selection was based on the detection of an appropriate ligand (Figure 1). A molecule becomes a ligand when it fulfills the following conditions: (i) it is a

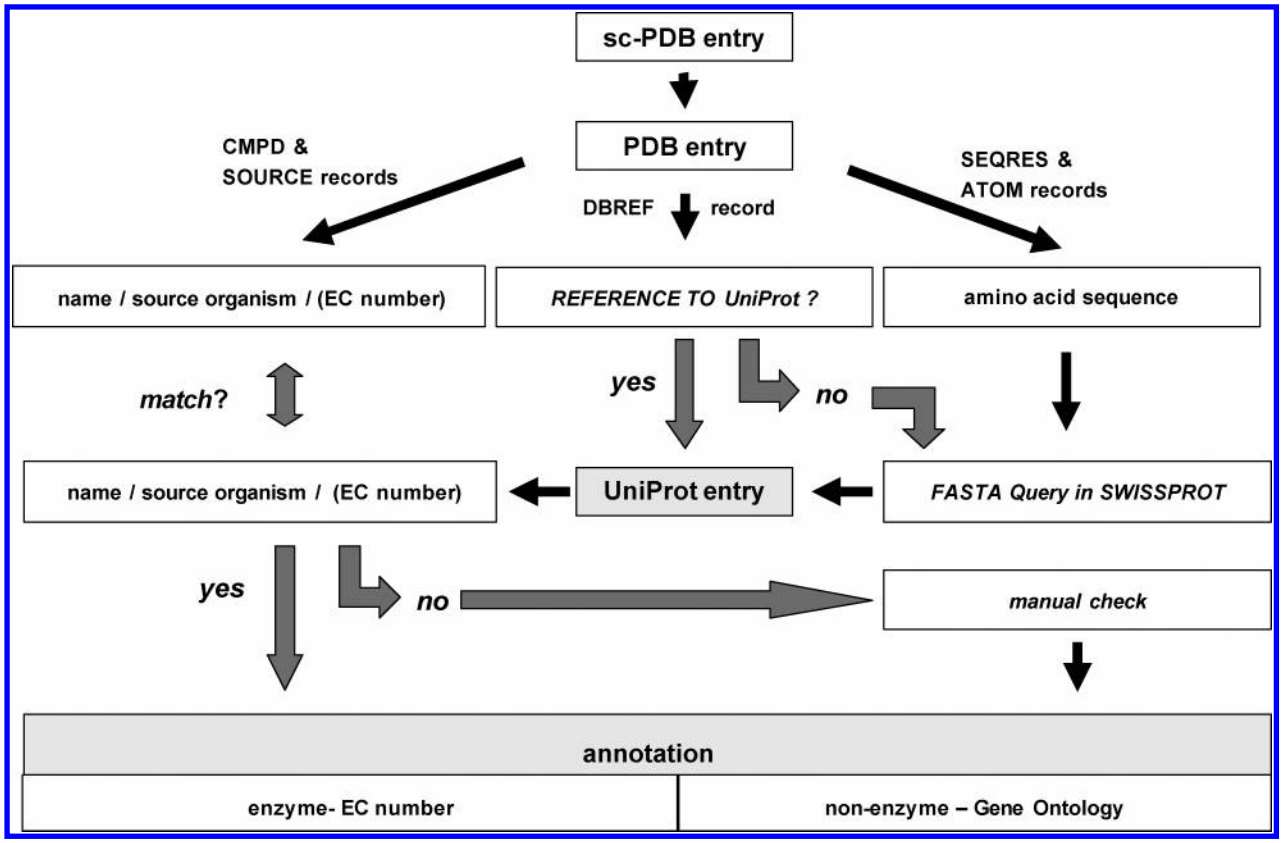

Figure 2. Flowchart of the sc-PDB database annotation.

small-molecular-weight molecule being either a nucleotide, a peptide, an endogenous ligand (e.g., metabolite) or drug but not a water molecule, a metal ion, or an "unwanted molecule" (e.g., solvent, detergent, or protein prosthetic group); (ii) it is not covalently bound to surrounding proteins; or (iii) it has a limited solvent-exposed surface. The binding site is defined by all the protein residues with at least one atom within 6.5 Å of any ligand atom. Protein residues consist of amino acids, as well as cofactors, prosthetic groups, and important ions involved in catalysis or fold stabilization.

Practically, both "ATOM" and "HETATM" records of the coordinate section were scanned to build all molecules described in the PDB file. Consecutive residues were connected provided that one of the following interatomic distances was found: S−S ≤ 2.06 Å, S−X ≤ 1.92 Å, or X−X ≤ 1.70 Å, where S is a sulfur atom and X is any heavy atom. The type of each molecule is then assigned in agreement with the name of its constituting residues: A "protein" includes more than eight natural amino acids. "Prosthetic groups" (heme, porphyrin, chlorophyll, bacteriopheophytin, and ubiquinone) as well as "ions" (Fe, Zn, Ca, Mn, Co, Gd, and Mg) and "cofactors" were identified using three different knowledge-based lists of PDB HET groups. "Prosthetic group" and "ion" molecules were appended to the protein molecule. A "peptide" contains less than eight residues including at least one amino acid. "Nucleotide" comprises standard nucleic acids (C, G, A, T, U, and N). The "compound" type was defined by default for the molecules that do not belong to the "protein", "peptide", "cofactor", or "nucleotide" classes. The molecule list is then cleaned up. In the case that it contains both a peptide and protein, "SEQRES" records and chain names of the PDB entry were considered in order to verify that the peptide is not a part of the protein. Large nucleotides (> 3-mer) were discarded from the molecule list. Other compounds have to pass a series of topological filters in order to remain in the list. Accepted molecules should contain at least one carbon atom, should have a molecular weight (heavy atom only) ranging from 70 to 800 Daltons, should not be covalently bound to any protein atom, should expose less than 50% of their surface to the solvent, and should show no overlaying molecules in the PDB entry. Compounds whose "HET" record indicates water or any "unwanted" molecule are also rejected. Last, if multiple copies of a given compound are present in a single PDB file, only the first one is kept in the molecule list.

The final list of molecules is evaluated to predict a potential binding site(s). If the list is composed of two or more molecules and includes at least one protein, separate coordinate files are generated from the PDB entry for the protein, active site, and ligand. The ligand can be a nucleotide, a peptide, a cofactor, or a compound. The cofactor becomes a ligand if the list is only composed of two molecules, namely, protein and cofactor; otherwise, it is included in the protein molecule. Upon the detection of multiple bound molecules (excepted for cofactors), related protein binding sites are merged if the center of mass of the corresponding molecules are within 5 Å (e.g., reaction products); otherwise, the complete entry is discarded form the database. For the sake of clarity, all molecules bound to active sites, namely, nucleotides, peptides, and organic molecules, will be called ligands from here on.

Binding Site Annotation. The sc-PDB was carefully annotated according to information found in the PDB, the Universal Protein resource (UniProt),[13] and the Gene Ontology (GO) database (Figure 2).[14] When not detectable from

Table 1. Descriptors Used to Annotate Proteins and Ligands of the sc-PDB

protein	ligand
name	molecular weight
UniProt entry	number of H-bond donors and acceptors
simplified generic name	AlogP
source organism	polar surface area
functional mutations (yes/no)	number of rotatable bonds and cycles
functional class	number of Lipinski rules violations
Enzyme Commission number	buried surface area (percentages)
Gene Ontology based cluster	3-D coordinates of the mass center
ion, cofactor, and prosthetic group	

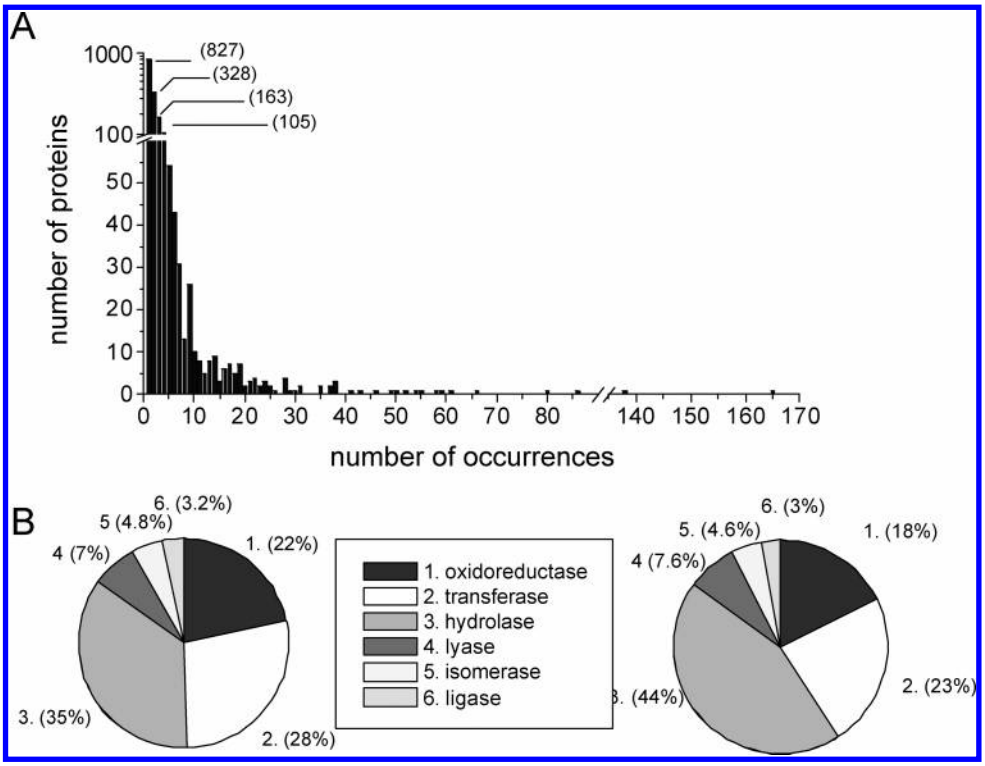

Figure 3. Functional classification of 6 415 sc-PDB protein entries. (A) Database redundancy. Each bar represents the number of proteins that have the same amount of copies in the database. Breaks are applied to the *x* and *y* axes. A logarithmic scaling after the break in the *y* axis allows a compact display of extreme values (indicated between brackets). (B) Distribution of the enzymatic proteins according to the E. C. number in sc-PDB (left, 5479 entries) and in the PDB (right, 14 507 entries). Number of entries as well as related percentage is given for each class.

the "DBREF" record of the PDB entry, the cross-reference to the UniProt database was searched by looking at the most similar sequences (identity above 97% over more than 90% of the target sequence length and expectation below 10 e^{-20}) and retrieved from the Swiss-Prot database[15] using the FASTA program implemented in the Wisconsin Package, version 10.2.[16] FASTA queries used a BLOSUM50 matrix[17] with gap and extension penalties set to −14 and −2, respectively. The input sequence was that of the protein chain(s) involved in the binding site. A unique UniProt identifier was unambiguously assigned to each protein chain of the binding site. Data from PDB and UniProt entries were merged to provide annotation for every sc-PDB entry. If the protein name, enzyme commission (E. C.) number, whenever present, and source organism indicated in the UniProt file did not match those found in the PDB file, the sc-PDB entry annotation was performed manually. The description of the protein(s) that composes the binding site includes the UniProt entry name, protein name found in the sequence file description, the source organism (species, reign), functional mutations, and a simplified generic name (designed in order to easily identify different proteins with identical function, e.g., insulin receptor precursor becomes insulin receptor).

Annotations served to derive a functional hierarchical classification. First, enzyme proteins were segregated according to E. C. rules.[18] Briefly, the E. C. number distinguishes four levels of classes on the basis of the reaction catalyzed, the chemical group involved in the catalysis, and the nature of the substrate. Similarly, nonenzyme protein classification implies two levels that correspond to the different biological processes (e.g., replication/transcription/traduction and signal transduction) and molecular function and is based on the detection of keywords as defined by Gene Ontology annotations.[14] All entries classified in the "immunity" (e.g, immunoglobulins), "oxygen transport" (e.g., globins), and "toxin" classes were removed from the sc-PDB because of their low druggability.[19]

Ligand Annotation. Atomic coordinates of sc-PDB ligands were converted into readable 2-D SD[20] and 3-D MOL2[21] file formats depending on the number of "HET" cards describing them. For single "HET" card-defining ligands, 1-D SMILES strings[22] were retrieved from MSDSite[6]

Table 2. Most Frequently Encountered Proteins in sc-PDB

name	E. C.	number of different UniProt names	number of repeats
trypsin	3.4.21.4	8	165
HIV-1 Pol protease	3.4.23.16	11	138
dihydrofolate reductase	1.5.1.3	9	86
thrombin	3.4.21.5	2	80
glutathione *S*-transferase	2.5.1.18	30	66
streptavidin	na[a]	1	61
thymidylate synthase	2.1.1.45	3	59
flavodoxin	na[a]	7	58
carbonic anhydrase II	4.2.1.1	1	55
nitric-oxide synthase	1.14.13.39	6	54
cell division protein kinase 2	2.7.1.37	2	52
guanyl-specific ribonuclease T1	3.1.27.3	1	50

[a] na: not applicable, in the case that the protein is not an enzyme.

and then converted into 2-D SD and 3-D MOL2 files using Corina 3.0.[23] Multiple "HET" card-defining ligands were converted directly from PDB to 3-D MOL2 format using PRODRG.[24] A visual check was performed for the complete set of ligands to ensure correct atom and bond types. Chemical descriptors and classification by chemotype (Table 1) according to predefined SMARTS strings[25] were then performed using Pipeline Pilot 4.5.[26] The chemical space covered by the sc-PDB ligands was assessed by a classification into graph-based maximum common substructures using the ClassPharmer 3.4 program.[27] The clustering procedure used the "medium" setting for homogeneity and redundancy parameters and allowed fuzzy ring closure. The annotation of sc-PDB entries also includes ligand name and structural properties, as well as structural features characterizing the protein−ligand complex. Accessible surface areas for the ligand (free and bound forms) as well as for the selected active site (free and bound forms) were computed with the SAVOL3 routine implemented in the SYBYL 7.0 package.[21] The percentage of the ligand surface buried upon binding was stored in the final SQL database.

RESULTS

The sc-PDB contains 6415 binding sites for small-molecular-weight ligands. We only considered X-ray structures, which roughly represent 85% of the starting 31 308 PDB entries. Then, by fixing a resolution threshold to 2.5 Å, we again reduced the number of entries. Practically, 21 207 PDB files were parsed to retain only relevant ligand−protein complexes. Selection was based on ligand properties, and a unique druggable cavity was assigned to each complex. Actually, a druggable cavity was found in about one-third of the selected PDB entries. This proportion is consistent with that found in the Binding MOAD database.[10] About

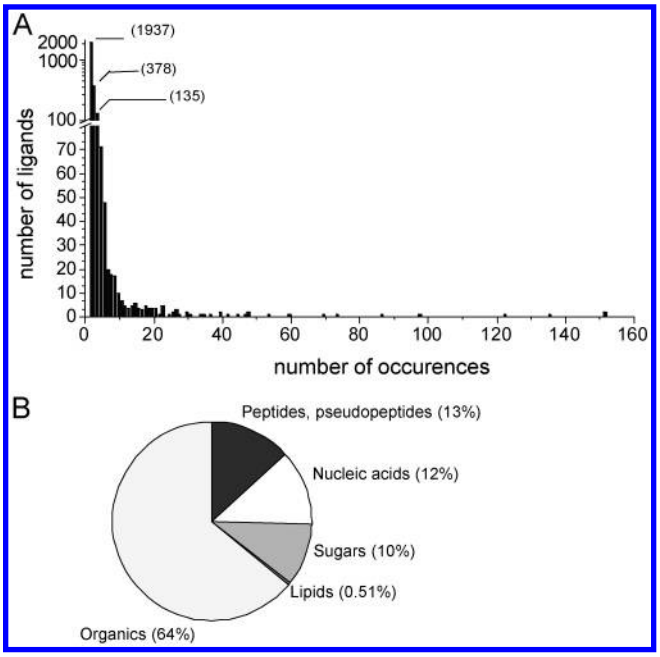

Figure 4. Structural classification of 2721 unique sc-PDB ligand entries. (A) database redundancy. Each bar represents the number of ligands that have the same amount of copies in the database. An axis break associated with a change in scaling (from linear to logarithmic) was applied to the *y* dimension to a allow compact display of extreme values (indicated between brackets). (B) Distribution of the ligands into structural clusters. The peptides and pseudopeptides class contains natural and nonnatural amino acids. Nucleic acids incorporate common purine and pyrimidine bases. Nucleic acids are either bases, nucleosides, nucleotides, nucleotide phosphate like compounds, ribonucleic/deoxyribonucleic acids, or related molecules like the cofactor nicotinamide adenine dinucleotide (NAD). Sugars consist of usual linear and cyclic oses, as well as oligopyranosides. Fatty acids, tricerids, glycerophospholipids, etherglycerophospholipids, sphingolipids, cerides, steroids, and terpens compose the lipid class. All other ligands are included in the organic class.

Table 3. Most Frequently Encountered Ligands in sc-PDB

name	HET code	number of repeats		type[b]
		sc-PDB	PDB[a]	
flavin-adenine dinucleotide	FAD, FAE	152	454	NA/COF
nicotinamide-adenine dinucleotide	NAD, NAH	152	422	NA/COF
adenosine 5′-diphosphate	ADP	135	405	NA/COF
flavin mononucleotide	FMN	122	243	NA/COF
adenosine 5′-triphosphate	ATP	97	240	NA/COF
guanosine 5′-diphosphate	GDP	86	195	NA/COF
nicotinamide-adenine dinucleotide phosphate	NAP	73	230	NA/COF
phosphoaminophosphonic acid-adenylate ester	ANP	69	118	NA
S-adenosyl-L-homocysteine	SAH	59	110	NA
phosphoaminophosphonic acid-guanylate ester	GNP	53	84	NA
benzamidine	BEN, BDN, BAM	47	58	organic
gluthation	GTT, GSH	47	71	organic
glucose	GLC	45	108	sugar

[a] From PDBsum queries using HET groups. [b] Ligand types are defined in the legend of Figure 4. NA and COF are abbreviations for nucleic acids and cofactors, respectively.

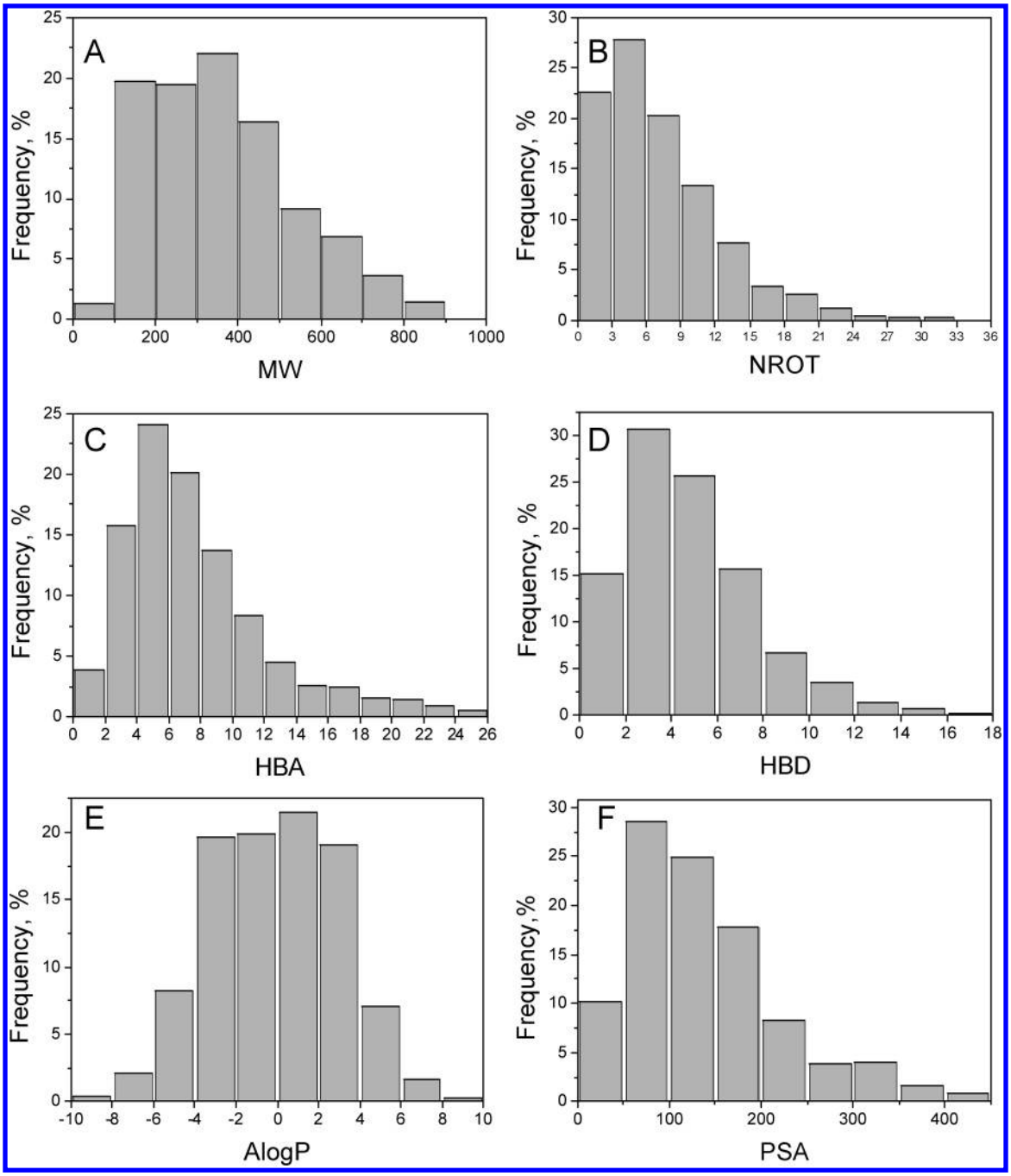

Figure 5. Physicochemical descriptors of the 2721 unique ligands of the sc-PDB. (A) Molecular weight (MW) in Daltons; (B) number of rotatable bonds (NROT); (C) number of H-bond acceptors (HBA); (D) number of H-bond donors (HBD); (E) calculated AlogP; (F) polar surface area (PSA) in Å^2.

500 X-ray complexes were not retained in the sc-PDB because the selected ligand was covalently bound to the protein. Of course, more binding sites could be selected by increasing the resolution threshold (1665 additional sites would have been selected by increasing the upper resolution threshold to 3.0 Å), but the quality of the overall data would be poorer.

Binding sites mostly consist of single peptidic chains, except for 483 entries. Dimeric and trimeric sites were reported for 473 and 10 entries, respectively. Multimeric binding sites either involve identical chains (typical examples of homo-dimeric proteins are HIV-1 protease or human NADPH-quinone oxidoreductase) or two distinct chains of a hetero-multimeric protein (e.g., the human hexameric hormone insulin). Four binding cavities are located at the interface of a binary complex (e.g., the complex between bovine cis−trans isomerase and serine/threonine protein phosphatase 2B, 1tco PDB entry). Upon annotation, 2193 different protein sequences were identified using UniProt entry names. However, since a single protein sequence file may encode several proteins (e.g., viral polyprotein) or several protein domains—and conversely, several protein sequence files may describe variants of a single protein—a generic name was given to each sc-PDB entry, according to data collected in the PDB, UniProt, and GO databases. The 1706 assigned generic names were used to evaluate the redundancy of the sc-PDB (Figure 3A). One unique entry was found for less than the half of the proteins, whereas

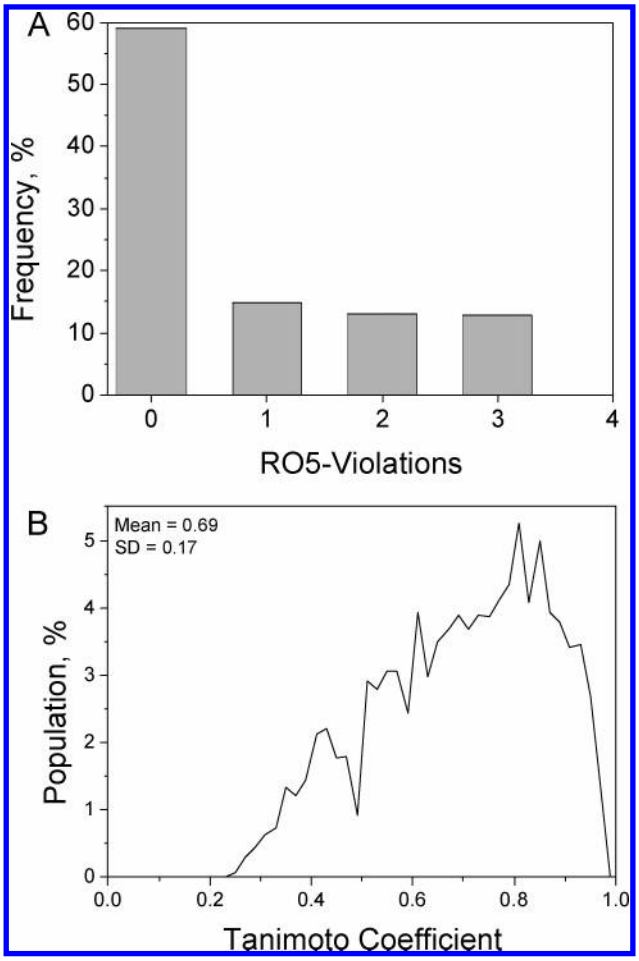

Figure 6. Chemoinformatic analysis of sc-PDB ligands. (A) Number of violations of the Lipinski's rule of five; (B) molecular diversity (self-similarity plot).

few proteins (about 6.8%) are highly redundant with more then 10 occurrences and cover 45% of the database. Actually, more than 50 entries are available for 10 proteins (Table 2). It is noteworthy that Pol proteins of the human immunodeficiency virus (HIV) solely constitute 2.5% of the sc-PDB, although the proportion of viral proteins in the database is limited to 5.3%. About 40% of the organisms represented in the sc-PDB are prokaryotes. A total of 53% belong to eukaryotes and include 68% of mammals (human being the predominant species). The remaining 2.8% represent archebacteria. The function-based classification of sc-PDB entries distinguishes 85.4% of enzymes and 14.6% of nonenzymatic proteins. By way of comparison, enzymes are present in less than 50% of the entries of the PDB March 2005 release. Therefore, our selected PDB subset is enriched in enzymes, which are the principal molecular targets of most drugs. The distribution of enzymatic entries of the sc-PDB into E. C. classes is similar to that observed in the PDB (Figure 3B). In both cases, the hydrolase family is the most populated, yet a small reduction of the hydrolase proportion in the sc-PDB is beneficial to the oxidoreductase class. The statistics on the overall distribution of enzymes indicate that 823 E.C. numbers are covered by the database.

The sc-PDB contains 2721 unique ligands within the 6415 complexes. In 224 binding sites were found two different ligands. The most frequent ligands are principally cofactors (Table 3). Together, FAD, NAD, ADP, and FMN represent about 8% of the total entries. Although few ligands are highly repeated, most of them are only observed once in the database (Figure 4A). A classification of all ligands by chemotype is proposed (Figure 4B). A large majority of the ligands (64%) are small-molecular-weight organic compounds. The remaining molecules are equally distributed into peptide, nucleic acid, and sugar classes (please note that our classification is based on generic SMARTS strings, thereby allowing a fuzzy classification). The physicochemical descriptors of the unique ligands shown in Figure 5 suggest that most ligands of the sc-PDB meet a "druglike" defini-

```
HEADER    HYDROLASE                                16-JUL-04   1W3L
TITLE     ENDOGLUCANASE CEL5A FROM BACILLUS AGARADHAERENS IN
TITLE    2 COMPLEX WITH CELLOTRI DERIVED-TETRAHYDROOXAZINE
/ . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . /
SEQRES   1 A  303  ASP ASN ASP SER VAL VAL GLU GLU HIS GLY GLN LEU SER
SEQRES   2 A  303  ILE SER ASN GLY GLU /. . . . .  . . . . . . . . . .
 . . . . . . . . . . . . . . . . . . . . . . / THR GLY GLY TRP THR GLU
SEQRES  23 A  303  ALA GLU LEU SER PRO SER GLY THR PHE VAL ARG GLU LYS
SEQRES  24 A  303  ILE ARG GLU SER
HET    GOL  A1304       8
HET    GOL  A1305      18
HET    SO4  A1306       5
HET    BGC  A1307      11
HET    BGC  A1308      11
HET    OXZ  A1309      10
/ . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . /
LINK         C1  BGC A1307                 O4  BGC A1308     1555   1555
LINK         C1  BGC A1308                 O4  OXZ A1309     1555   1555
/ . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . /
```

Figure 7. Selected pieces of the 1w3l PDB entry. The title indicates that the ligand of the protein is "cellotri derived-tetrahydrooxazine". The HET section reveals six nonstandard residues: two GOL, SO4, two BGC, and OXZ. GOL and SO4 are glycogen and sulfate molecules, respectively. The two BGC and OXZ are β-D-glucoses and tetrahydrooxazine, respectively, and compose the true protein ligand. According to HET records, all six groups are assigned to the same chain as the protein (chain A). An analysis of LINK or CONECT records is required to properly connect the three residues of the ligand.

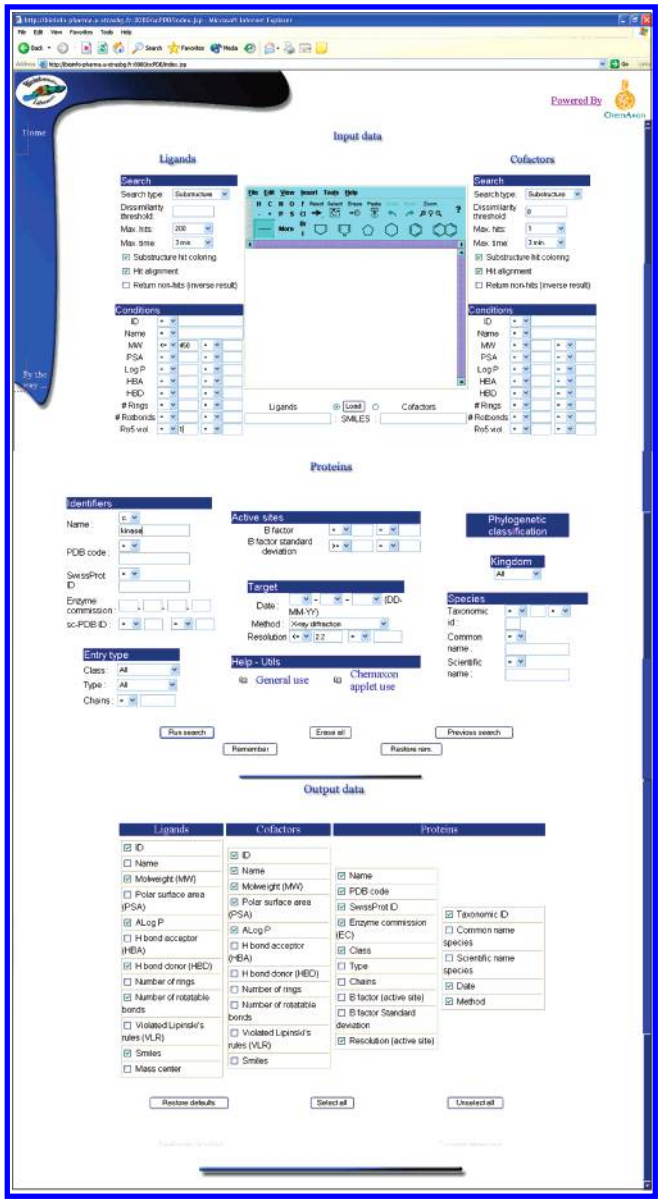

Figure 8. Example of a query input to sc-PDB: ligand-related fields (top panel); target-related fields (middle panel); choice of output data (bottom panel).

tion.[28] About 75% of the ligands violate only one Lipinski rule of five,[29] with 60% presenting no violation at all (Figure 6A). The self-similarity plot using the Tanimoto metric and SciTegic FCFP_4 fingerprints[26] indicates a moderate diversity of ligands (Figure 6B) with a mean Tanimoto coefficient of 0.69.

DISCUSSION

Efficiency and Accuracy of Ligand Detection. In the first sc-PDB release,[11] ligand detection was based on the analysis of the HET groups in selected PDB entries. HET groups define all the molecules, part molecules, or nonstandard residues whose coordinates are reported in the HETATM records of PDB files. A total of 5625 different HET groups were listed in PDBsum in February 2005. Because of the limitations of the atom naming conventions used by the PDB, large HET groups are broken into subgroups. The content of the HET records (e.g., the chain identifier or the token PART_OF in the text field), as well as the residue sequences in SEQRES records, were evaluated to relate individual components of a large HET group. However, the automated assembly of large ligands remains a tricky task, and we did note that ligand identification based on PDB HET and SEQRES records is only sometimes misleading (see example in Figure 7).

In the present work, both ligand and receptor binding sites were detected by an acute analysis of all molecules described in the PDB file coordinate section (ATOM and HETATM records). Ligand molecules that are composed of several residues were built according to interatomic distances between residues. Moreover, extensive automated curation using HET codes and ligand computed descriptors were performed to remove invalid ligands. Last, ligands were filtered out according to both physicochemical features and the protein binding mode (e.g., about 500 X-ray complexes were not retained in the sc-PDB because the ligand was covalently bound to the protein), and PDB entries with incomplete ligand atom coordinates were discarded. Crystal water molecules were not retained as part of the cavities because their locations are highly dependent on the ligand of interest. Indeed, recent developments of several docking programs (e.g., FlexX[30] and Gold[31]) include the automatic placement of water during the docking procedure. Serial docking of ligands to our collection of active sites should, thus, not be penalized that much by removing water molecules from the input PDB files.

In sc-PDB, a binding site is derived from the bound ligand. Assuming that precedented ligand occupancy is a key factor for druggability, we have chosen not to systematically compute all possible protein cavities[32,33] to better control the selection of binding sites.

File Curation. Special attention was given to accurately describe each ligand in standard molecular formats for small-molecular-weight compounds. Visual checks were systematically performed to ascertain atomic valence, hybridization state, and ionization state. Significant advances have been recently reported in the field of molecular interpretation of 3-D coordinates.[34,35] However, the success of the automatic perception highly depends on the quality of the input PDB structure. Indeed, no hydrogen atoms are present in the X-ray structure of PDB ligands. Moreover, large coordinate errors and highly strained geometries are often observed. Therefore, even a good converter cannot easily cope with imperfect input coordinates. The manual check of all ligands was indeed a cumbersome task, but it had to be done only once. This operation will be less time-consuming for the sc-PDB periodic update (we expect about 500 additional entries every 6 months by considering the actual growth rate of the PDB).

Since proteins and active sites of the sc-PDB have to be correctly processed by docking programs, coordinates extracted from the PDB entries were subjected to curation in order to produce standardized files. Residues were (re)-ordered according to the protein sequence. IUPAC atom nomenclature and the atom numbering scheme were checked for all amino acid residues. Other residues were left unchanged if covalently bound to any protein amino acid residue. Incomplete amino acid residues were automatically removed from the protein files. The sc-PDB annotation indicates incomplete amino acids in close vicinity of the ligand(s). In the case of alternate locations, only the coordinate set with the highest occupation rate was kept.

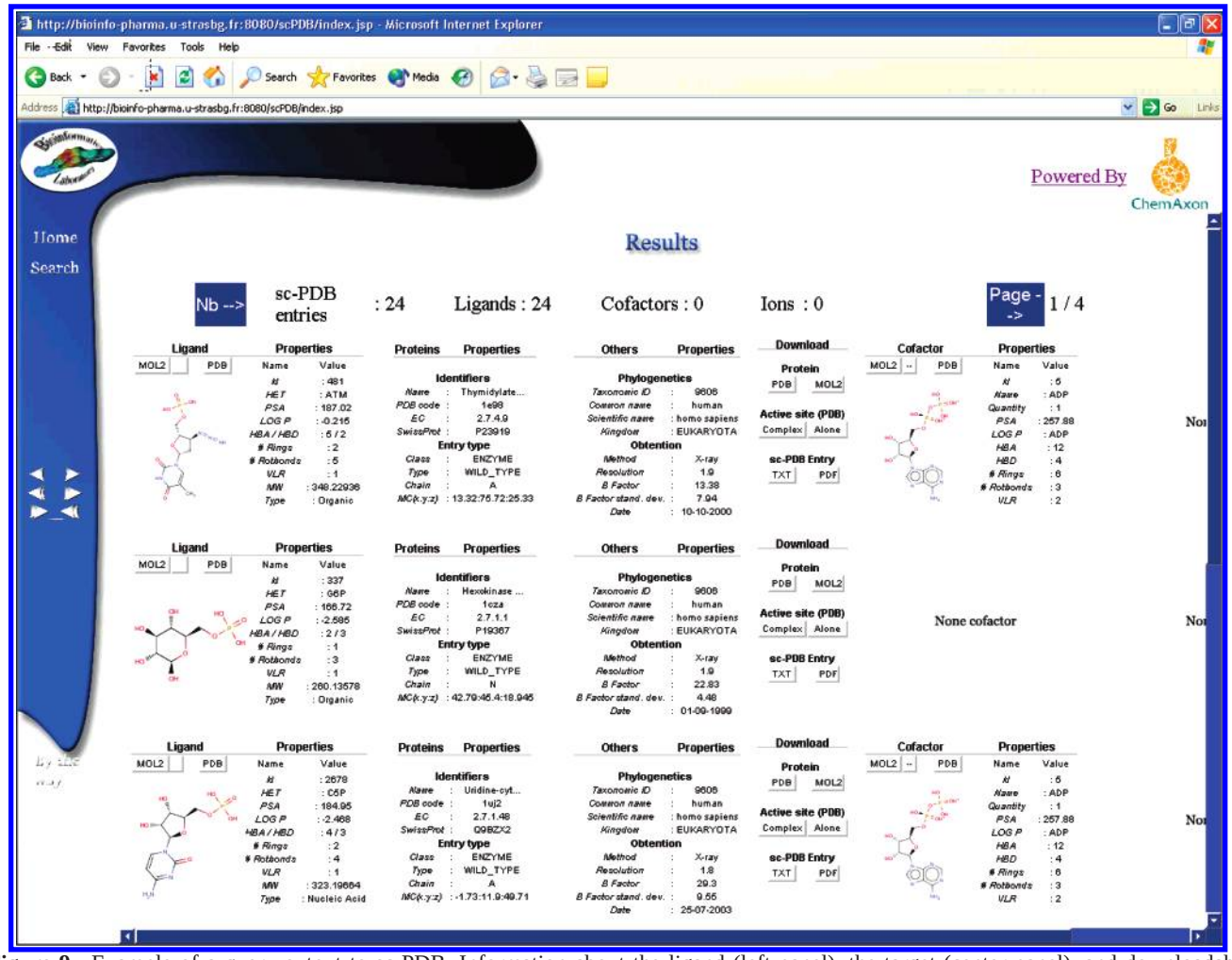

Figure 9. Example of a query output to sc-PDB. Information about the ligand (left panel), the target (center panel), and downloadable cleaned coordinates (right panel) in various file formats is directly accessible.

Water molecules, whose location is ligand-dependent, were rejected from the binding site description to allow unbiased structure-based virtual screening.

Accuracy of Functional Annotation. Accuracy and uniformity of annotation alleviate the comparison of data across the sc-PDB, hence, allowing the rapid and efficient processing of the data resulting from an inverse docking strategy. Because of the heterogeneity of PDB file contents, a simple keyword search is not adapted to the sc-PDB annotation. Moreover, functional annotation of binding sites cannot directly rely on information found in the PDB header, because of missing or erroneous data. The use of the protein sequence enables the retrieval of the UniProt sequence file, which subsequently affords links to Enzyme and Gene Ontology (GO) entries. A recently published GO annotation of PDB files was also implemented.[36] The enzyme classification regroups enzymes with similar chemical functionality. Within these classes, proteins are grouped into homologous protein "families" on the basis of sequence. GO annotation provides the molecular function, biological process, and cellular component of a gene product using controlled vocabulary. Integration of the data provides protein descriptors (Table 1) as well as functional classification for enzyme and nonenzyme proteins. Assigning reliable information to enzymatic peptidic chains still remains difficult. Actually, E. C. numbers found in both UniProt and PDB files are often assigned to whole protein chain sequences rather than functional units (e.g., sites for catalysis, noncompetitive inhibition, or allosteric modulation). Further sc-PDB developments will aim at the clustering of the different binding cavities of a given enzyme by the evaluation of either sequence or structure similarity (see further improvement of the sc-PDB) and the use of the SCOPEC database resource.[37]

Example of Query. Queries can be performed from ligand, cofactor, or protein attributes. For example, one can combine searches on the "ligand" (e.g., any ligand of molecular weight lower than 450, showing one violation of Lipinski rules) and "protein" tables (e.g., all kinase X-ray structures of resolution lower than 2.2 Å) to retrieve information about the corresponding complexes in a user-customizable way (Figure 8). The sc-PDB output page is straightforward to analyze (Figure 9). Molecular properties as well as downloadable 3-D coordinates (protein, active site, ligands, and cofactor if any) in standard molecular formats are directly accessible. Hyperlinks to external databases (SwissProt, PDB, and PubChem) will be soon implemented to assist the user navigating in sc-PDB-related resources.

Further Improvements. We are currently working on a nonredundant data set of the sc-PDB repertoire by choosing a representative protein from each functional family of enzymes present in the database. Family homogeneity is assessed by sequence identity. Representatives are chosen by best resolution, low B factor, absence of mutation, and size of the ligand. In the near future, the description of the

binding sites will include simple physical parameters (volume and mouth opening of the protein cavity including cofactor and ions) that can serve to establish another classification of sc-PDB protein entries. Sophisticated methods based on the analysis of amino acid composition, sequence-dependent (motif search) or independent (residues propensities, spatial organization of pharmacophoric features, and local structural similarity), may also give evidence for unexpected protein similarity, which can be highly relevant for drug discovery.

The next sc-PDB update will take advantage of the ongoing efforts of the Research Collaboratory for Structural Bioinformatics (RCSB) to make data uniform in the PDB,[38,39] Standardization of nomenclature and usage as well as additional checks to rectify sequence and E. C. errors were conducted for all PDB files. Corrections were released in the mmCIF format for the PDB entries. Our scripts will be modified to analyze mmCIF files, thereby improving sc-PDB data accuracy and consequently facilitating integration with other database resources required for annotation. The sc-PDB would also benefit from the use of the new website of the RCSB PDB, which gives access to new features such as the simplified protein name or the mol file of HET groups.

CONCLUSION

A curated database of 6415 binding sites has been created by parsing PDB files and carefully identifying, for each entry, the protein(s) to which it belongs as well as bound small molecules (ions, cofactors, and ligands). Importantly, the ligand is considered from a pharmacological point of view. Therefore, we assume that most cavities stored in the sc-PDB are druggable from a structural point of view.[19] Three-dimensional coordinates in readable formats are provided for the protein, the active site, and the bound small molecules. The database can be browsed by either target-based or ligand-based queries, therefore, enabling a logical chemogenomic interface between all stored data. Four potential applications of increasing importance in modern structure-based drug discovery are directly possible with the current version of the sc-PDB: (i) establishing large docking benchmarks to evaluate the accuracy of docking tools,[40] (ii) predicting the most likely targets of any ligand by browsing the collection of protein active sites through serial docking,[11] (iii) delineating the chemical similarity of binding sites by structural alignment techniques,[32,41] and (iv) setting up chemogenomic links between ligand and target spaces and, thus, facilitating the design of focused or targeted compound libraries.[42] The sc-PDB is regularly updated and available at http://bioinfo-pharma.u-strasbg.fr/scPDB/

ACKNOWLEDGMENT

The authors acknowledge the French Ministry of Science and Technology, the National Center for Scientific Research (CNRS), and the Alsace-Lorraine Genopole for funding.

REFERENCES AND NOTES

(1) Tickle, I.; Sharff, A.; Vinkovic, M.; Yon, J.; Jhoti, H. High-throughput protein crystallography and drug discovery. *Chem. Soc. Rev.* **2004**, *33*, 558−565.

(2) 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.

(3) Stuart, A. C.; Ilyin, V. A.; Sali, A. LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures. *Bioinformatics* **2002**, *18*, 200−201.

(4) Ivanisenko, V. A.; Pintus, S. S.; Grigorovich, D. A.; Kolchanov, N. A. PDBSite: a database of the 3D structure of protein functional sites. *Nucleic Acids Res.* **2005**, *33*, D183−D187.

(5) 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.

(6) Golovin, A.; Oldfield, T. J.; Tate, J. G.; Velankar, S.; Barton, G. J.; Boutselakis, H.; Dimitropoulos, D.; Fillon, J.; Hussain, A.; Ionides, J. M.; 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, G. J.; Tagari, M.; Tromm, S.; Vranken, W.; Henrick, K. E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res.* **2004**, *32*, D211−D216.

(7) Kleywegt, G. J.; Jones, T. A. Databases in protein crystallography. *Acta Crystallogr., Sect. D* **1998**, *54*, 1119−1131.

(8) Laskowski, R. A.; Chistyakov, V. V.; Thornton, J. M. PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Res.* **2005**, *33*, D266−D268.

(9) Feng, Z.; Chen, L.; Maddula, H.; Akcan, O.; Oughtred, R.; Berman, H. M.; Westbrook, J. Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics* **2004**, *20*, 2153−2155.

(10) Hu, L.; Benson, M. L.; Smith, R. D.; Lerner, M. G.; Carlson, H. A. Binding MOAD (Mother Of All Databases). *Proteins* **2005**, *60*, 333−340.

(11) Paul, N.; Kellenberger, E.; Bret, G.; Muller, P.; Rognan, D. Recovering the true targets of specific ligands by virtual screening of the protein data bank. *Proteins* **2004**, 54, 671−680.

(12) Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2* (22), 3204−18.

(13) Apweiler, R.; Bairoch, A.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M. J.; Natale, D. A.; O'Donovan, C.; Redaschi, N.; Yeh, L. S. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* **2004**, *32*, D115−D119.

(14) Harris, M. A.; Clark, J.; Ireland, A.; Lomax, J.; Ashburner, M.; Foulger, R.; Eilbeck, K.; Lewis, S.; Marshall, B.; Mungall, C.; Richter, J.; Rubin, G. M.; Blake, J. A.; Bult, C.; Dolan, M.; Drabkin, H.; Eppig, J. T.; Hill, D. P.; Ni, L.; Ringwald, M.; Balakrishnan, R.; Cherry, J. M.; Christie, K. R.; Costanzo, M. C.; Dwight, S. S.; Engel, S.; Fisk, D. G.; Hirschman, J. E.; Hong, E. L.; Nash, R. S.; Sethuraman, A.; Theesfeld, C. L.; Botstein, D.; Dolinski, K.; Feierbach, B.; Berardini, T.; Mundodi, S.; Rhee, S. Y.; Apweiler, R.; Barrell, D.; Camon, E.; Dimmer, E.; Lee, V.; Chisholm, R.; Gaudet, P.; Kibbe, W.; Kishore, R.; Schwarz, E. M.; Sternberg, P.; Gwinn, M.; Hannick, L.; Wortman, J.; Berriman, M.; Wood, V.; de la Cruz, N.; Tonellato, P.; Jaiswal, P.; Seigfried, T.; White, R. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **2004**, *32*, D258−D261.

(15) Boeckmann, B.; Bairoch, A.; Apweiler, R.; Blatter, M. C.; Estreicher, A.; Gasteiger, E.; Martin, M. J.; Michoud, K.; O'Donovan, C.; Phan, I.; Pilbout, S.; Schneider, M. The SWISS−PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **2003**, *31*, 365−370.

(16) *Wisconsin Package*, version 10.2; Genetics Computer Group: Madison, WI 53711.

(17) Henikoff, S.; Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 10915−10919.

(18) Bairoch, A. The ENZYME database in 2000. *Nucleic Acids Res.* **2000**, *28*, 304−305.

(19) Hopkins, A. L.; Groom, C. R. The druggable genome. *Nat. Rev. Drug Discovery* **2002**, *1*, 727−730.

(20) SD file format, MDL Information Systems Inc.: San Leandro, CA.

(21) MOL2 file format, Tripos Inc.: St. Louis, MO.

(22) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(23) *Corina*, version 3.0; Molecular Networks GmbH: Erlangen, Germany.

(24) Schuttelkopf, A. W.; van Aalten, D. M. PRODRG: a tool for high-throughput crystallography of protein−ligand complexes. *Acta Crystallogr., Sect. D* **2004**, *60*, 1355−63.

(25) *SMARTS*; Daylight Chemical Information Systems, Inc.: Mission Viejo, CA.

(26) *Pipeline Pilot*, version 4.5; SciTegic: San Diego, CA.

(27) *ClassPharmer*, version 3.4; Bioreason Inc.: Santa Fe, NM.

(28) Charifson, P. S.; Walters, W. P. Filtering databases and chemical libraries. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 311−323.

(29) 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.* **2001**, *46*, 3−25.
(30) Rarey, M.; Kramer, B.; Lengauer, T. The particle concept: placing discrete water molecules during protein−ligand docking predictions. *Proteins* **1999**, *34*, 17−28.
(31) Verdonk, M. L.; Chessari, G.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Nissink, J. W.; Taylor, R. D.; Taylor, R. Modeling water molecules in protein−ligand docking using GOLD. *J. Med. Chem.* **2005**, *48*, 6504−6515.
(32) An, J.; Totrov, M.; Abagyan, R. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol. Cell. Proteomics* **2005**, *4*, 752−761.
(33) Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graphics Modell.* **1997**, *15*, 359−363.
(34) Labute, P. On the Perception of Molecules from 3D Atomic Coordinates. *J. Chem. Inf. Model.* **2005**, *45*, 215−221.
(35) Wolber, G.; Langer, T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* **2005**, *45*, 160−169.
(36) Ponomarenko, J. V.; Bourne, P. E.; Shindyalov, I. N. Assigning new GO annotations to protein data bank sequences by combining structure and sequence homology. *Proteins* **2005**, *58*, 855−865.
(37) George, R. A.; Spriggs, R. V.; Thornton, J. M.; Al-Lazikani, B.; Swindells, M. B. SCOPEC: a database of protein catalytic domains. *Bioinformatics* **2004**, *20*, I130−I136.
(38) Bhat, T. N.; Bourne, P.; Feng, Z.; Gilliland, G.; Jain, S.; Ravichandran, V.; Schneider, B.; Schneider, K.; Thanki, N.; Weissig, H.; Westbrook, J.; Berman, H. M. The PDB data uniformity project. *Nucleic Acids Res.* **2001**, *29*, 214−218.
(39) Westbrook, J.; Feng, Z.; Jain, S.; Bhat, T. N.; Thanki, N.; Ravichandran, V.; Gilliland, G. L.; Bluhm, W.; Weissig, H.; Greer, D. S.; Bourne, P. E.; Berman, H. M. The Protein Data Bank: unifying the archive. *Nucleic Acids Res.* **2002**, *30*, 245−248.
(40) 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.
(41) Schmitt, S.; Kuhn, D.; Klebe, G. A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* **2002**, *323*, 387−406.
(42) Savchuk, N. P.; Balakin, K. V.; Tkachenko, S. E. Exploring the chemogenomic knowledge space with annotated chemical libraries. *Curr. Opin. Chem. Biol.* **2004**, *8*, 412−417.

CI050372X